ISSN : 2456-3307

IJSR CSEIT

विज्ञानं ब्रह्म

International Conference on Machine Learning & Computational Intelligence - ICMLCI-2017

Organised by

Department of Computer Science & Engineering
Shri Mata Vaishno Devi University, Katra, J&K 182320, India

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN COMPUTER SCIENCE, ENGINEERING AND INFORMATION TECHNOLOGY

Volume 2, Issue 7, September-2017

Email: editor@ijsrcseit.com

International Conference on Machine Learning and Computational Intelligence

[ICMLCI-2017]

27-28 Sept, 2017

Sponsored by: TEQIP - III

Organized by

Department of Computer Science & Engineering
Shri Mata Vaishno Devi University, Katra, J&K 182320, India

In Association with
International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307
Volume 2, Issue 7, September-2017

Published By
Technoscience Academy
(The International Open Access Publisher)

[www.technoscienceacademy.com]

About Us

The Department of Computer Science & Engineering has been set up to impart training of the highest standards to the students, in the field of computers, thus preparing them to meet the exacting demands of the highly competitive global industrial market. The Department is currently focused on providing B.Tech., M.Tech., M.C.A. & PhD programs which cover exhaustively the area of Computer Science & Engineering.

Aim & Objectives

- Bring researchers and experts together to discuss and share their experiences
- Share the current and new research topics and ideas
- Improve and enhance personal, enterprise, national and international awareness
- Provide a platform to present and discuss recent advancements
- Increase international collaborations among university–industry-institutions

Theme of the Conference

The International Conference on Machine Learning & Computational Intelligence(ICMLCI-2017) aims to provides an opportunity for the researchers, engineers, developers and practitioners from academia and industry to discuss and address the experimental, theoretical work and methods in solving problems and to share their experience, exchange and cross-fertilize their ideas in the fields of Machine Learning and Computational Intelligence. This conference will be useful for research scholar and professionals working in data science, big data, machine learning, artificial intelligence and predictive modeling.

In the recent years, machine learning has emerged as powerful techniques for solving problem involve huge loads of data. Machine learning provides algorithms and tools that are useful for solving classification and regression tasks in the supervised learning problems, clustering and frequent pattern mining tasks in unsupervised learning problems, tasks of controlling the behavior of machines in reinforcement learning problems.

Technical Publishing Committee Member

Prof. Suresh Dara,Department of CSE, BVRIT, Narsapur Telengana,India
Dr. Baijnath Kaushik,Department of CSE,SMVD University, Katra, J&K, India

Organizing Committee

Dr. Manoj Gupta,SMVDU,Jammu,India
Dr. Abhishek Gupta,SMVDU,Jammu,India
Dr. Sunanda Gupta, SMVDU,Jammu,India
Dr. Sakshi Arora, SMVDU,Jammu,India
Mr Manoj Kumar, SMVDU,Jammu,India
Mrs Sonika Gupta,, SMVDU,Jammu,India
Mr Sanjay Sharma, SMVDU,Jammu,India
Mrs. Pooja Sharma, SMVDU,Jammu,India
Mr Deo Paraksh, SMVDU,Jammu,India
Mr Anuj Mahajan, SMVDU,Jammu,India
Mr Sudesh Bhadu, SMVDU,Jammu,India
Mr Rajnish Kumar , SMVDU,Jammu,India
Mr Sushil Trisal , SMVDU,Jammu,India
Mr Naveen Kumar , SMVDU,Jammu,India
Mr Deepak Sharma , SMVDU,Jammu,India
Mr Himmat Raj Sharma , SMVDU,Jammu,India
Mrs Nisha Gupta , SMVDU,Jammu,India
Mrs Priyanka Sharma , SMVDU,Jammu,India
Technical Advisory Committee
Dr. Joost van de Weijer, Computer Vision Centre (CVC), Barcelona
Prof. Antonio M. López, Computer Vision Centre (CVC), Barcelona
Dr. Bijoy Chand Chatterjee, Postdoctoral Fellow, UEC, Japan

CONTENTS

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Probabilistic Modeling in Machine Learning and Artificial Intelligence

Sajal Kaushik, Pulkit Kogat, Dr. Narina Thakur
Bharati Vidyapeeth's College of Engineering, A-4, Paschim Vihar, Rohtak Road, New Delhi, India

ABSTRACT

Probabilistic modeling plays a vital role in inferencing from huge datasets with high probability of uncertainity. This paper introduces most common probabilistic models applicable in machine learning found in litreature. An extensive litreature survey on Bayesian Networks, Markov Models,Hidden markov Models and stochastic grammars is captured under this single formalism. It also discusses a generic formalism called Bayesian Programming. The following paper presents various probabilistic modeling techniques used in practical applications of machine learning and machine learning related areas.

Keywords : Bayesianprogramming(bp), pertitentvariables (P), decomposition(d), Bayesian networks(bn), searching(s).

I. INTRODUCTION

The idea of probabilistic framework in machine learning provides models to explain the observed data set .Therefore, a machine can exploit such models to make better predictions about future datasets, and take decisions accordingly that are rational to the given predictions. Uncertainty plays a major role in inferencing. Observed data can be consistent with many models, and therefore which model is appropriate, can be thought of. Similarly, predictions about future data and the future consequences of actions are uncertain. Probability theory provides a framework for modelling uncertainty.

Machine learning is playing a vital role in every industry.machine learning involves training a model on particular dataset.With the huge generation of data, there is uncertainity in data which can be resolved to a larger extent with the help of probabilistic models.

II. LITREATURE REVIEW

Probabilistic modeling [1] is a mainstay of modern machine learning research, providing essential tools for analyzing the vast amount of data that have become available in cognitive science. [2]With the increasing amount of data, uncertainity of predicting the result from the given data has increased. Therefore,[3] requirement of probabilistic models to predict the output for test data. This section discusses various types of probabilistic models used in machine learning.

III. TYPES OF MODELS

A. BAYESIAN MODELS

Bayesian Models(composed of bayesian networks) have been evoloved as a basic approach for dealing with large datasets of probabilistic and uncertain information. These are the outcome of the integration between graph theory and various theories of probability. They are defined by the Bayesian algorithm of

$$
P\left\{\begin{array}{l}
D\left\{\begin{array}{l}
S\left\{\begin{array}{l}
\text{Variables}\\
\quad X^1,\ldots,X^N\\
\text{Decomposition}\\
\quad P(X^1\ldots X^N)=\prod_{i=1}^{N}P(X^i|\,Pa^i),\\
\quad \text{with } Pa^i\subseteq\{X^1,\ldots,X^{i-1}\}\\
\text{Forms: any}
\end{array}\right.\\
\text{Identification: any}
\end{array}\right.\\
\text{Question: } P(X^i\mid Known)
\end{array}\right.
$$

Figure 3: The BN formalism rewritten in BP.

Figure 3. The P(pertitent) variables have no constrains and have no specific meaning. The D(decomposition) variables, on the contrary, are clearly identified: it is a product of distributions of one variable Xi , conditioned by a conjunction of other variables, called its “parents”, P a i , P a i ⊆ {X 1 , . . . , X i−1 }.This presumes that variables are ordered, and fully certains that applying Bayes’ rule correctly defines the joint distribution .Also note that Xi indicates that one and only one variable. Thus, the above model, which can fit in a BP, does not in a BN:

P (A B C D) = P (A B)P (C | A)P (D | B).

In a BN, if A and B are to sight together on the LHS of a term, as in P (A B), then they have to be merged into a single variable Ai, Bi, and cannot be consequently separated, as in P (C | A) and P (D | B).

A bijection exists within joint probability distributions defined by such a decomposition and directed acyclic graphs: nodes are associated to variables, and edges are associated to conditional dependencies. Using graphs in probabilistic models leads to an efficient way to define hypotheses over a set of variables, an economic representation of a joint probability distribution and, most importantly,an easy and efficient way to do probabilistic inference.

The parametric forms are not constrained theoretically, but in BN commercial softwares they are very often restricted to probability tables, or tables and constrained Gaussians.

B. HIDDEN MARKOV MODELS

Hidden markov models are the advance models containing bayesian filters

$$
P\left\{\begin{array}{l}
D\left\{\begin{array}{l}
S\left\{\begin{array}{l}
\text{Variables}\\
\quad S_0,\ldots,S_T,O_0,\ldots,O_T\\
\text{Decomposition}\\
\quad P(S_0\ldots S_T\ O_0\ldots O_T)=\\
\qquad P(S_0)P(O_0\mid S_0)\\
\qquad \prod_{i=1}^{T}\left[P(S_i\mid S_{i-1})P(O_i\mid S_i)\right]\\
\text{Forms}\\
\quad P(S_0)\ \equiv\ \text{Matrix}\\
\quad P(S_i\mid S_{i-1})\ \equiv\ \text{Matrix}\\
\quad P(O_i\mid S_i)\ \equiv\ \text{Matrix}
\end{array}\right.\\
\text{Identification: Baum-Welch algorithm}
\end{array}\right.\\
\text{Question: } P(S_0\ S_1\ldots S_t-1\mid S_t\ O_0\ldots O_t)
\end{array}\right.
$$

Figure 2

Variables are assumed to be discrete. Therefore, the transition model P (S i | S i−1) and the observation model P (O i | S i) are both specified using probability matrices (conditional probability tables).Variants exist about this particular point: when the observation variables are continuous, the formalism becomes known as “semi-continuous HMMs” .In this case, the observation model is associated either with a Gaussian form, or a Mixture of Gaussian form.

C. MARKOV MODELS

In probability theory, a Markov model is a stochastic model used to model randomly and site changing systems where it is assumed that future states are dependent on the present state not on the events that occurred previously (it assumes the Markov property). Generally, this assumption with the hardcoded markov model helps us to improve inferencing and computatability otherwise it is intracable. For this reason, in the fields of predictive modelling and probabilistic forecasting, it is necessary to use markov models for better predictions.

D. STOCHASTIC GRAMMARS

PCFGs extend context-free grammars similar to how hidden Markov models extend regular grammars.Probability is assigned to each production. The probabilities derivation uses the product of product derivations.These probabilities can be viewed as parameters of the model, and for large problems it is convenient to learn these parameters via machine learning.The validity of probabilistic grammar is limited by the context of its training dataset.

PCFGs have application ranging from nlp to rna molecules to design of programming languages. Weighing factors of scalability and generality result in designing of efficient PCFG.

Table 1

OBJECTIVE	**TECHNIQUES USED**				
	Generic probabilistic modelling	**Bayesian Model**	**Markov models**	**Hidden Markov modes**	**Stochastic grammars**
Improving PILCO(a reinforcement learning model)[12]		✓			
Improving ensemble learning[11]	✓				
Analysis of neuroimaging data using fsl[13]		✓			
To improve exploratory behaviour in student’s model[14]	✓				
Sampling and Bayes' Inference in Scientific Modelling and Robustness[15]		✓			
Bayesian exponential random graph modelling of interhospital patient referral networks[16]		✓			
Estimation of distributive algorithms[17]		✓			
Opportunities for Personalization in Modeling Students[18]		✓			
Towards a probabilistic formalisation of case based inference[19]	✓				
Probabilistic modelling of joint orientation[20]	✓				
survey of probabilistic models, using the Bayesian Programming methodology as a unifying framework[21]		✓			

Global behaviour patterns using probabilistic latent semantic analysis[22]	✓				
2-dimension dynamic Bayesian network for large-scaledegradation modelling with an application bridges network[23]		✓			
Complexity of inference of probabilistic models[24]	✓				
Constrained Bayesian networks [25]		✓			
Reliabilty modelling with dynamic Bayesian networks[26]		✓			
A probabilistic approach to modelling uncertain arguments[27]	✓				
Probabilistic modelling ,inference and learning using logical theories[28]	✓				
Probabilistic models of cognition[29]					✓
Probabilistic sensitivity analysis of complex models[30]		✓			
Statistical inference and probabilistic modelling for constained based NLP[31]					✓
Recognisition of patterns[32]			✓		
Speech recognisition[33]				✓	
Learning and detecting activities from movement trajectories using the hierachial hidden markov model[34]				✓	
Large margin hidden markov models for automatic speech recognisition[35]				✓	
Hidden markov models in computer vision[36]				✓	

IV. DESCRIPTION OF TABLE-1 AND PIE-CHART

The above table describesreal life problems in which probabilistic modelling is used ranging from speech recognisition[35] ,computer vision[36],nlp to distributed algorithms specifically in all machine learning related fields.It provides a abstract of the problem and technique of probabilistic modelling used.pie-chart simply presents the percentage of technique used in various applications mentioned in the literature review of the paper.table-1 presents the wide-ranging applications in which probabilistic modelling can be used in all machine learning related fields. Further proving that there is a huge scope in on improving the probabilistic models for removing the uncertainity in the data.

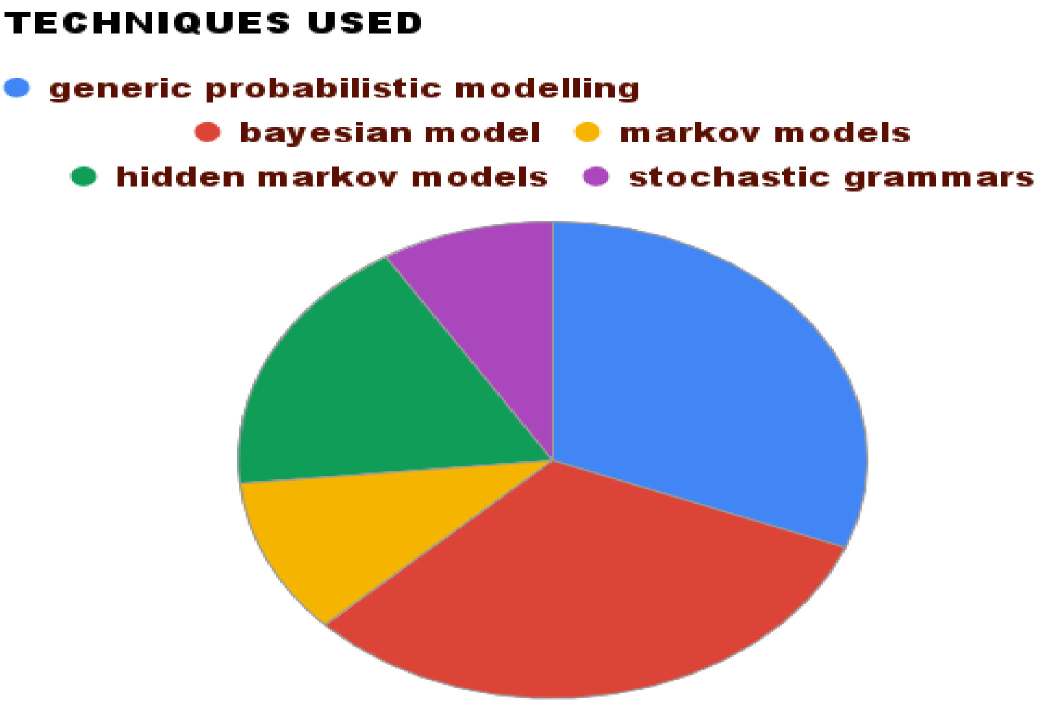


Figure 4

V. RECOMMENDATIONS

- ✓ it has immense work in computer vision where there is lot of uncertainity in the data.
- ✓ Stochastic grammars are mostly used in recommender systems where they usually generate a sentence.
- ✓ In concepts like pca in dimension reduction, probabilistic modelling is used.
- ✓ Probabilistic modeling will play a important role in nlp related field.

VI. CONCLUSION

Probabilistic approaches in machine learning is a very active research area with wide-ranging impact beyond conventional pattern-recognition problems .The key difference between problems in which a probabilistic approach is important and problems that can be solved using non-probabilistic machine-learning approaches is whether uncertainty has a central role. Moreover, most conventional optimization-based machine-learning approaches have probabilistic analogues that handle uncertainty in a more principled manner.

VII. REFERENCES

[1]. Bishop, C. M. Pattern Recognition and Machine Learning (Springer,2006).

[2]. Murphy, K. P. Machine Learning: A Probabilistic Perspective (MIT Press, 2012).

[3]. Ghahramani, Z. Bayesian nonparametrics and the probabilistic approach to modelling. Phil. Trans. R. Soc. A 371, 20110553 (2013).

[4]. Jaynes, E. T. Probability Theory: the Logic of Science (Cambridge Univ. Press, 2003).

[5]. Koller, D. & Friedman, N. Probabilistic Graphical Models: Principles and Techniques (MIT Press, 2009).

[6]. Doya, K., Ishii, S., Pouget, A. & Rao, R. P. N. Bayesian Brain: Probabilistic Approaches to Neural Coding (MIT Press, 2007).

[7]. Jordan, M., Ghahramani, Z., Jaakkola, T. & Saul, L. An introduction to variational methods in graphical models. Mach. Learn. 37, 183–233 (1999).

[8]. Pfeffer, A. Practical Probabilistic Programming (Manning, 2015).

[9]. Bishop, C. M. Model-based machine learning. Phil. Trans. R. Soc. A 371, 20120222 (2013).

[10]. Mansinghka, V., Selsam, D. & Perov, Y. Venture: a higher-order probabilistic programming platform with programmable inference. Preprint at http://arxiv. org/abs/1404.0099 (2014).

[11]. MacKay, David JC. "Developments in probabilistic modelling with neural networks-ensemble learning." In Neural Networks: Artificial Intelligence and Industrial Applications. Proc. of the 3rd Annual Symposium on Neural Networks,pp.191-198. 1995.

[12]. Gal, Yarin, Rowan Thomas McAllister, and Carl Edward Rasmussen. "Improving PILCO with bayesian neural network dynamics models." In Data-Efficient Machine Learning workshop, vol. 951, p. 2016. 2016.

[13]. Woolrich, Mark W., Saad Jbabdi, Brian Patenaude, Michael Chappell, Salima Makni, Timothy Behrens, Christian Beckmann, Mark Jenkinson, and Stephen M. Smith."Bayesian analysis of neuroimaging data in FSL." Neuroimage 45, no. 1 (2009): S173-S186.

[14]. Bunt, Andrea, and Cristina Conati. "Probabilistic student modelling to improve exploratory behaviour." User Modeling and User-Adapted Interaction 13, no. 3 (2003): 269-309.

[15]. Box, GEORGE EP. "Sampling and Bayes' inference in scientific modelling and robustness." Journal of the Royal Statistical Society A 143 (1980): 383-430.

[16]. Caimo, Alberto, Francesca Pallotti, and Alessandro Lomi. "Bayesian exponential random graph modelling of interhospital patient referral networks." Statistics in Medicine (2017).

[17]. Gallagher, Marcus, Ian Wood, Jonathan Keith, and George Sofronov. "Bayesian inference in estimation of distribution algorithms." In Evolutionary Computation, 2007. CEC 2007. IEEE Congress on, pp. 127-133. IEEE, 2007.

[18]. Garcia, Patricio, Analia Amandi, Silvia Schiaffino, and Marcelo Campo. "Evaluating Bayesian networks' precision for detecting students' learning styles." Computers & Education 49, no. 3 (2007): 794-808.

[19]. Hullermeier, Eyke. "Toward a probabilistic formalization of case-based inference." In IJCAI, pp. 248-253. 1999.

[20]. Kulatilake, Pinnaduwa HSW, Tien H. Wu, and Deepa N. Wathugala. "Probabilistic modelling of joint orientation." International Journal for Numerical and Analytical Methods in Geomechanics 14, no. 5 (1990): 325-350.

[21]. Diard, Julien, Pierre Bessiere, and Emmanuel Mazer. "A survey of probabilistic models using the bayesian programming methodology as a unifying framework." (2003).

[22]. Li, Jian, Shaogang Gong, and Tao Xiang. "Global Behaviour Inference using Probabilistic Latent Semantic Analysis." In BMVC, vol. 3231, p. 3232. 2008.

[23]. Kosgodagan, Alex, et al. "A 2-dimension dynamic Bayesian network for large-scale degradation modelling with an application to a bridges network." (2017).

[24]. Jaeger, Manfred. "On the complexity of inference about probabilistic relational models." Artificial Intelligence 117, no. 2 (2000): 297-308.

[25]. Beaumont, Paul, and Michael Huth. "Constrained Bayesian Networks: Theory, Optimization, and Applications." arXiv preprint arXiv:1705.05326(2017).

[26]. Weber, Philippe, and Lionel Jouffe. "Reliability modelling with dynamic bayesian networks." IFAC Proceedings Volumes 36, no. 5 (2003): 57-62.

[27]. Hunter, Anthony. "A probabilistic approach to modelling uncertain logical arguments." International Journal of Approximate Reasoning 54, no. 1 (2013):47-81.

[28]. Ng, Kee Siong, John W. Lloyd, and William TB Uther. "Probabilistic modelling, inference and learning using logical theories." Annals of Mathematics and Artificial Intelligence 54, no. 1-3 (2008): 159-205.

[29]. Chater, Nick, Joshua B. Tenenbaum, and Alan Yuille. "Probabilistic models of cognition: Conceptual foundations." Trends in cognitive sciences 10, no. 7 (2006): 287-291.

[30]. Oakley, Jeremy E., and Anthony O'Hagan. "Probabilistic sensitivity analysis of complex models: a Bayesian approach." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66, no. 3 (2004): 751-769.

[31]. Riezler, Stefan. "Statistical inference and probabilistic modelling for constraint-based nlp." arXiv preprint cs/9905010 (1999).

[32]. Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.

[33]. Nilsson, Mikael, and Marcus Ejnarsson. "Speech recognition using hidden markov model." (2002).

[34]. Nguyen, Nam Thanh, Dinh Q. Phung, Svetha Venkatesh, and Hung Bui. "Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model." In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 2, pp. 955-960. IEEE, 2005.

[35]. Sha, Fei, and Lawrence K. Saul. "Large margin hidden Markov models for automatic speech recognition." In Advances in neural information processing systems, pp. 1249-1256. 2007.

[36]. Bunke, H. and Caelli, T. eds., 2001. Hidden Markov models: applications in computer vision (Vol. 45). World Scientific.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Implementation of Neuro-Fuzzy Decision Tree Based Malignant Tumor Detection System

Sanjeev Kumar
Sam Higginbottom University of Agriculture, Technology and Sciences,Allahabad, ABESIT, Allahabad, Uttar Pradesh, India

ABSTRACT

The paper designs a technique to classify the tumor as malignant or benign. The designed system works on various attributes of tumor like tumor thickness, shape, size etc. The classification process completes in three phases; the phase 1 classifies the attributes as cat1 or cat2 on the basis of information gain. Then in phase 2 cat1 attributes are used to select the class of tumor by using the RBF neural network while the cat2 attributes uses the fuzzy to select the class of tumor. The results of both techniques are collaborated by using the fuzzy inference system in the phase 3. The effectiveness of the technique is easily identified by using results. Finally Comparing accuracy between Neuro Fuzzy system and decision tree.

Keywords: Breast Cancer, Malignant, Benign, Tumor, RBF, Fuzzy, Decision Tree

I. INTRODUCTION

Breast cancer develop from breast cell. Breast lumps and abnormal mammogram are first sign of breast cancer. Some of symptoms of breast cancers are change in contour ,size, texture ,temperature of breast .Any change in size of nipple like dimpling ,itching ,nipple retraction and unusual discharge of blood are sign of breast cancer. Rate of breast cancer is much higher in developed country than Developing country. Researcher assume that life style and eating habits are one of the reason.In USA ,232340 female and 2240 male are suffered from breast cancer per year as said by NATIONAL CANCER INSTITUTE of which near about 40000 died. In U.K women suffering from breast cancer increased by 6% over last decade 1999 to 2001 and 2008 to 2010[1].It has been observed that around 550,000 to 570,000 people suffering from breast cancer in U.K[2].A women has 12.5% risk of breast cancer during her life time in U.S.A[3].Breast cancer is one of important factor of cancer death among women. Mammography is one of the most Effective technique for identifying breast cancer. since quality of image by mammography technique is poor and noisy .hence it is difficult to interpret mammography image. Calcification and masses are important abnormalities in breast cancer. Calcification are small mineral deposit with in breast tissue which appear white spot on film. Calcification are of two type macro calcification and micro calcification. The presence of micro calcification is primary sign of breast cancer. Micro calcification are tiny deposit of calcium whose general size ranges from .1mm to 1mm. and average size is .3mm.when three or more deposit of calcium come together it create a micro calcification cluster. Micro calcification is high frequency component and high frequency noise with lower frequency background. Correct segmentation of each micro calcification is challenged by micro calcification shape and size variability superimposed surrounding tissue with high frequency noise [4].

According to radiological view of micro calcification cluster, an area 1cm^2 contain no fewer than three micro calcification [5]. Mammography enables detection of micro calcification at early stage since spatial resolution of mammography is very high. Only certain kind of micro calcification cluster are cancer with a high probability of malignancy [6][7].There are two types of tumor in breast called malignant and benign. only malignant tumor is cancerous. The malignant and benign tumors are classified by using a grey level based method segmentation of individual micro calcification can obtain such as region growing[9]and grey level threshold in region of interest ROI[10].There are many standard technique used for parameter free segmentation radial gradient based method[11],watershed algorithm[12], morphologic operation[13],Bayesian pixel classification and markov random field model[14].Again wavelet transform method is used for segmentation of micro calcification due to detecting power of spatial localize high frequency component[15].A multi scale active ray for detecting segmentation due to micocalcification variability[16].There are a no. of CAD (Computer added diagnostic) which incorporates expert knowledge to radiologist for detection of micro calcification, but accurate interpretation of micro calcification is still difficult. Monika shinde [17] develop a new method for classification of mass and normal tissue. Expected Maximization method separate mass tissue from normal breast tissue by using digitized mammogram. Law texture analysis was done by using raw data and summary data. By using EM method 63% sensitivity and 89% specificity is achieved. In this method 300 image are analyzed of which 203 are benign and 97 are malignant. Renato campanini et.al.[18] proposed a idea for mass detection by using digital mammogram. They were not extracted any feature for detection of ROI(region of interest) and exploited all information available on the image. Detection task is performed as two class pattern recognition .According to them sensitivity of system was 80% with a FPR(false positive rate)1.1marks per image. Classification of micro calcification Cluster explain feature selection and then classify by using suitable classifier until now no any standard feature set that accurately classify more than 90% .

RBF:-Radial basis function network is used mainly for pattern classification. Radial basis network uses radial basis function as activation function. Radial basis function is a classification and functional approximation neural network. Network uses nonlinearities function such as sigmoid and Gaussian function. The structure consist of three layers: The input layer is made up of source node that connect network to its environment. The second layer consist of hidden unit applies a nonlinear transformation from the input space to hidden (feature) space. In most application dimensionality of only hidden layer is high. This layer is trained in unsupervised manner. The output layer is linear ,designed to supply the response of the network to activation pattern applied to input layer. This layer is trained in supervised manner. The response of such is positive for all value of y. The response decrease to zero as |y|->0 $f(y)=e^{-y^2}$

Derivative of above function is $f'(y)= -2ye^{-y^2}= -2yf(y)$. Graphical representation is given in figure 1.

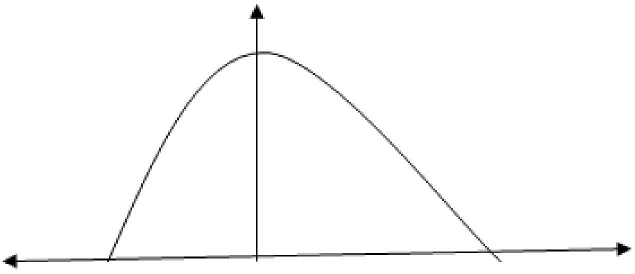


Figure 1 . Gaussian kernel function.

Each node is found to produce identical output for input within fixed radial distance from the centre of kernel when Gaussian potential function is used.

Hence they are radically symmetric and name is radial basis function network. Whole network form a linear combination of nonlinear function. Architecture of RBF is given in figure 2.

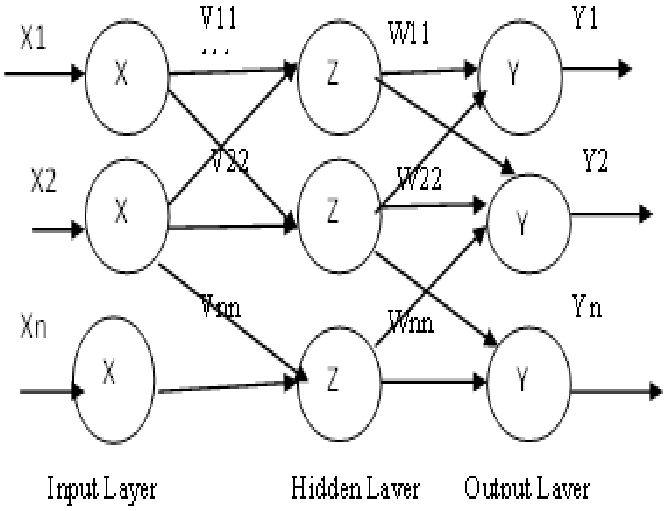


Figure 2. Architecture of RBF

II. PROPOSED SYSTEM

The existing work uses the RBF to classify the tumor as benign or malignant. The main drawback of the RBF NN is the high false positive rate. Moreover, if the feature space is large then the RBF network is able to select the important input variables. The present system has selected the important input variables in the first step. Then the RBF NN is used to classify the tumor by using these important variables. The remaining input elements are used as input to fuzzy system to classify the tumor as the malignant or benign. The output of both system is given as input to fuzzy to get the final output of the network. The following overall architecture explains the process:

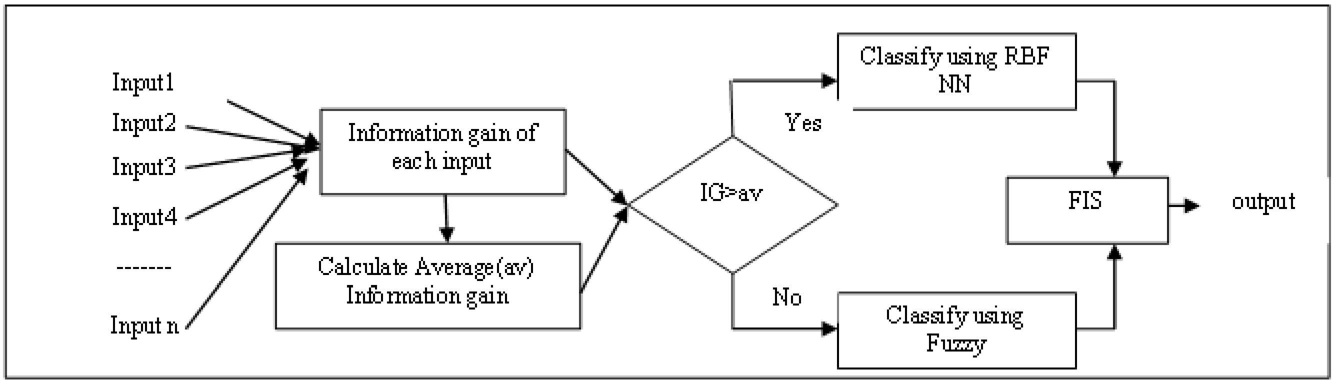


Figure 3. Block diagram for breast cancer tumor classification

The figure 3 gives the brief of the present system. The whole working of the present system can be classified in three phases namely information gain, classification, decision. All three phases are explained below:

Information Gain (Phase 1): In this phase the input features (feature present in the dataset) are used to evaluate the information gain then the category of each input instance is decided based on information gain. Information gain is the amount of information gained due to the given attribute to predict the class of sample. In other words, it is the amount of information available when the only feature is used to predict the class. The larger the gain, lower is the entropy i.e. uncertainty. It

can be given as:

$$IG(D,a) = H(Da) - \sum_{j=values\ of\ D}^{|Dj=v|/|D|} H(Dj = v) \qquad (1)$$

Where H(Da) denotes the entropy of the dataset D. The a is attribute selected and v is value of the attribute & $|Dj = v|/|D|$ is the fraction of dataset having values v.

The

$$H(Da)=-p\pm(D)\log_2 p\pm(D) \quad (2).$$

Here p±D is the probability of same and different classes in the given set. The average of the information gain calculated using equation (1) for each input is calculated by using equation (3)

$$av = 1/n * \sum_{a=1}^{n} IG(D,a) \qquad (3)$$

Each input is classified in two categories i.e. cat1 and cat2 by using the av value. The input having the information gain greater than the av value is categories to cat1 otherwise as cat2 also shown by equation (4).

$$if\ IG(D,a) > av\ then\ a = cat1\ else\ a = cat2 \qquad (4)$$

This category of input instance is given as input to the phase 2. If the cat2 attributes are

Classification (Phase 2): This phase is used to classify the input tumor sample as benign or malignant by using the fuzzy and the neural network. This phase works in two sub phases, one for the cat1 and other for the cat2. The input instances belong to cat1 are classified by using the RBF neural network while the cat2 instances are classified by using the fuzzy.

Classification using RBF neural network

The details of RBF are already given in the section 2. The RBF neural network used in this system is trained against the Wisconsin breast cancer data set downloaded from the UCI repository[]. This trained network is used for the classification purpose. The radial basis function is used as the activation function and the output is obtained by using the equation (5).

$$y = f(\textstyle\sum_{i=1}^{n} xi * wi) \qquad (5)$$

Where the xi represents the ith input attribute and the wi is the weight obtained using the training phase. f(.) shows the activation function. The output y is either 1 or 0 representing the malignant and benign respectively.

Classification using the fuzzy

All the input attributes of category cat2 are given as the input to the fuzzy system. The name of attribute can be different but the proper fuzzy system the crisp range, linguistic range and the mapping of the variables is taken same for each variable. The crisp range of each attribute is 1-10 and the linguistic range is L,M,H. The membership function for mapping is shown in figure 4.

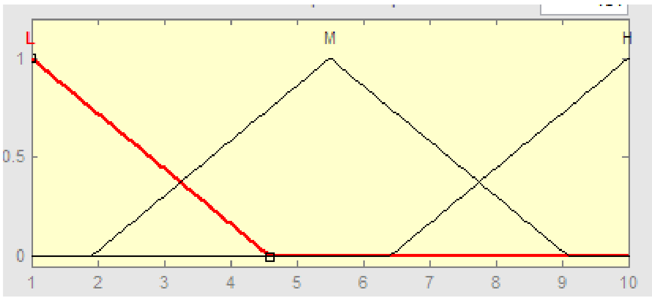


Figure 4. Membership function for mapping.

The output of the fuzzy system will decide the class of the tumor sample.

Decision(Phase 3) This phase collaborates the classification output of the cat1 and cat2 by using a fuzzy inference system. Various type of FIS system exists, Madami fuzzy system is used shown in figure 1.

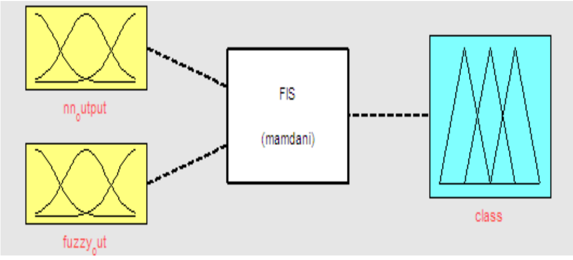
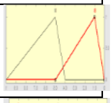
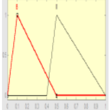
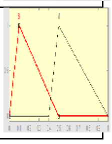


Figure 5. Fuzzy System

The FIS used for the decision making takes two input variable i.e. NN output and fuzzy output and provides one output i.e. class of tumor. The detail of variable is given in the table 1.

Table 1. FIS System

Sr. no.	**Attribute Name**	**Variable Type**	**Crisp Range**	**Linguistic Range**	**Membership function**
1	Nn_output	Input	0-1	B,M	
2	Fuzzy_output	Input	0-1	B,M	
3	Class	Output	0-1	B,M	

The FIS system has 4 rules to decide the class given in the figure 2:

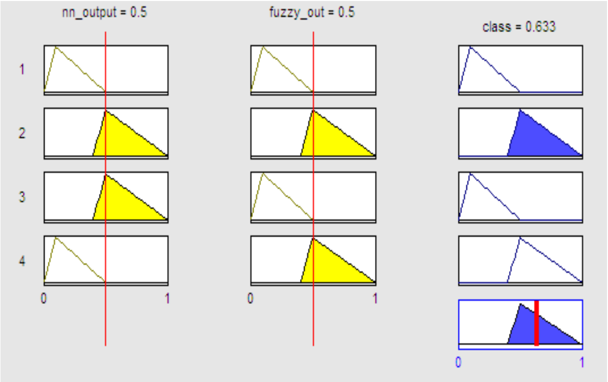


Figure 6. Fuzzy Rules

The system will decide the output class of the tumor on the basis of this system. The present system is used to classify the brain tumor sample as the benign or malignant. The implementation of the system is discussed in the next section.

Implementation

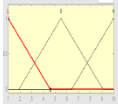
The system described in above section is implemented using the MATLAB. The script editor along with the fuzzy toolbox is used for the implementation of the system. The system is implemented on the Breast Cancer Wisconsin (Diagnostic) Data Set downloaded from UCI repository[]. The dataset has 11 attributes given in the table 2.

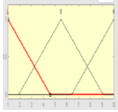
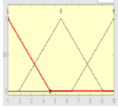
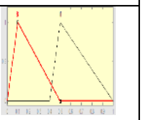
Table 2

Sr. No	**Attribute Name**	**Range**
1	Clump Thickness	1-10
2	Single Epithelial Cell Size	1-10
3	Bland Chromatin	1-10
4	Bare Nuclei	1-10
5	Normal Nucleoli	1-10
6	Marginal Adhesion	1-10
7	Uniformity of Cell Size	1-10
8	Mitoses	1-10
9	Uniformity of Cell Shape	1-10
10	ID number	Any number
11	Class	0-1

These attributes are given as input to the present system. The phase 1 classify the attributes in cat1 and cat2 based on information gain. The cat1 attributes are Marginal Adhesion, Clump Thickness, Mitoses, Bare Nuclei, Bland Chromatin, Uniformity of Cell Size while the remaining are the cat2 attributes. Then in the phase 2 classify the tumor. The RBF NN classify by using the cat1 attributes and the result tumor as M. While the input detail to the fuzzy system for classification is as follow:

Table 3

Sr. no.	**Attribute Name**	**Variable Type**	**Crisp Range**	**Linguistic Range**	**Membership function**
1	Uniformity of Cell	Input	1-10	L,M,H	

	Shape				
2	Single Epithelial Cell Size	Input	1-10	L,M,H	
3	Normal Nucleoli	Input	1-10	L,M,H	
4	Class	output	0-1	B,M	

The fuzzy system uses the 27 rules to decide the class and the resulting class is B. Then the output of the both system is given to phase 3 i.e. decision phase. The result of decision phase is B as the input to this phase is M and the B. Overall the tumor is found to be of benign. The processed is carried over all the 699 instance of the dataset and the result is discussed in the next section.

III. RESULT AND DISCUSSION

The implementation of above system is done and the results are compared by using various parameters shown in the table 4 and figure 7 and 8.

Table 4. Comparison of Various algorithms using various parameters

Algorithm	Classification accuracy	TP rate	FP rate	Precision
RBF	96.87	0.967	0.033	0.967
HRBF	97.09	0.972	0.028	0.972
Proposed	98.02	0.980	0.020	0.980

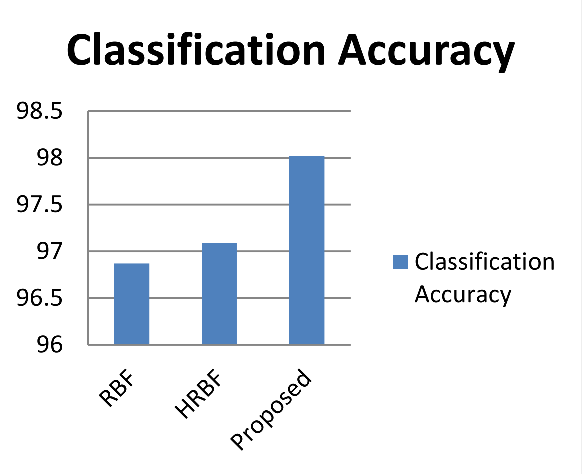


Figure 7. Accuracy Comparison of various Algorithms RBF, HRBF and proposed algorithm

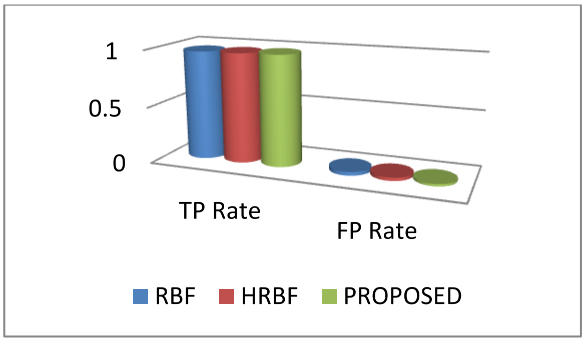


Figure 8. Comparison of proposed ,HRBF and RBF algorithm

The figure 7 and fig 8 compares the accuracy, TP rate and the FP rate respectively. The enhancement in the accuracy and TP rate with the degradation in the FP rate signifies that the cancer can be effectively identified by using the proposed technique.

Decision tree implementation:-

Decision Trees require least tuning, and does automatic feature selection. And it can also deal with noisy or incomplete data.. But it has high classification error rate for more number of classes. And for large dataset calculation growth is exponential. In our dataset on analyzing decision tree we see that nearly 75% datasets are classified as malignant or benign on basis of worst perimeter. Without pruning dataset the tree over fits, thus we apply pre pruning for more accurate results. And it yields better results at maximum depth = 2.

Normal - 0.923076923077@ maximum depth = 5

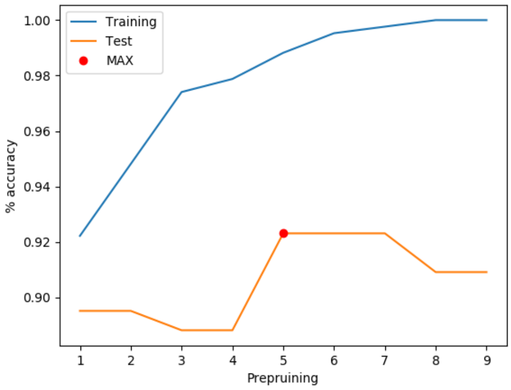


Figure 9. Decision tree for accuracy

IV. CONCLUSION

This paper presents a system to detect the breast cancer by using the neural and fuzzy. The system is implemented using the MATLAB and analyzed against database downloaded from the UCI repository. The approximate 1% increase in the accuracy is identified by using the described system. Moreover, the system also decreases the FP rate. In future , the meta-heuristics techniques can be used to enhance the accuracy to 100%. Accuracy by Neuro Fuzzy system is more than Decision tree

V. REFERENCES

[1]. Ferlay, J., Shin, H. R., Bray, F., Forman, D., Mathers, C., & Parkin, D. M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008.International journal of cancer, 127(12), 2893-2917.

[2]. Helvie, M. A., Chang, J. T., Hendrick, R. E., & Banerjee, M. (2014). Reduction in late- stage breast cancer incidence in the mammography era: Implications for overdiagnosis of invasive cancer. Cancer, 120(17), 2649-2656.

[3]. Siegel, R., DeSantis, C., Virgo, K., Stein, K., Mariotto, A., Smith, T., ... & Ward, E. (2012). Cancer treatment and survivorship statistics, 2012. CA: a cancer journal for clinicians, 62(4), 220-241.

[4]. Cheng, H. D., Cai, X., Chen, X., Hu, L., & Lou, X. (2003). Computer-aided detection and classification of microcalcifications in mammograms: a survey.Pattern recognition, 36(12), 2967-2991.

[5]. Ma, Y., Tay, P. C., Adams, R. D., & Zhang, J. Z. (2010, September). A novel shape feature to classify microcalcifications. In Image Processing (ICIP), 2010 17th IEEE International Conference on (pp. 2265-2268). IEEE.

[6]. Sickles, E. A. (1986). Breast calcifications: mammographic evaluation.Radiology, 160(2), 289-293.

[7]. Arnold, R., Langer, P., Rothmund, M., Klöppel, G., Kann, P. H., Heverhagen, J. T., ... & Stinner, B. (2013). Endokrine Tumoren des gastroenteropankreatischen Systems. In Praxis der Viszeralchirurgie (pp. 497-628). Springer Berlin Heidelberg.

[8]. Islam, M. J., Ahmadi, M., & Sid-Ahmed, M. A. (2010). An efficient automatic mass classification method in digitized mammograms using artificial neural network. arXiv preprint arXiv:1007.5129.

[9]. Chan, H. P., Sahiner, B., Lam, K. L., Petrick, N., Helvie, M. A., Goodsitt, M. M., & Adler, D. D. (1998). Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces. Medical Physics, 25(10), 2007-2019.

[10]. I. Leichter, R. Lederman, S. Buchbinder, P. Bamberger, B.Novak and S. Fields, "Optimizing parameters for computeraided diagnosis of microcalcifications at mammography",

[11]. Cheng, H. D., Cai, X., Chen, X., Hu, L., & Lou, X. (2003). Computer-aided detection and classification of microcalcifications in

mammograms: a survey. Pattern recognition, 36(12), 2967-2991.

[12]. Moradmand, H., Setayeshi, S., & Targhi, H. K. (2012). Comparing methods for segmentation of microcalcification clusters in digitized mammograms. arXiv preprint arXiv:1201.5938.

[13]. Ganesan, K., Acharya, U., Chua, C. K., Min, L. C., Abraham, K. T., & Ng, K. B. (2013). Computer-aided breast cancer detection using mammograms: a review. Biomedical Engineering, IEEE Reviews in, 6, 77-98.

[14]. Paquerault, S., Yarusso, L. M., Papaioannou, J., Jiang, Y., & Nishikawa, R. M. (2004). Radial gradient-based segmentation of mammographic microcalcifications: observer evaluation and effect on CAD performance. Medical physics, 31(9), 2648-2657

[15]. Heinlein, P., Drexl, J., & Schneider, W. (2003). Integrated wavelets for enhancement of microcalcifications in digital mammography. Medical Imaging, IEEE Transactions on, 22(3), 402-413.

[16]. Unser, M., Aldroubi, A., & Laine, A. (2003). Guest editorial: wavelets in medical imaging. IEEE Transactions on Medical Imaging, 22(LIB-ARTICLE-2003-004), 285-288.

[17]. Shinde, M. (2003). Computer aided diagnosis in digital mammography: Classification of mass and normal tissue (Doctoral dissertation, University of South Florida).

[18]. Renato Campanini, Danilo Dongiovanni, Allessandro Riccardi, “A Novel Featureless Approach to Mass Detections in Digital Mammograms based on Support Vector Machines”, Department of Physics, University of Bologna, Italy, 2003.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Problems and Testing the Assumptions of Linear Regression: a Machine Learning Perspective

Ayoosh Kathuria
Sam Higginbottom University of Agriculture, Technology and Sciences, Allahabad, ABESIT, Allahabad, Uttar Pradesh, India

ABSTRACT

Linear Regression is perhaps one of most well-known algorithms in statistics and Machine Learning. Despite its widespread use in machine learning applications, the importance of testing the assumptions of linear regression is often trivialised in machine learning literature. However, the predictions of linear regressions cannot be trusted unless its assumptions are met. An attempt has been made to attract the attention of the community towards this understated aspect of putting linear regression into practice. This paper serves as an endeavour to shed some light on ways to test the assumptions of linear regressions and how to remedy the violations if there are any.

Keywords : Time Series Prediction, Regression Analysis, Linear Regression, Machine Learning

I. INTRODUCTION

Linear Regression is often the very first algorithm to be taught in any machine learning curriculum. The algorithm, which is borrowed from statistics models the variable we are trying to predict (called the target variable) as a weighted linear combination of one or more inputs variables. Despite its simplicity, it has been deployed across a vast array of real life problems including forecasting stock market trends [1], weather forecasting [2], analysis of automobile engine performance [3], optimising targeted advertising [4] to name a few. The vanilla version of algorithm works by minimising the least squares function, and the performance is judged by the value of the Pearsons coefficient of correlation [5], also referred to as R-squared value (Adjusted R-square for model having multiple input variables. In this paper, whenever we come across the term R-squared, it is to be understood we mean adjusted R-squared in case of multiple input variables) However, most of the machine learning practitioners often focus on squeezing as much predictive power as they can out of a model, and are often less concerned about the explanatory power of the features used as input. It is also to be noted that the predictions of the model can be trusted only if the assumptions of algorithm are satisfied. In case they are violated, the evaluation metrics, however stellar they might be, are no guarantee for the effectiveness of the model. Hence, it is absolutely fundamental that these assumptions should be rigorously tested while evaluating the performance of Linear Regression. Though these techniques are well documented in statistics literature [6], their coverage in machine learning literature leaves much to be desired.

The object of this paper is to attract the attention of the community over emphasising the need to test these assumptions, and consequently incorporating methods to fix the violations, if any. This paper is divided into five parts. First, we start by laying a theoretical framework of Linear Regression that helps the reader appreciate why the assumptions are so crucial to the assessment of the model. Second, we perform ordinary least squares regression on Google Stock data from the past ten years. Third, we explore ways to test for the violations of the assumptions and how to fix them. Fourth, we will apply these techniques to fix the violations made by our model and contrast the performance of the new models with the previous one. Finally, we put forth our concluding remarks, and explore the options of using methods other than Linear Regression (LR).

II. THEORETICAL FRAMEWORK

Given training examples (X, Y), LR tries to establish a linear relationship between the inputs $x^{(i)} \in X$ and their corresponding labels (value of target variable) $y^{(i)} \in Y$ such that

$$y(i) = \theta^T x(i) + \varepsilon(i) \qquad (1)$$

where θ is the weights vector that parametrises the model, and $\varepsilon^{(i)}$ is the error term that captures either the unmodelled effects (such as features pertinent to predicting target variable that we failed to include in our model), or random noise.

Assumptions of Linear Regression

1) There exists a linear relationship between target variable and each of the input variables. The weights, $\theta_i{}^0$s $\in \theta$ associated with each variable are independent of each other. The effects each the input variable has on the target variable are additive in nature.
2) The errors $\varepsilon^{(i)}$s exhibit homoscedasticity or constant variance against time, the target variable as well as the input variables.
3) The errors, $\varepsilon^{(i)}$s are independent and identically distributed (IID).
4) The errors $\varepsilon^{(i)}$s are normally distributed with mean zero and variance σ^2.

Loss Function

Using assumption (4), we can write the probability density of $\varepsilon^{(i)}$ as:

$$p(\varepsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\varepsilon^{(i)})^2}{2\sigma^2}\right) \qquad (2)$$

This implies that

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \qquad (3)$$

The above equation represents the probability density of an arbitrary $y^{(i)}$ given a $x^{(i)}$ and is parametrised by θ. This function, called the likelihood function, depends on θ. We then move to choose a value of θ that maximises the value of this function. This will give us a model that agrees best to our training data.

The joint probability density for the training set can be simply written as the product of probability densities of each training examples, as we have assumed them to be IID (assumption 2). Since we have assumed errors exhibit homoscedasticity, all of them share a common variance σ^2

The likelihood function for the training set could then be written as

$$L(\theta) = \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta) \qquad (4)$$

$$L(\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\varepsilon^{(i)})^2}{2\sigma^2}\right) \qquad (5)$$

Taking log on both sides of (5) and simplifying it can be shown that maximising likelihood is same as minimising the equation.

$$\frac{1}{2}\sum_{i=1}^{m}(y^{(i)} - \theta^T x^{(i)})^2 \qquad (6)$$

This is the standard least squares loss function we minimise during training in LR. The objective of the above treatment was to show how assumptions of LR factor into derivation of the loss function.

Predicting Google's stock price

LRs is often used for predictive analysis of trends in stock markets. We have picked a simple problem where we will use LR for predicting the stock price of Google using the data from the past 10 years. Our goal is to predict the closing price of the stock. The data has been acquired from Quandl, an online platform that hosts financial data. The input features include opening price (Open), highest price (High), lowest price (Low), Closing Price (Close), Volume Traded (Volume), Ex-Dividend, Split Ratio, Adj. Open, Adj. High, Adj. Low, Adj. Close, Adj. Volume collected 10 days prior to the day for which the prediction has to be made. The adjusted ones account for stock splits (One stock becomes two, and the value of each stock is halved), whereas the regular ones do not, so we are going to drop those features. We’ll now refer to the Adjusted features without the ”Adjusted” prefix.

Feature selection

We are going to limit the feature selection process to dropping all but one amongst the sets of highly correlated features, though feature selection forms one of the most critical parts of the model training process, we are not going to dwell too much on it here for sake of focusing on the main topic of the paper.

Table 1. Correlation Matrix For Input Features

	Open	High	Low	Close	Volume
Open	1.0000	0.9999	0.9998	0.9997	-0.5599
High	0.9999	1.0000	0.9998	0.9999	-0.5583
Low	0.9998	0.9998	1.0000	0.9999	-0.5630
Close	0.9997	0.9999	0.9999	1.0000	-0.5608
Volume	-0.5599	-0.5583	-0.5630	-0.5608	1.0000

By looking at Table 1 we can clearly see Open, High, Low, Close are highly correlated. Let us drop all of them but Close variable while training our model. We have also dropped Volume to keep things simple in accordance to the principle of Occam’s Razor. We will test another model later which has Volume later in the paper.

Model Evaluation

Table 2. Model Evaluation Metrics

	weight	Std err	t-statistic	P-value	95.0% Conf. Int.
Intercept	4.8595	1.292	3.761	0.000	[2.326 7.393]
Close	0.9641	0.003	294.316	0.000	[0.958 - 0.971]

Adjusted R-squared: 0.966

This model achieves a Adj. R-squared of 0.966. A lot of ML practitioners may take it as a conclusive proof of the high efficiency of the model. However, we could also look at the 95.0% Conf. Int or the confidence Intervals for the weights. These are the values which a particular weight may take about 95 out of every 100 times we randomly sample the data from a population and fit the model to it. One should be alarmed if zero also falls in the confidence interval of a weight. This could suggest that there’s a considerable probability that the value of the weight is zero, which implies there’s no relationship between the feature and our target. Another metric that can tell us about such a problem is the p-value. It basically measures the likelihood of our data, given the null hypothesis that the weight is equal to zero. In other words, it measures whether the relationship we observe is merely a statistical fluke. born out of a sampling error

Since we're using 95% confidence intervals, the p-value for all the coefficients must be less than 0.05 [7]. We could also look at the T-statistic, which measures the number of standard deviations a weight distribution's mean is away from zero. Typically, for 95% confidence intervals, T-value should be more than 2 in magnitude. The current model fulfils all the above criteria. High value of R-square, all the p-values below 0.05, and all the t-values above 2.

At this point the reader might be edging towards concluding the model does a very good job generalising to the dataset. However, we now proceed to check whether the assumptions of LR are violated or not.

Testing the assumptions of LR

Assumption of Linearity

The very premise of testing this assumption can be a tricky bargain. Assuming that there is indeed a linear relationship between the target and the input variable forms the core of our belief that LR is a suitable choice to solve the problem at hand. If we are willing to test this assumption, it means we are ready to consider the case where a linear relationship might not hold, which in turn renders this complete analysis redundant. One might even think that this assumption is merely a leap of faith. However, in practical cases, such assumptions are often backed by domain expertise and guiding principles like the Occams razor.

In fact, a lot of ML practitioners believe a good measure of whether this assumption holds is the R-squared itself. However, R-squared is merely the percentage of the target variable variation that is explained by the straight line we have fit. It only describes how well are the input and the target variables correlated. It does not confirm a causal relationship between the target variable and the input variables. Thus, R-squared cannot be solely use to establish our assumption. R-squared simply tells us the quality of the linear relationship, assuming a linear relationship does exist.

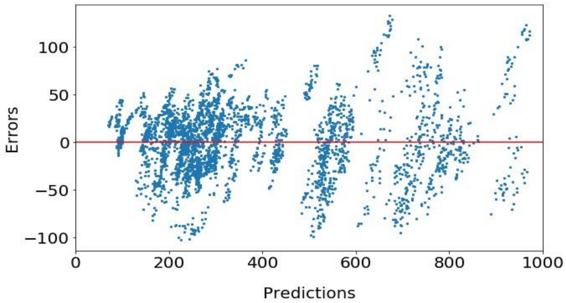


Figure 1. Errors v/s Predicted Values plot for the Google Stock price Model with only Close as the input variable

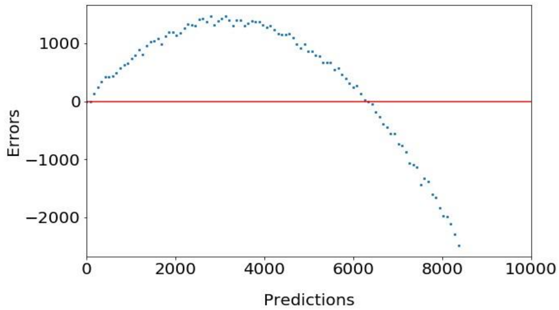


Figure 2. Errors v/s Predicted Values plot for a LR model with only x as the input variable. The dataset has a non-linear (quadratic) relationship between the inputs and the labels, described by y=x^2+10x

How to diagnose: Non-linearity can be detected by plotting errors vs the predicted values of the target variable. Figure 1 shows the error v/s predicted values plot for our model.

In Figure 1, we see the errors roughly have a zero mean, and are randomly distributed around the mean. This makes a strong case for the assumption of linearity. The reader may recall that error term is attributed to unmodelled effects, as well as random noise. Had we tried to model a non-linear relationship using LR, the non-linear effects of input variables would have showed up in error term. In such a case, errors would have been systematic in nature.

To get a better insight let us use LR on a dataset having a non-linear (quadratic) relationship between the input and the target variable, described by y = x^2 + 10x. Figure 2 shows

Figure 3. Predicted values v/s errors for a LR model with x and x^2 as the input variables. The dataset has a non-linear (quadratic) relationship between the inputs and the labels, described by y=x^2+10x

what the Errors v/s Predicted Values plot looks like when we fit a LR model to the dataset.

Remedy: The very first thing one can do is try to apply a non-linear transformation to one or more variables to linearise the relationship. For example, if the target variable is an exponential function of the inputs, applying log transformation to the input variables will linearise the relationship. If a small percentage changes in one or more input variables induces a proportionate percentage change in value of the target variable, the relationship between the inputs and the target variable is a multiplicative one. In such case, a log transformation may be applied to a both the input and the target variables.

One can also try to add another input variable which is simply a non-linear transformation of one of the input variables used in the model. However, such methods could often lead to overfitting, and reguarisation must be used appropriately. One can also come up with a new variable that is a combination (for example, product) of two or more input features used in the model. The cusp of engineering a new input variable to to account for any unmodelled effects.

Assumption of homoscedasticity

This assumption can be tested by looking at the plots of errors vs the predicted value of the target variable, as well as the errors v/s time plot, shown in figure 4 in case of a time series data. By looking at figure 1, we can easily conclude that this assumption is violated as the errors do not have a constant variance across different values of the predicted variable. In particular, the variance seems to increase as the predicted value of the target variable increases.

Again, the violation of this assumption is very evident as we observe the errors do not exhibit a uniform variance.

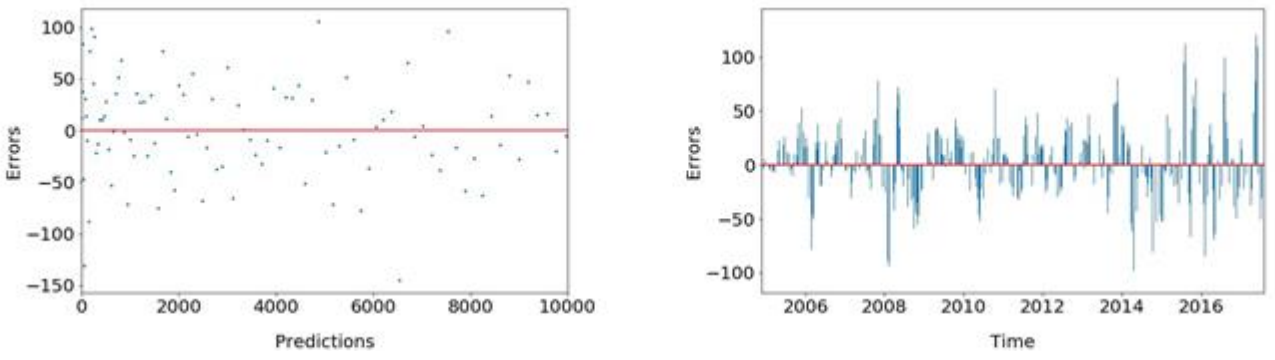


Figure 4. Errors v/s Time plot for the Google Stock Price Model with only Close as the input variable

Remedy: If the target variable can take only positive values and variance of the errors increases, probably proportionately, as the predicted value of the target variable increases, applying a log transformation to the target variable may stabilise the variance of the errors. Such a transformation helps because such sort of errors are consistent in terms of percentage growth, rather than absolute terms. Heteroscedasticity can also arise owing to violations of the assumptions of linearity and/or independence, in which case it may be fixed as a consequence of fixing those problems.

Table 3. Model Evaluation Metrics

Lag	1	2	3
Autocorrelation	0.973	0.946	0.920

In case of time-series data, one may also note a periodic trend in the variances of errors. The variance of errors maybe roughly uniform for periodic intervals. Such a problem may be solved by introducing an additional variable in our model that accounts for seasonal patterns. It maybe also the case the we deal with larger values for some of our input variables in some particular part of the season resulting in errors of larger magnitude. In that case too, applying a log transformation to target variable can help solve the issue.

Assumptions of Independence

This assumption can be tested by the use of a errors v/s time plot, shown in figure 4. This assumption is clearly violated in plot shown in figure 4. We conclude this by observing that positive errors are followed by positive errors, and negative errors are followed by negative ones for long intervals. This idea can be captured more formally by a mathematical quantity called autocorrelation.

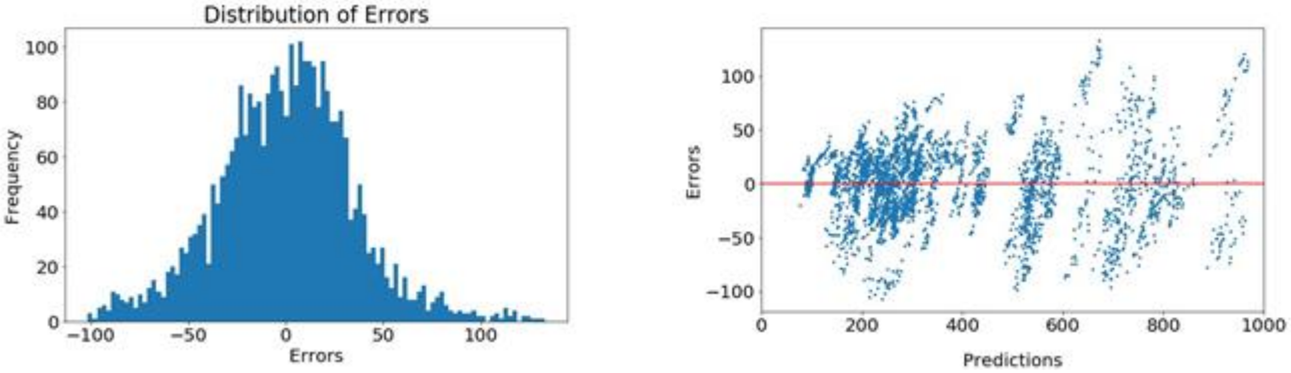


Figure 5. Error histogram for the Google Stock Price Model with only Close as the input variable

Autocorrelation is basically the serial correlation between the errors separated by a fixed amount of time interval (called the lag). The autocorrelations for most lags should fall between $+/-\left(\frac{2}{\sqrt{n}}\right)$, where n is the size of the training set. (0.035 for our model). The autocorrelation for errors of our model are given in the table are shown in table 3. This assumption of LR is clearly violated as autocorrelations of our model are away above the threshold that must be adhered to.

Remedy: Mild cases of autocorrelation maybe addressed by adding a time-lagged version of either the target or one of the input variables. If there's significant autocorrelation at the lag n, one can use a variable lagged by n time intervals to address the issue. There might be seasonal autocorrelation in time series data, wherein errors belonging to the same season may be correlated. A seasonally lagged variable can be added to the model to address this issue.

Assumption of Normality

This assumption can be simply tested by plotting a histogram of errors. The histogram of errors of our models are shown in figure 5 The reader can see the distribution is not perfectly normal, and seems a bit negatively skewed. Violations of this assumptions arise to due to non-linearity, or the presence of outliers.

Remedy: Most of the techniques that remedy non-linearity remedy the violation of this assumption too. A non-linear transformation of variables is often sought as the cure to this problem. As far as the question of outliers go, one must ponder over the question of keeping them in the training dataset or not. To resolve that issue, we must ask ourselves whether they denote merely a statistical fluke or do they represent rare phenomenon which could repeat itself in future.

Figure 6. Predicted v/s Error plot for model having Close and Volume as input variables

5. Fixing violations of assumptions in Google Stock Data Model

Our model fares well on the assumption of linearity as we observe the errors are randomly distributed about the zero mean line in figure 1. However, we had omitted the Volume input variable in our model. We could try a model having Volume as an additional input variable to see if we can get better results on the linearity front. The errors v/s predictions graph is plotted in figure 6. We see no considerable improvements and hence, we stick with our earlier model in spirit of keeping the model simple.

As seen in figure 1 and 4, the assumption of homoscedasticity is clearly violated. In figure 1, we see that the variance of errors increases as the value of the stock price increases. It maybe also noted a similar trend is noted in figure 4, where the variance grows as we progress through time (It can be observed that the stock price has risen as we proceed through time too). Such a violation suggests errors are consistent in percentage rather than absolute value. As suggested, earlier we apply a log transformation to our target variable. Figure 7 shows the errors v/s time plot,

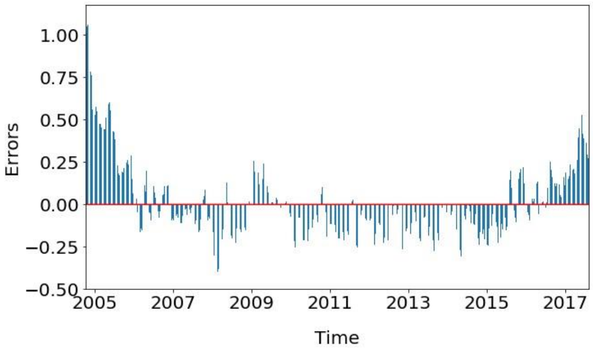


Figure 7. Error vs Time plot for model with Close as input and logged target variable

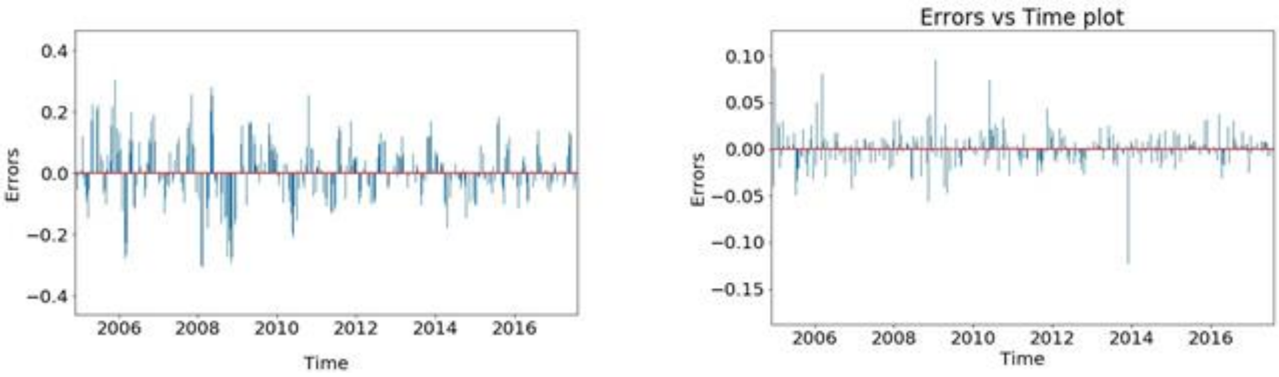


Figure 8. Error vs Time plot for model with logged Close as input and logged target variable

which shows improved homoscedasticity of errors in general except a bunch errors of higher variance in the beginning.

Interestingly, it’s a practical observation that stock prices grow with almost constant percentages over time, and we might even try to test a model with both the target as well as the input variable logged (Since Close is also a measure of stock price). However, it would mean that effect of the input variables is multiplicative rather than additive on the target variable, that is, a small percent change in the input induces a proportionate percentage change in target variable. Figure 8 shows the error-vs-time plot which exhibits improved homoscedasticity with lesser autocorrelation (still alarmingly high). So, we ditch earlier model for this one now.

Table 4. Model Evaluation Metrics

Lag	1	2	3
Autocorrelation	0.008	-0.003	-0.019

Autocorrelation is still a big problem with our models. One of the ways to fix autocorrelation is to add a lagged variable of our target function (Note, that now we’re referring to a model where we have logged both the target and the input variables). The reader might have noted that the only input variable in our model is Closed is nothing but the value of our target value ten days prior to the day for which we want to make our prediction. However, it is advisable to arrest autocorrelation at the smallest lag as possible in order to prevent it from percolating to higher lags as well. We find significant autocorrelation at lag 1, and thus add a variable, which is nothing but the target variable lagged by one day. When we plot the Error vs time graph, as shown in the figure 9, we see that the auto-correlation has significantly improved. This fact is confirmed by the autocorrelation figures shown in Table 4.

Figure 9. Error v/s Time plot for model with logged Close and 1-lagged target variable as input and logged target variable

We plot an error histogram in figure 10, and the normality assumption is much more appropriately satisfied than figure 5. As mentioned earlier, the assumption of normality, if violated, is often remedied as a by-product of addressing the other violations. The evaluation metrics for our current model are listed in Table 5

Table 5. Model Evaluation Metrics

	weight	Std err	t-statistic	P-value	95.0% Conf. Int.
Intercept	0.0007	0.003	0.195	0.845	[-0.0060 .008]
log(Close)	0.0319	0.003	10.555	0.000	[0.026 0.038]
log(lagged)	0.9679	0.003	323.678	0.000	[0.962 0.974]

Adjusted R-squared: 0.999

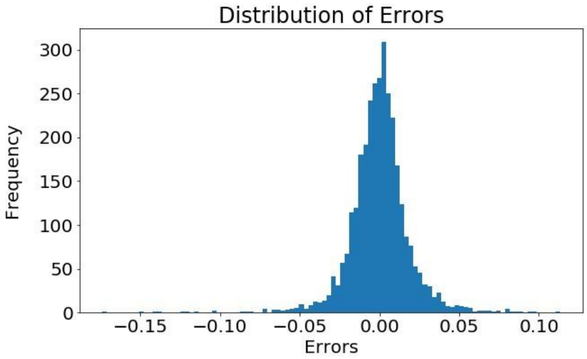


Figure 10. Error histogram for model with logged Close and 1-lagged target variable as input and logged target variable

III. CONCLUSION

So far, we have built a LR model that gives stellar results on the standard evaluation metrics. However, as we tested our model for violations of assumptions of LR, we found that assumptions of homescedasticity and independence were seriously violated. The assumption of normality was also violated to a lesser extent. We then applied appropriate steps to address the issues.

It may be noted that even though we came up with a model that agrees with the assumptions, the results are far from perfect. In fact, there are a couple of anomalies observed throughout our graphs that need to be addressed. We see that even after applying remedying the violations, we observe errors of very high variance in the beginning of the 10-year period, where the value of the stock price was relatively small. We also see high variance around 2008, which can be attributed to sudden steep dip in stock prices owing to the 2008 financial prices.

These anomalies may simply be outliers that may be removed from the data. But we must ask ourselves whether the outliers simply represent statistical flukes or some rare phenomenon that should be accounted for nonetheless. An example is the possibility of a financial crisis happening like one happened in 2008. The crisis was bought about by the liquidity in USA housing market, and therefore it more or less becomes a matter of domain expertise in deciding what our input features should be. It must be noted that the purpose of our analysis was not to build a high quality model, but merely to demonstrate how violations of assumptions of LR can be detected and fixed.

Finally, we may even conclude that perhaps LR may not be the best method to attack the problem. For example, there are many intricacies of modelling stock markets that are far beyond the capabilities of LR. Stock markets are often prone to periods of high and low volatility. This might be the very reason we see high variances in the beginning. This is normal and is often addressed by using ARCH (auto-regressive conditional heteroscedasticity) models wherein the error variance is fitted by an autoregressive model [8].

One of the great difficulties with modelling Stock prices with LR happens to be related to the assumption of independence. In the derivation of the loss function, we assumed our training examples are IID. However, that is quite not the case in real life. A stock's price on a particular day may be effected by its performance during previous days or months. In such a case, one might think of applying a model which takes into account the effect of the previous values of the target variable into consideration while trying to model its current value. Recurrent Neural Networks are one such example and have shown immensely better results when used for this task [9]

However, this does not discount the value of LR as a valuable modelling tool in any way. The simplicity of LR helps it dodge the curse of overfitting [10] which is one of the biggest problems while training a model in machine learning. Even if one can get past the problem of overfitting in complex models, they can often lack the explanatory power of LR. For instance, while using non-linear regression, we can no longer calculate p-values, and confidence intervals are not guaranteed to be calculable, making it hard to interpret the explanatory power of input variables. Even if LR is not well-suited to attack the problem, it can give us valuable insights which may be used later while testing complex models.

IV. ACKNOWLEDGMENTS

We would also like to thank Sumanta Sarathi Sharma, DoPC, Shri Mata Vaishno Devi University for his valuable insights that greatly helped improve the manuscript.

V. REFERENCES

[1]. T. H. B. Farhad Soleimanian Gharehchopogh and S. R. Khaze, "A linear regression approach to prediction of stock market trading volume: A case study," International Journal of Managing Value and Supply Chains (IJMVSC), 2013.

[2]. Paras and S. Mathur, "A simple weather forecasting model using mathematical regression," Indian Research Journal of Extension Education Special Issue, vol. 1, 2012.

[3]. e. a. Gopal, R., "Experimental and regression analysis for multi cylinder diesel engine operated with hybrid fuel blends," THERMAL SCIENCE, vol. 18, no. 1, pp. 193–203, 2014.

[4]. Y. Lifshits and D. Nowotka, "Estimation of the click volume by large scale regression analysis," vol. 1, 2007.

[5]. K.S.U.Libraries, "Spss tutorials Pearson correlation," http://libguides.library.kent.edu/SPSS/PearsonCorr, July 21, 2017.

[6]. R. Nau, "Testing the assumptions of linear regression," http://people.duke.edu/rnau/testing.htm.

[7]. I. D. J. Bross, "Critical Levels, Statistical Language and Scientific Inference," in Foundations of Statistical Inference, V. P. Godambe and D. A. Sprott, Eds. Toronto: Holt McDougal, 1971.

[8]. D. AL-Najjar, "Modelling and estimation of volatility using arch/garch models in jordans stock market," Asian Journal of Finance & Accounting, vol. 8, no. 1, 2016.

[9]. S. V. Murtaza Roondiwala, Harshal Patel, "Predicting stock prices using lstm," International Journal of Science and Research (IJSR), 2015.

[10]. G. C. Cawley, Over-Fitting in Model Selection and Its Avoidance. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 1–1. Online]. Available: https://doi.org/10.1007/978-3-642-34156-41

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Research Issues in Biological Data Mining: A Review

Kanica Sachdev, Manoj Kumar Gupta
Computer Science and Engineering Department SMVDU, J&K, India

ABSTRACT

Biological data is evolving at a very fast rate in the recent years. Large datasets of biological data are now available for analysis and inference. Biological data mining techniques help in the understanding of this data to help biologists to study and visualize the relation between this data under different conditions. This paper presents the biological data mining research areas and the corresponding tools that have been developed in these areas. It studies the various techniques of biological data mining data to provide an idea of the current state of research and introduces future directions for researchers to work in these fields.

Keywords: Biological Data Mining, Visual Data Mining, Biclustering, Pathway Analysis

I. INTRODUCTION

Data mining refers to the analysis of the existing data present in databases in order to generate certain new information [1]. It involves the identification of various associations and patterns from the complex and heterogeneous data by applying various techniques of statistics and machine leaning. It is basically categorized into two sections: descriptive and predictive [2]. Descriptive data mining refers to the characterization and depiction of the existing data. Predictive data mining on the other hand refers to interpretation of the existing information for the purposes of prediction. There have been tremendous advancements in the field of data mining in the past decade. Numerous methods and tools have been developed for the clustering of extensive amount of data, analyzing spatial/ temporal data, sequential and structured pattern analysis, outlier analysis from existing databases etc [3].

Biological data too has seen immense growth in the last decade. Data relating to medical issues and diseases has been increasing rapidly. With the development of new technologies related to medical research and applications, the biological data has been expanding in volume as well as diversity [4]. The scale of current biological data has already gone beyond petabytes and exabytes of storage. Figure 1 shows the accelerated growth in biological data around the year 2011. The research in the field of biology has caused the generation of explosive amounts of medical and clinical data that includes DNA microarrays, to biomedical images, to patient and health records [5]. This has created the need to develop more efficient algorithms and techniques to handle the complex and heterogeneous data, to integrate the varied data from different sources and to develop principles for the manipulation of this data [6]. New methods and tools need to be established that can help analyze this huge volume of biological information so as to develop a better understanding of biological processes. Data mining of biological information and databases hence becomes the greatest challenge for the researchers [7].

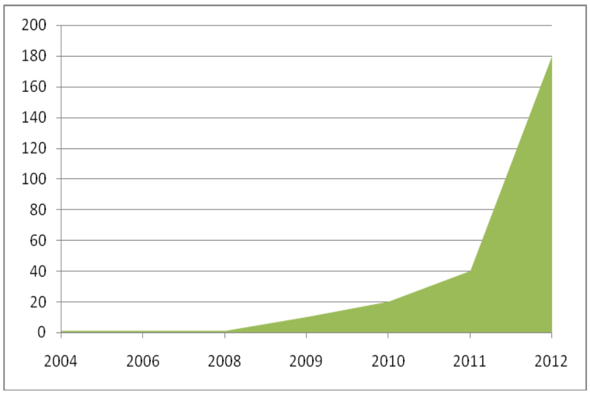


Figure 1. Growth of biological data (in terabases) over recent years [8]

This paper aims to highlight the important research issues in the data mining of biological data. Numerous issues like preprocessing/ cleaning of data, visual analysis of data, pathway analysis of biological data, biclustering etc exist. The goal of this paper is to introduce certain prominent research areas and the noteworthy contributions in those areas. It can serve as a starting point for researchers to understand the fields of research related to biological data mining. Section I gives a basic introduction of biological data mining and its need. Section II highlights the research areas and the prominent works of researchers in those areas. Section III concludes and summarizes the paper.

II. RESEARCH ISSUES IN BIOLOGICAL DATA MINING

Due to the recent growth in medical processes and biological data, many researchers have aimed to develop new techniques that can help in the mining and understanding of this data. Various research issues exist in this field of biological data mining. This paper aims to highlight certain pressing fields of research in this area i.e. visual data mining, biclustering and biological pathway analysis. Visual data mining refers to the representation of data in a form that is easy to analyze and comprehend. Biclustering refers to the grouping of data based on various expression values in order to identify patterns and structures. Biological pathway analysis is also somewhat related to visual data mining. It helps to determine the sequence of evolution of various molecules and genes in order to discover certain changes in them. The following sub sections explain each of these issues and highlight some of the major works done in these areas. Future directions and scope of these areas is also proposed for researchers to give directions for further research.

Visual Data Mining

Visual data mining is a combination of information visualization and computer graphics in the field of life sciences. It refers to the representation of various forms of information like macromolecular structures, genes, sequences, magnetic resonance imaging records etc. It helps to perceive and communicate data, to develop new ideas, as well as to apprehend the biological processes [9]. Data visualization is an important research area in the field of data mining of biological data. The need of data visualization of biological data arises due to various reasons. Firstly, biological data is huge in terms of volume as well as type. The human genome, for example, consists of 3 billion base pairs [10]. Secondly, various biological technologies producing data in the form of DNA microarrays, serial analyses of gene expression (SAGE) etc are developing expeditiously. Hence it becomes difficult to analyze this large quantity of data. Also, visualization tools need to be developed to integrate the heterogeneous sources of data and to model various biological systems. They are required in order to visualize the raw data present in the form of textual annotations tables, images etc and the distributed information stored in diverse spatially and temporally differing data sets. Thus, visual data mining aids the knowledge discovery and comprehension of biological data.

Numerous techniques and tools have been proposed for biological data visualization. The earliest contributions in this field include ACeDB system [11] and the Entrez web browser [12]. These systems integrated the data, the database management system and the user interface as a complete tool provided to the user. The former is a software package coded in C language which is used to handle the physical as well as genetical data and DNA sequences of animals, plants and prokaryotes. It contains a web interface that is flexible to adapt to any database schema. Entrez web browser from NCBI (National Centre for Biotechnology Information) has also integrated the searching and retrieval modules in which the web interface gives access to all datasets concurrently by entering a single query. It is capable of retrieving related structures and sequences. It visualizes various chromosome maps and gene sequences. These integrated systems helped the biologists to easily mine data visually. However, these systems used data that contained only a snapshot of information occurring at the distribution time which later becomes obsolete. The data may be regularly updated over the internet but it is a tedious job and may be prone to errors. Also, these software packages need to be locally installed on a machine and hence cannot be accessed from any location.

Consequently, the BioViews browser applet was introduced which was written in Java [13]. It represents the biological features on physical maps and DNA sequences. The API is connected to various datasets and can retrieve diverse features and present the hyperlinked data on the features that are selected. The browser was built on top of extensible graphic components that can be reused by other programmers without knowing the internal coding details. The widgets in this browser also provide the feature of semantic zooming so as to view the data at differ detailed levels for a better understanding. Zomit was another architecturally independent applet tool that was developed [14]. The earlier systems used in the visualization of biological data generated new pages when a link from a particular page was followed. They provided no relationship between the two linked pages. Zomit overcame this drawback and helped to keep track of the various views for helping the biologists to maintain semantic link between different views.

Other tools that were developed for biological data visualization include Apollo [15] and GBrowse [16]. Apollo is a Java application that allowed the biologists to view genome annotations as well as edit them. It is an interactive tool for biologists that helps them to evaluate the data related to each annotation. GBrowse included many features like scrolling and zooming different regions of a genome, enabling/ disabling tracks, altering the relative order and appearance of tracks etc. It contained open source components and had a simple installation and integration process. Subsequently, GDVTK (Genome Data Visualization Toolkit) was developed [17]. Unlike Apollo that was a standalone application and needs to be locally installed, GDVTK was developed as a library. In comparison to GBrowse, it required less CPU time and memory as it handled the web requests using Java Servlet. GBrowse on the other hand used CGI (Common Gateway Interface) to handle requests that creates a corresponding page on the server to serve each request. However, an important limitation of GDVTK was that it required certain level of technical expertise due to the complexities of J2EE.

Recent tools include IGV (Integrative Genomics Viewer) [18] and BioCircos.js [19]. IGV can handle diverse datasets efficiently and provides an effortless user experience. The main emphasis of IGV is to support array based as well as next generation sequencing data. IGV may be used to visualize the genomic data from public database, however its main focus is to help biologists to visualize and understand their individual data or the data from their contemporaries. Thus IGV provides efficient data visualization on standalone desktop systems. BioCircos.js is a lightweight script used for interactive visualization of biological data. It helps in visualizing bimolecular interactions, gene variations and genomic features. It contains Background module to show axis circles and Text module for the annotations. It supports numerous platforms and can be used on all major web browsers. A comparison of various data visualization tools on the basis of their architecture and main distinguishing feature is given in Table 1.

Although many data visualization tools have been developed for biological data, certain open issues still need to be addressed. The data produced by various experiments contain huge amount of noise. This inserts uncertainty in the visualization representations. This uncertainty needs to be addressed so that the biologists can clearly understand and apprehend data. Also, the quality of various graphical visualization models need to be measured for a better comparison and understanding. Although some measures for comparing the effectiveness of these models have been discussed [20], more comparison measures need to be formulated for an efficient comparison. Additionally, it is difficult to represent the evolving changes in data. Optimizing the display space and the visual considerations in order to represent this high dimensional dynamic information is still a challenge.

Table 1. Summary of data visualization tools

Tool	Architectural details	Main Feature
ACeDB	Database management system with user interface coded in C	Integrated data, database and user interface
Entrez	Web Browser	Can be accessed from any location
BioViews	Java Applet	Built on top of graphic components and contains reusable modules
Zomit	Java Applet	Keeps track of different views
Apollo	Java application	Allows user to view/edit genome annotations
GBrowse	Interactive web pages coded using Java Servlets	Includes features like scrolling, zooming etc.
GDVTK	Java based application framework	Requires less CPU time and memory
IGV	Java application	Mainly for biologists to analyze their individual data
BioCircos.js	JavaScript	Lightweight script with diverse features

Biclustering

Once the data has been visualized using biological data visualization tools, it is still hard to comprehend the results. The process of extracting useful information from the visual data is still a challenging task. An important step in analysis of this data is the grouping of genes that display similar properties or patterns. This grouping/ clustering was shown to be beneficial for the purpose of identification, classification and annotation. However the process of clustering has certain drawbacks. Firstly, it is based on the presumption that similar genes exhibit same properties over all set of conditions. Although, this presumption may hold true when the data is gathered from a single experiment, but it is not correct when the data is accumulated from various experiments and diverse conditions. Secondly, clustering process divides the data into disjoint groups which assumes that each gene belongs to only one biological process or function. This may not always be the case [21].

To overcome the limitations of clustering, subset of genes with similar properties across a subset of conditions are identified. This is achieved by the process of biclustering. The data is organized in the form of matrix with the rows and columns representing these subsets. In the biclustering of biological data, each row represents ne gene and each column represents one condition. Each cell of this matrix shows the expression level of a gene under a specific condition [22]. Biclustering helps to achieve the major objectives of analysis of gene expression data i.e. identifying genes with similar expressions under numerous conditions, identifying conditions with similar gene expressions and classifying new genes based on the expression of other genes [23].

Many biclustering algorithms have been proposed for the classification of genetical data to identify local patterns, where similar properties are being shown by a subset of genes. Cheng and Church suggested an algorithm for biclustering based on MSR (Mean Squared Residue) [24]. The method begins by excluding the rows and colmns where the value of MSR is very high. When the value of MSR becomes greater than or equal to a threshold value, the rows and columns whose residue has a value lesser than the bicluster value are included back. If more than one bicluster s to be selected, the selected biclusters are masked and the process repeats iteratively.

OPSM (Order Preserving Submatrix) is a deterministic greedy algorithm to detect biclusters [25]. Since all biclusters are order preserving, this method represents a bicluster as an order preserving submatrix. The other algorithms aimed to classify all set of genes across all set of experiments, but this algorithm aims to identify a subset of genes in a subset of experiments. It creates biclusters by developing each bicluster iteratively using a probabilistic score that every bicluster will develop to a certain size. The best biclusters at each step of iteration are retained.

MOTIFS is another non deterministic greedy algorithm that constructs biclusters from a dataset with conserved rows [26]. Firstly it constructs intervals by comparing the significance of the interval to uniform distribution. Then, a seed column is selected randomly. For each of the seed columns, the algorithm identifies rows having the same state. Thus, this algorithm detects biclusters with constant row values. This representation has many applications. If the various clusters correspond to various diseases, we can find out the genes that are conserved in many classes but have different values in different classes. These genes can serve as drug targets. Also, the information about the highly expressed genes can be used in pathway identification.

Prelić et al. proposed a divide and conquer algorithm BiMax that searches binary matrix for rectangles consisting of 1s [27]. The entire data matrix is converted into binary form by any thresholding/binarization method. It starts with the entire matrix of data and iteratively divides it into checker board format. Another important proposed algorithm was QUBIC (QUalitative BIClustering) [28]. This algorithm can detct all significant biclusters. Also, it is a fast and efficient algorithm that can construct biclusters from thousands of genes under thousands of condition in only a few minutes. It represents the data in the form of bipartite graphs and identifies heavy subgraphs. The data is first converted to a discrete form and then the biclusters are produced recursively from a seed edge of the graph.

Bayesian biclustering and spectral biclustering was also proposed for constructing biclusters [30, 31]. Bayesian biclustering used Gibbs sampling to detect biclusters. It handled the problem of missing data using Monte Carlo imputation. Spectral biclustering constructed checkerboard structures using eigen vectors for each gene expression. These eigen vectors can be identified using SVD (singular value decomposition) or linear algebra techniques.

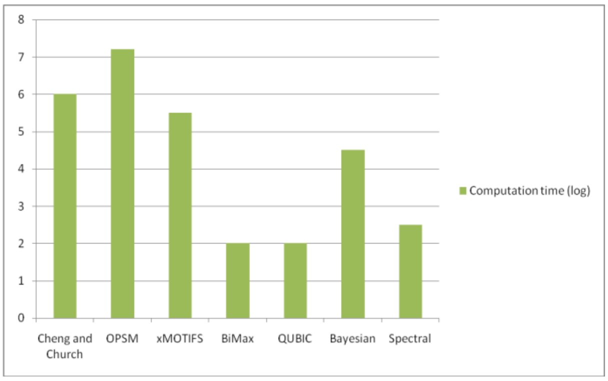


Figure 2. Comparison of biclustering algorithms in terms of running time [29]

Figure 2 shows a comparison of the biclustering algorithms in terms of computation time for a dataset comprising of 4000 rows. As it can be seen BiMax and QUBIC method work efficiently and have a lesser computation time compared to other methods.

Many other algorithms and techniques have been proposed and used for the biclustering of gene data and many other combinations of techniques have been employed for the same. Only a few notable methods have been mentioned in this paper. Biclustering of biological data has several applications in data analysis, data mining and filtering. Various other potential applications can also be uncovered by applying these algorithms in other research areas. Future work in this area can include the comparative study of these biclustering methods with already known biological data for the purpose of validation. The algorithms can also be modified to develop more accurate and efficient techniques. The significance of the extracted biclusters also needs to be studied as the selection of a large number of biclusters may be difficult to analyze in real life applications [32].

Biological Pathway Analysis

Biological pathways are defined as a sequence of interactions in the molecules that result in any transition or modification in the cell. These pathways can lead to the collection of molecules like fats/ proteins. The pathways are associated in certain processes like transmission of signals and gene expression regulation. Biological pathways depict how one or more molecules can be used by the organisms to form new products that are essential for sustenance [33]. These pathways are rarely linear in structure. Mostly they consist of several steps which may contain many intermediates. This makes them complicated to visualize and apprehend.

The visualizations of these pathways are static and can be constructed manually. However, this is a time consuming process. Also manual construction does not give additional information about the pathway and its members. This requires an automation for the construction of biological pathways. Many tools exist for the visualization and generation of these pathways.

IPA (Ingenuity Pathway Analysis) is used to analyze the biological pathways that encompasses four algorithms: Upstream Regulator Analysis (URA) to

find out the possible upstream regulators that are directly/ indirectly related to genes, Mechanistic Networks (MN) associates the regulators that belong to the same mechanism in hypothesis network, Causal Network Analysis (CNA) relates upstream regulators to molecules but focuses on the path having more than one link and hence gives a better analysis of possible causes of observed changes and Downstream Effects Analysis (DEA) analyzes the effect on biological functions that are connected to genes whose expression has changed [34]. IPA helps in various analysis applications. It helps to find the most relevant biological functions/ diseases for a particular set of genes. It also predicts the downstream effects of genes on biological functions/ diseases as well as the activation of upstream regulators. It can also be used to compare the affected pathways across the various experiments/ conditions.
GeneSpring is another tool developed by Silicon Gentics for pathway analysis [35]. It had various interactive features like an interface for an organized file management system, a collection of various clustering tools, has numerous data display methods, can handle input data from different formats, had automated annotation feature. It can be used for performing many functions like measuring similarity, hierarchal and k-means clustering, pathway analysis, constructing self organizing maps etc.

The pathway visualization tools generally do not integrate the molecular interactions with the state measurements so that they can be viewed on a common platform and can be studied over various parameters and biological attributes. Cytoscape was designed keeping this need in mind [36]. It presents an environment to combine bimolecular networks and states on a common environment. It can be used to integrate the diverse data, transfer annotations to desired node/ edge, provides automated graphical layout methods and supports graph selection and filtering.

VisANT is another tool that provides an online interactive interface for the interaction of biological data [37]. It provides tools for mining data and visualizing it in terms of pathway, related annotations, sequence and structure. The analyzed and inter related data can be combined and manipulated using numerous built in features of this software.

The proposal of different tools for biological pathway analysis is progressive and open ended due to the growth and development of their targeted applications. Future research can focus on developing more flexible softwares with easy download [procedures, consistent pathway names and description of pathway overlap mechanism. Also, development of a universal scheme for annotation can help to increase the interoperability between different tools. Future pathway analysis tools can also be developed that have increased computational efficiency. The datasets can also include diverse data like genomic databases, pathways, parameters etc to increase the flexibility of their operation.

III. CONCLUSION

This paper presents an overview of various open ended problems in the field of data mining of biological data. Three important issues i.e. visual data mining, pathway analysis and biclustering are mentioned. Although many tools and techniques have been proposed for these issues, still researchers can help to achieve better efficiency and accuracy and design tools with various features that can help scientists and biologists to analyze biological datasets and infer results from the same.

IV. REFERENCES

[1]. Romero, Cristóbal, and Sebastián Ventura. "Educational data mining: a review of the state of the art." IEEE Transactions on Systems, Man, and

Cybernetics, Part C (Applications and Reviews) 40, no. 6 (2010): 601-618.

[2]. Chen, Ming-Syan, Jiawei Han, and Philip S. Yu. "Data mining: an overview from a database perspective." IEEE Transactions on Knowledge and data Engineering 8, no. 6 (1996): 866-883.

[3]. Han, Jiawei. "How can data mining help bio-data analysis?." In Proceedings of the 2nd International Conference on Data Mining in Bioinformatics, pp. 1-2. Springer-Verlag, 2002.

[4]. Cook, Charles E., Mary Todd Bergman, Robert D. Finn, Guy Cochrane, Ewan Birney, and Rolf Apweiler. "The European Bioinformatics Institute in 2016: data growth and integration." Nucleic acids research 44, no. D1 (2015): D20-D26.

[5]. Li, Xiaoli, See-Kiong Ng, and Jason TL Wang, eds. Biological data mining and its applications in healthcare. Vol. 8. World Scientific, 2013.

[6]. Li, Yixue, and Luonan Chen. "Big biological data: challenges and opportunities." Genomics, proteomics & bioinformatics 12, no. 5 (2014): 187-189.

[7]. Yang, Qiang, and Xindong Wu. "10 challenging problems in data mining research." International Journal of Information Technology & Decision Making 5, no. 04 (2006): 597-604.

[8]. Marx, Vivien. "Biology: The big challenges of big data." Nature 498, no. 7453 (2013): 255-260.

[9]. O'Donoghue, Seán I., Anne-Claude Gavin, Nils Gehlenborg, David S. Goodsell, Jean-Karim Hériché, Cydney B. Nielsen, Chris North et al. "Visualizing biological data—now and in the future." Nature methods 7 (2010): S2-S4.

[10]. Venter, J. Craig, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith et al. "The sequence of the human genome." science 291, no. 5507 (2001): 1304-1351.

[11]. Dunham, I., R. Durbin, J. Thierry-Mieg, and D. R. Bentley. "Physical mapping projects and ACEDB." Guide to human genome computing (1994): 111-158.

[12]. Epstein, Jonathan A., Jonathan A. Kans, and Gregory D. Schuler. "WWW Entrez: A Hypertext Retrieval Tool for Molecular Biology." (1994).

[13]. Epstein, Jonathan A., Jonathan A. Kans, and Gregory D. Schuler. "WWW Entrez: A Hypertext Retrieval Tool for Molecular Biology." (1994).

[14]. Pook, Stuart, Guy Vaysseix, and Emmanuel Barillot. "Zomit: biological data visualization and browsing." Bioinformatics (Oxford, England) 14, no. 9 (1998): 807-814.

[15]. Lewis, Suzanna E., S. M. J. Searle, N. Harris, M. Gibson, V. Iyer, J. Richter, C. Wiel et al. "Apollo: a sequence annotation editor." Genome biology 3, no. 12 (2002): research0082-1.

[16]. Stein, Lincoln D., Christopher Mungall, ShengQiang Shu, Michael Caudy, Marco Mangone, Allen Day, Elizabeth Nickerson et al. "The generic genome browser: a building block for a model organism system database." Genome research 12, no. 10 (2002): 1599-1610.

[17]. Sun, Hao, and Ramana V. Davuluri. "Java-based application framework for visualization of gene regulatory region annotations." Bioinformatics 20, no. 5 (2004): 727-734.

[18]. Thorvaldsdóttir, Helga, James T. Robinson, and Jill P. Mesirov. "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration." Briefings in bioinformatics 14, no. 2 (2013): 178-192.

[19]. Cui, Ya, Xiaowei Chen, Huaxia Luo, Zhen Fan, Jianjun Luo, Shunmin He, Haiyan Yue, Peng Zhang, and Runsheng Chen. "BioCircos. js: an interactive Circos JavaScript library for biological data visualization on web applications." Bioinformatics 32, no. 11 (2016): 1740-1742.

[20]. Bertini, Enrico, Andrada Tatu, and Daniel Keim. "Quality metrics in high-dimensional data visualization: An overview and systematization." IEEE Transactions on Visualization and Computer Graphics 17, no. 12 (2011): 2203-2212.

[21]. Tanay, Amos, Roded Sharan, and Ron Shamir. "Discovering statistically significant biclusters in

gene expression data." Bioinformatics 18, no. suppl_1 (2002): S136-S144.

[22]. Tanay, Amos, Roded Sharan, and Ron Shamir. "Discovering statistically significant biclusters in gene expression data." Bioinformatics 18, no. suppl_1 (2002): S136-S144.

[23]. Madeira, Sara C., and Arlindo L. Oliveira. "Biclustering algorithms for biological data analysis: a survey." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 1, no. 1 (2004): 24-45.

[24]. Cheng, Yizong, and George M. Church. "Biclustering of expression data." In Ismb, vol. 8, no. 2000, pp. 93-103. 2000.

[25]. Ben-Dor, Amir, Benny Chor, Richard Karp, and Zohar Yakhini. "Discovering local structure in gene expression data: the order-preserving submatrix problem." Journal of computational biology 10, no. 3-4 (2003): 373-384.

[26]. Murali, T. M., and Simon Kasif. "Extracting conserved gene expression motifs from gene expression data." In Pacific symposium on biocomputing, vol. 8, pp. 77-88. 2003.

[27]. Prelić, Amela, Stefan Bleuler, Philip Zimmermann, Anja Wille, Peter Bühlmann, Wilhelm Gruissem, Lars Hennig, Lothar Thiele, and Eckart Zitzler. "A systematic comparison and evaluation of biclustering methods for gene expression data." Bioinformatics 22, no. 9 (2006): 1122-1129.

[28]. Li, Guojun, Qin Ma, Haibao Tang, Andrew H. Paterson, and Ying Xu. "QUBIC: a qualitative biclustering algorithm for analyses of gene expression data." Nucleic acids research 37, no. 15 (2009): e101-e101.

[29]. Eren, Kemal, Mehmet Deveci, Onur Küçüktunç, and Ümit V. Çatalyürek. "A comparative analysis of biclustering algorithms for gene expression data." Briefings in bioinformatics 14, no. 3 (2012): 279-292.

[30]. Gu, Jiajun, and Jun S. Liu. "Bayesian biclustering of gene expression data." BMC genomics 9, no. 1 (2008): S4.

[31]. Kluger, Yuval, Ronen Basri, Joseph T. Chang, and Mark Gerstein. "Spectral biclustering of microarray data: coclustering genes and conditions." Genome research 13, no. 4 (2003): 703-716.

[32]. Madeira, Sara C., and Arlindo L. Oliveira. "Biclustering algorithms for biological data analysis: a survey." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 1, no. 1 (2004): 24-45.

[33]. Wang, Kai, Mingyao Li, and Hakon Hakonarson. "Analysing biological pathways in genome-wide association studies." Nature reviews. Genetics 11, no. 12 (2010): 843.

[34]. Kramer, Andreas, Jeff Green, Jack Pollard Jr, and Stuart Tugendreich. "Causal analysis approaches in ingenuity pathway analysis." Bioinformatics 30, no. 4 (2013): 523-530.

[35]. Chu, Lillian, Eric Scharf, and Takashi Kondo. "GeneSpringTM: tools for analyzing microarray expression data." Genome Informatics 12 (2001): 227-229.

[36]. Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome research 13, no. 11 (2003): 2498-2504.

[37]. Hu, Zhenjun, Joseph Mellor, Jie Wu, and Charles DeLisi. "VisANT: an online visualization and analysis tool for biological interaction data." BMC bioinformatics 5, no. 1 (2004): 17.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Review on Missing Value Imputation Techniques in Data Mining

Arjun Puri, Dr. Manoj Gupta
Department of Computer Science and Engineering, Shri Mata Vaishno Devi University, Katra, Jammu and Kashmir, India

ABSTRACT

Now days, there are huge amount of data available for analysis, the main problem with the data is inconsistency. The inconsistent data (missing value) need to replace with most appropriate fit values. Some missing values are dependent on some known variable in the dataset need to be taken for further calculation. There are different methods to impute these missing values. In this paper, we discuss various technique based on their classification and also discuss their behavior in different datasets under different types of missing values.

Keywords : Missing value imputation, data mining, data preprocessing, Techniques for missing value imputation, MCAR, MAR, NMAR.

I. INTRODUCTION

In real world scenario, we are dealing with data and analysis of data. In order to deal with the extraction of knowledge from the given raw data, data mining is one of the important branch. There are many steps involved in the getting meaningful information from raw data. One of the important step is data preprocessing; the technique which helps in improving the quality of data and also improve mining results. One of the important issues in data preprocessing is missing value. It plays a vital role in deciding the computational results obtained by data preprocessing. Missing value can be caused from different sources like: sensor failure, corrupted datasets, incomplete survey etc. (Irfan Pratama, 2016). Inconsistence of data (missed values) is of different types, some of them are discussed below:

1. Missing completely at random (**MCAR**), if there is no dependency in the missing data is related to its known values. In this type of missing data we assume that a whole distribution of data is completely missed.
2. Missing at random (**MAR**), when the missing value depend on the already known value and does not depend upon missed value itself.
3. Not missing at random (**NMAR**), when missed value does not depend upon any given or missed value. [(Irfan Pratama, 2016), (Shichao Zhang, 2011), (Julián Luengo, 2012)]

These types of anomalies generally arise due to different sources like: MCAR can arise due to sensor recording failure because no data is dependence in between them whereas, MAR can arise during the survey question some question are not answered by the people but there are other questions related with them (Irfan Pratama, 2016).

In order to deal with the missing values there are many techniques developed so for, some of them usually ignore the missing values and some of them delete and some techniques use imputation. Broadly speaking, these techniques divided into two main types: conventional techniques (like mean, mode, median and deleting values) and modern techniques (hot deck, cold deck (Geeta Chhabra, 2017), classification techniques like SVM)[(Alireza Farhangfara, 2008), (Julián Luengo, 2012)]. In this research paper, we survey some of the techniques to deal with missing value imputation and compare them in contrast in the following sections: literature survey, methods deal with missing value imputation, discussion of different techniques on different datasets and conclusion.

II. LITERATURE SURVEY

Various researcher compare different techniques on different datasets analyze their outputs and also suggest that which technique is suitable for which dataset [(Alireza Farhangfara, 2008) (Geeta Chhabra, 2017) (Julián Luengo, 2012) (Schmitt P, 2015) (Irfan Pratama, 2016). Some other researchers develop technique to improve accuracy in imputation of values.

In Alireza Farhangfara et al(2008), in which a compartative study was made which includes six single and multiple imputation methods on 15 discrete incomplete datasets. In this paper researcher find that imputation improves by using classification techniques, except for the mean imputation method which shows poor results with high rate of missing values (50%). In this paper,researcher conclude that Naive-Bayes based imputation shows better result by using RIPPER classification on datasets with high amount of missing values, i.e. 40% and 50%. Researcher also show that multiple imputation polytomous regression method shows best result with SVM on different datasets. Finally, shows that the mean imputation is least beneficial.

In Sasi et al.(2016), an intelligent approach was suggested by the authors to deal with data of different types. In this authors take 3 different datasets(name: iris, credit and adult) and preform missing value imputation by using different approaches(name: Mean/Mode, K-Nearest Neighbor, Hot-Deck, Expectation Maximization, and C5.0 imputation techniques) by using and IITMV menthod which decide which dataset missing values and operated by which technique. In this paper authors, concluded that IITMV technique shows better results compared with C5.0 algorithm.

In Esther-Lydia Silva-Ramírez et al(2011), A methodology for data imputation by means of artificial neural networks has been proposed and empirically compared with three classic methods: mean/mode imputation, regression models and hot-deck. Fifteen datasets are used for elvaluation and observed that multilayer perceptron provide better results.

In Irene Erlyn Wina Rachmawan et al(2015), An algorithm from machine learning for missing value called Reinforcement Programming was proposed. Reinforcement programming shows better result as compared with zero imputation, Mean imputation and Genetic Algorithm.During evaluation researcher find that Reinforcement Programming could run better in solving Missing imputation.

In Schmitt.P et al(2015), a comparison of six imputation methods(Mean, KNN, SVD, maximization expectation imputation, bPCA and MICE) based on four real datasets (iris,e.coli,breast cancer1,breast cancer 2) of various sizes, under an MCAR assumption with missing values ratio percentage of missing values (from 5% to 45%increased by 10%). By using different evaluation techniques authors identifed performance of different techniques : Root mean squared error (RMSE), unsupervised classifcation error (UCE), supervised classifcation error (SCE) and execution time. While much attention has been paid to the imputation accuracy measured by RMSE. Results shows that, MICE technique is complex in structure and show better results in case of small datasets. While bPCA and FKM shows better results with large datasets.

In Ibrahim Berkan Aydilek et al (2013), a hybrid method was proposed by the resaercher by using support vector regression and genetic algorithm with fuzzy clustering to estimate missing values. A Complete train data were clustered based on their similarity, and fuzzy principles were used during clustering. Therefore, each missing value becomes a member of more than one cluster centroids, which yields more sensible imputation results. Six datasets with different characteristics and also with differnet missing value ratios were used in this paper, and resulted showed that better result obtained as compared to other techniques.

Some of the techiques for missing values imputation discussed during literature survey are as under.

2.1 CLASSFICATION OF MISSING VALUE TECHNIQUES

There are so many approach and techniques developed so far to deal with missing values. Researchers develop many techniques ranges from simple to complex. Researcher generally makes division but these techniques some of them deal with low missing values and some of them deal with higher missing values. These techniques are discussed as under:-

1. **Mean imputation:** In this technique, mean of missing value is calculated by using the corresponding attribute value. This technique is faster over other techniques. It shows good result when data is small, but result is not good for big data. This model is helpful for only MAR but not useful for MCAR [(Irfan Pratama, 2016), (Jason Van Hulse, 2008), (Sasi, 2016)].
2. **Hot deck imputation:** this method is used for categorical data and it is beneficial for big data and not for small data. In this method, missed valued is replaced by the most similar values of that attribute, this method becomes problematic when there is no other similar data is available [(Irfan Pratama, 2016), (Alireza Farhangfara, 2008), (Sasi, 2016) (Esther-Lydia Silva-Ramírez, 2011)].
3. **K-nearest Neighbor imputation (KNN):** This technique used Euclidean distance to determine the similarity between two values and replace the missing one with similar one .The main benefits of this approach are as given as under:
   •KNN is useful for datasets having both qualitative and quantitative attribute values.
   •There is no need for creating a predictive model for each attribute of missing data and helpful for multiple missing values.
   The main drawback of the KNN approach is that, whenever the KNN looks for the most similar instances, the algorithm searches through all of the data set (Sasi, 2016).
4. **Regression Imputation:** This technique is applied by using known values for the construction of model and calculates the regression between variables and then applied that model to calculate the missing values. This technique gives more accurate results than mean imputation (Jason Van Hulse, 2008).
5. **REPTree imputation:** REPTree is a decision tree used for the analysis of independent variables in comparison with quantitative dependent variables. In this process recursive technique are applied to complete the incomplete dataset with least error by using reduced error pruning by using variance. (Jason Van Hulse, 2008).
6. **Support Vector Regression:** This method is extension of Support vector machine. In Support vector machine generally missing values are ignored first and then rest of data is feed to train the system and then missing values are filled with the trained system (Irfan Pratama, 2016). By using regression with support vector machine classifier efficiency will increase (Alireza Farhangfara, 2008).

7. **Fuzzy mean imputation:** It is the technique which uses fuzzy in the calculation of missing value with the help of clustering in the known value and finding which missing value belong to which cluster there are two different ways to calculate fuzzy mean one is K-mean and other is C-mean. C-mean is better than K-mean in most of cases (Schmitt P, 2015).
8. **Reinforcement Programming:** It is generally used for dynamic approach for the calculation of missing values by using machine learing approaches. It has capability of convergence and to solving imputation problem by using exploration and exploitation (Irene Erlyn Wina Rachmawan, 2015).
9. **Nonparametric Iterative Imputation algorithm (NIIA):** It is an iteratively imputing the missing values in a dataset. It works as follows:
   Identify some missing values and then compute the values of all complete values used to estimate these incomplete values. Then these missed value imputed is used for further analysis of other incomplete instances and the process repeat until the values of dataset are completely filled. (Shichao Zhang, 2011)
10. **Multilayer Perceptrons:** Multilayer perceptrons is the technique to develop by using artificial neural networks. It runs as on multilayer and also use different learning processes to train the network (Esther-Lydia Silva-Ramírez, 2011).

III. DISCUSSION

In this paper we review different techniques and with different dataset and analysis that which technique is giving best result in which type of dataset. This collaborative information is represented with the help of following table.

Table 1. List of various papers on missing value imputation techniques.

Research paper	**Datasets**	**Techniques**	**Remarks**
Geeta Chhabra et. al. (2017)	Iris	1. Predictive Mean Matching 2. Multiple Random Forest Regression Imputation. 3. Multiple Bayesian Regression Imputation 4. Multiple Classification and Regression Tree (CART). 5. Multiple Linear Regression using Non-Bayesian Imputation. 6. Multiple Linear Regression with Bootstrap Imputation.	A comparison of different approaches of MICE methods on iris datasets. Efficiency gain with Multiple Imputation combined with Bayesian Regression is that it is able to make better use of the available information by accommodating non linearities among the predictors.
Ibrahim Berkan Aydilek et al (2013)	1. Iris 2. Haberman 3. Glass 4. Musk1 5. Wine 6. Yeast	1. SvrFcmGa (proposed) 2. FcmGa 3. SvrGa 4. ZeroImpute	Dataset inconsistence can be ranged from 10% to 25% and analyze that SVRFCMGA (Fuzzy C-mean with Support vector Regression and Genetic algorithm) perform better than other.

Sasi et al. (2016)	1. Iris 2. Credits 3. Adults	1. Mean/Mode. 2. Hot Deck. 3. Expectation Maximization. 4. K neighbor nearest.	In this paper, authors compare C5.0 with this new developed technique known as IITMV and also show its performance on different data sets.
Esther-Lydia Silva-Ramírez et al. (2011)	1.Cleveland 2.Heart 3. Zoo 4. Buhl1-300 5.Glass 6.Ionosphere 7.Iris 8.Pima 9. Sonar 10.WaveForm2 11.Wine 12.Hayes-Roth 13. Led7 14.Solar 15. Soybean	1. mean/mode 2. Regression. 3. Hot deck. 4. ANN.	Result shows that Multilayer perceptrons (MLP) with different learning rules show better results with quantitative datasets as compared with classical imputation methods. In this paper, type of missing value is missing completely at random (MCAR) is taken.
Schmitt P et al.(2015)	1. Iris 2. E. coli 3. Breast cancer 1 4. Breast cancer 2	1. Mean 2. K-nearest neighbors(KNN) 3. Fuzzy K-means (FKM) 4.Singular value decomposition(SVD) 5.Bayesian principal component analysis (bPCA) 6.Multiple imputations by chained equations (MICE).	Results show that different techniques are best at different datasets and different size. MICE is useful for small datasets but for big datasets bPCA and FKM are better one.

In the above table, different researchers compare different techniques on the bases of RSME and calculate difference between correct dataset with incorrect datasets, also predict the efficiency of particular techniques. In Geeta Chhabra et al(2017), discuss various techniques regarding MICE and concluded that is with Multiple Imputation combined with Bayesian Regression gives better efficiency than other techniques, where as in Schmitt P et al.(2015), compare different techniques and concluded that MICE technique are useful for small dataset error replacement. In Ibrahim Berkan Aydilek et al(2013), research made a comparsion between different hybrid techniques on different datasetswith change in missing values (ranges from 10% to 25%), and concluded that SvrFCmGA gives better preformance than other techniques (FcmGa, SvrGa, ZeroImpute). In Schmitt P et al. (2015), for big datasets bPCA and Fuzzy k mean gives better result. In Sasi et al. (2016), author compute different types of datasets on different techniques and gives classification of dataset that which technique suits what kind of datasets and also proposed and test his technique with C5.0 technique.Now a days, the development of new method is done with combining different techniques together.

IV. CONCLUSION AND FUTURE WORK

Missing value is one of the challenge in the fields of data analysis. In this paper, we discussed various techniques dealing with the imputation depending on different datasets and different missing value type (MCAR, MAR) and study the behaviour of different techniques with different percentage of missing values (10%,20%,40%, etc.), find out that there is no such one technique to deal with all datasets. In study, we reach over the conclusion that many research are trying to combine many techniques together to implement intelligently on different datasets and uses a decision algorithm to pick one out of them.

In future work, we need to develop techniques for unclassified datasets (such as, estate estimation problem for nonlinear stochastic timedelay systems with missing measurements) having better efficiency and accuarcy. Moreover, while analysing we found that there is need of intelligent system which make decision regarding which techniques is suitable for which type of datasets.

V. REFERENCES

[1]. Alireza Farhangfara, L. K. (2008). Impact of imputation of missing values on classification error for discrete data. Pattern Recognition , 3692-3705.

[2]. Esther-Lydia Silva-Ramírez, R. P.-M.-C.-D.-d.-l.-V. (2011). Missing value imputation on missing completely at random data using multilayer perceptrons. Neural Networks , 121–129.

[3]. Geeta Chhabra, V. V. (2017). A Comparison of Multiple Imputation Methods for data with Missing Values. Indain Journal of Science and Technology , 1-7.

[4]. Ibrahim Berkan Aydilek, A. A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Information Sciences , 25–35.

[5]. Irene Erlyn Wina Rachmawan, A. R. (2015). Optimization of Missing Value Imputation using Reinforcement Programming . International Electronics Symposium (IES), (pp. 128-133).

[6]. Irfan Pratama, A. E. (2016). A Review of Missing Values Handling Methods on Time-Series Data. International Conference on Information Technology Systems and Innovation (ICITSI) (p. 6). Bandung-Bali : IEEE.

[7]. Jason Van Hulse, T. M. (2008). A comprehensive empirical evaluation of missing value imputation in noisy software measurement data. Journal of System and Software , 691-708.

[8]. Julián Luengo, S. G. (2012). On the choice of the best imputation methods for missingvalues considering three groups of classification methods. , Knowledge Information System , 77–108.

[9]. Sasi, T. A. (2016). Intelligent Imputation Technique for Missing Values . International Conference on Advances in Computing, Communications and Informatics (ICACCI), (pp. 2441-2445). Jaipur, India.

[10]. Schmitt P, M. J. (2015). A Comparison of Six Methods for Missing Data Imputation. Journal of Biometrics and Biostatistics , 2-6.

[11]. Shichao Zhang, Z. J. (2011). Missing data imputation by utilizing information within incomplete instances. The Journal of Systems and Software , 452–459.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Parallel Corpora : A Much-Needed Linguistic Resource for Low Computational Resource Languages

Preeti Dubey
Assistant Professor, Department of Computer Science, J&K Higher Education Department, India

ABSTRACT

Natural language Processing (NLP) is one of the upcoming research areas of computer science. There are many applications of NLP, but in the last decade, most of the effort in this field is inclined towards machine translation. A lot of work is available for the machine translation of English and Hindi. Some work is also undertaken for the translation of Indian languages, therefore; there has been a revolutionary research in development of text in machine readable form. Currently efforts are being made for developing large parallel corpora for most Indian languages, which is a much-needed linguistic resource for the development of Statistical Machine Translation systems. This paper introduces the concept of parallel corpus, its need and application in natural language processing. The various projects undertaken for the development of parallel corpus, followed by tools where parallel corpus is applied is also presented. The need of development of this resource for languages with low computational resources is also discussed.

Keywords : Text Corpus, Speech Corpus, Parallel Corpora, Natural Language Processing, Low Resource Languages

I. INTRODUCTION

Machine Translation Systems are in great demand and are widely in use. For the past few years, a number of Machine Translation Systems has been developed for Indian as well as foreign languages. The efficiency of a machine translation system depends upon the accuracy rate of the output produced by the system. Therefore, machine translation is not mere dictionary based substitution of words of one natural language into another natural language, but it needs to preserve the meaning of the sentences just like a human translator. There are many available approaches that can be used for machine translation. Some famous approaches are: Direct, Indirect and statistical. Recently machine translation systems are also being developed based on deep learning methods. The statistical approach of MT is widely used as most systems developed using this method have highly accurate results.

II. THE STATISTICAL APPROACH

of machine translation is being widely used for the purpose of achieving efficient outputs. These systems require a large parallel corpus and the working is based on statistical methods like the Bayes' Theorem. The text to be translated is matched with that in the corpus and translation is done with the text has maximum frequency. Some statistical machine translation systems that display highly accurate results have been developed for the following language pairs: Hindi-Punjabi, Punjabi-English, English-Urdu, Telugu, Gujarati-English, Bengali -

English etc. As read in the literature, the SMT output is coarse due to lack of corpora for Indian languages or due to small size of the corpus. As studied by NJ Khan et.al. [6], the results of SMT system that takes the Indian language (Hindi, Urdu, Bengali, Tamil, Malayalam, Telugu) sentences as input and it generates corresponding closest translation in English. The translation of over 800 sentences were evaluated using automatic evaluation metric i.e. BLEU evaluation. The reported average BLEU score was 10% to 20% for all the languages. It is concluded by the authors in their study that the quality of translation is directly dependent on the scope and quality of parallel language corpora.

Statistical methods are not only being used for the development of machine translation systems but also for the evaluation of machine translation systems. Evaluation of the output produced by the machine translators is very important. Earlier the evaluation of these systems was completely manual i.e. the evaluation was done by linguists manually. Therefore, it was time consuming and a cumbersome process. Presently many statistical evaluation tools are available. Some widely used automatic evaluation tools are: BLUE, NIST etc.

The major requirement of any statistical tool is the parallel corpus. The accuracy any statistical tool whether a statistical machine translation system or a statistical evaluation system depends on the size of the corpus used by it.

III. CORPUS

A corpus is a collection of text or phrases of a language that can be used as a sample of the language. Corpus can be text as well as spoken. Collection of spoken/speech corpus is difficult as compared to text corpus. Corpus can also be a part of a larger corpora, such a corpus is called sub corpora. Sub corpora can also be domain specific for example corpora containing only technical text or corpora for the medical domain, tourism etc.

A parallel corpus is a collection of texts, translated into one or more languages. If it involves two languages such that one of the corpora is an exact translation of the other, then it is referred as a bilingual corpus If some corpora involves more than one language such that one of the corpora is an exact translation of the more than one language, it is called multilingual parallel corpora.

IV. NEED FOR PARALLEL CORPUS

To enhance research on computational linguistics, there is a great need to generate linguistic resources which can further be used for developing tools that can be used for languages that are computationally low-resourced. Dogri is one such language which has how computational resources. It is a language used in the state of Jammu and Kashmir. It is a constitutional language of India. Presently, there is no work done so far related to the technological development of Dogri and which is the need of the hour. Only one tool that is the **Hindi to Dogri machine translation system** developed by the author in 2014 is available for the automatic translation of Hindi Text into Dogri text.

V. CHARACTERISTICS OF CORPUS

I. The corpus should be as large as possible, since the accuracy of the system developed depends on the size/quantity of the corpus used.

II. It should have a variety of text/speech samples. The efficiency of the system also depends on the variety of samples in the corpus. Therefore, the quality of the corpus must be varied.

VI. INITIATIVES TAKEN FOR THE DEVELOPMENT OF PARALLEL CORPUS FOR INDIAN LANGUAGES

- **GyanNidhi Parallel Text Corpus**

It contains million pages multilingual parallel text corpus in English and 11 Indian languages. It is a useful resource that can be used for as improving translation system, and also be useful for other applications such as spell checkers dictionaries. Kiran Pala, Sriram Chaudary, Lakshmi Narayana kodavali and Keshav Singhal (2008) have worked on the alignment of English to Hindi texts in Gyan Nidhi parallel corpus at sentence level.

- **ILCI (Indian Languages Corpora Initiative)**

This project is funded by Technology Development for Indian Languages (TDIL) unit of Ministry of Communication and Information Technology (MCIT) for building parallel corpus for major Indian languages including English.

The project is aimed at building parallel corpora for Hindi (SL). It is focused on two domains namely: health and tourism.

- **EMILLE Corpus**

The EMILLE Corpus is a collaborative effort by the EMILLE Project (Enabling Minority Language Engineering), Lancaster University, UK, and the Central Institute of Indian Languages (CIIL), Mysore, India. The EMILLE/CIIL Corpus (ELRA-W0037) is distributed free of charge for use in non-profit-making research only.

- **TIDES**

It is a Hindi-English corpus which was originally collected for the DARPA-TIDES surprise language contest in 2002. It was later refined at IIIT Hyderabad and provided for the NLP Tools Contest at ICON 2008. It contains 50K sentence pairs taken mainly from news articles.

- **WMT (Workshop on Machine Translation)**

In 2014, WMT introduced English-Hindi as an experimental, low resource language pair.

- **The Hindi-Punjabi parallel corpus**:

was developed using the existing Hindi to Punjabi machine translation system developed by Vishal Goyal. Vishal Goyal and Pardeep Kumar (2010) have contributed by developing the parallel corpus for this language pair.

VII. CONCLUSION & FUTURE SCOPE

In current situation of NLP, research is progressing for Indian languages that have the required linguist resources for their automization, whereas the computationally low resources languages are still struggling Low resourced languages are the languages for which the computational resources required for the automatic translation of two languages are not available. Computational resources like machine readable dictionary, corpora etc are very important for the development of NLP tools. It is very challenging in terms of time and money to start from scratch. Dogri is one such low resourced language. The only NLP tool for this language is the HINDI-Dogri machine translation system developed by the author. The author is also working on the development of this resource using the existing HINDI-DOGRI machine translation system. to make way in this field. The plight is that a computer researcher has a financial constraint to develop the linguist resources and on the other hand, a linguist lacks the computational knowledge. The state of art is that the research on NLP for these computationally low resourced languages can progress only if the required resources are developed, enabling the regional languages researchers to find flaws in the available methods and develop new techniques/algorithms. Therefore, the development of these resources must be encouraged for the processing of every low resourced Indian language.

VIII. REFERENCES

[1]. Akshar Bharati, Dipti Misra Sharma, Rajeev Sangal et al., (15th December, 2006), AnnCorra: Annotating Corpora, Guidelines for POS and Chunk Annotation for Indian Languages. Retrievedfrom http://researchweb.iiit.ac.in/~rashid.ahmedpg08/ilmtdocs/chunk-posann-guidelines-15-Dec-06.pdf (15th December, 2006) AnnCorra: Annotating Corpora, Guidelines for POS and Chunk Annotation for Indian Languages. Retrievedfrom http://researchweb.iiit.ac.in/~rashid.ahmedpg08/ilmtdocs/chunk-pos-ann-guidelines-15-Dec-06.pdf

[2]. Ben Langmead. (n.d.) Hidden Markov Models. Retrieved from http://www.cs.jhu.edu/~langmea/resources/lecture_notes/hidden_markov_models.pdf

[3]. PAN Localization. (n.d.). Retrieved from http://www.panl10n.net/english /Outputs%20 Phase%202/CCs/ Nepal /MPP/Papers/2008/Report% 20on% 20Nepali%20Computational%20Grammar.pdf

[4]. Chirag Patel et. al., Part-Of- Speech Tagging for Gujarati Using Conditional Random Fields, Proc. Of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, 2008, pp.117-122.

[5]. NJ KHAN,et.al, ‘ machine translation approaches and survey for indian languages' https://arxiv.org/ftp/arxiv/papers/1701/1701.04290.pdf

[6]. Mutatis Iqbal, et.Al ‘ English to Kashmiri machine translation system, International journal of Advance Research in Computer Science & technology (IJARCST 2015),vol:3, issue2 (Apr. - Jun. 2015), ISSN : 2347 - 8446 (Online) ISSN : 2347 - 9817 (Print)

[7]. Raghavendra Udupa U, et. Al, “ An English-Hindi Statistical Machine Translation System”, Part of the Lecture Notes in Computer Science book series (LNCS, volume 3248), LNAI 3248, pp. 254–262, 2005. https://link.springer.com/chapter/10.1007/978-3-540-30211-7_27

[8]. Tej Bahadur Shai et al. 2013. Support Vector Machines based Part of Speech Tagging for Nepali Text, International Journal of Computer Applications, May 2013, Vol: 70-No. 24, pp. 0975-8887.

[9]. Prajadip Sinha et al. Enhancing the Performance of Part of Speech tagging of Nepali language through Hybrid approach, International Journal of Emerging Technology and Advanced Engineering.2015 Vol 5(5).

[10]. Antony P J et al. 2011.Parts of Speech Tagging for Indian Languages: A Literature Survey, International Journal of Computer Applications, 2011, Vol. 34(8), pp. 0975-8887.

[11]. Amruta Godase, “MACHINE TRANSLATION DEVELOPMENT FOR INDIAN LANGUAGES AND ITS APPROACHES”, International Journal on Natural Language Computing (IJNLC) Vol. 4, No.2,April 2015, ISSN: 2278-1307

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

GLCM Parameters and their Relationship for Dermatological Image Analysis

Chaahat Gupta[1], Dr. Naveen Kumar Gondhi[2], Dr. Parveen Kumar Lehana[3]
[1]Research Scholar, SMVDU, CSE Katra, India
[2]Assistant Professor SMVDU, CSE Katra, India
[3]Professor, University of Jammu, Department of Electronics Jammu, India

ABSTRACT

Analysis of texture is one of the significant perspectives of human vision which is used to segregate objects and surfaces. Texture based features are widely used in analysis of various images for medical diagnosis. Dermatology is the branch of medical science for analysing and treatment of skin abnormalities. Dermatological diseases are the most universal diseases occurring worldwide in people of all ages. These days, image processing techniques are generally used in various medical fields for improving classification, identification and treatment stages of various dermatological diseases. In this paper, texture analyses for 3 Pyoderma diseases (Boil, Carbuncle, and Impetigo Contagiosa) are done using GLCM. 4 different image features (Energy, Correlation, Contrast and Homogeneity) are extracted for three color channels from the given input images. Contrast measures the coarseness of texture in an image. Correlation calculates the linear dependency of gray levels on neighboring pixels. Energy calculates the textural uniformity of the image. Homogeneity finds out the distribution of elements in an image. The mean of Contrast, Correlation, Energy and Homogeneity are calculated. Also, standard deviations of these parameters are found. The histograms show that textural features for individual diseases are different from each other. Hence, it shows promising results that different diseases can be classified and identified into separate categories using machine learning algorithms. In our future work, we will use Gaussian Mixture Model for classification and identification of various categories of dermatological diseases.

Keywords : Dermatology, feature selection, Gray Level Co-occurrence Matrix, Image segmentation, Skin abnormalities.

I. INTRODUCTION

The use of digital images has increased at a brisk speed over the past decade. Photographs, hard copy media and printed text are now frequently converted into digital form. Various medical imaging techniques also generate images directly in digital form. The recognition, labeling, quantitative measurement, etc. of specific structures are involved in the analysis of medical images. Therefore, to provide clinical information about an object in terms of its size and shape, image segmentation, feature extraction and classification are important techniques needed to give the desired information.

Skin protects humans against germs and plays an important part in monitoring body metabolisms against foreign bodies. Skin lesion recognition has become a popular topic for research [33] [34]. Dermatology is the branch of medical field for analysing and treatment of skin abnormalities and inconsistencies. It deals with hair, nails, skin and its diseases. Dermatological diseases are the most universal diseases occurring worldwide in people of all ages Human skin is one of the crucial areas to synthesize and analyze due to its complexity.

The main objective of our work is to analyze textural patterns for 3 Pyoderma diseases (Boil, Carbuncle and Impetigo Contagiosa) using the GLCM. Four different image features (Contrast, Correlation, Energy and Homogeneity) are extracted for three color channels from the given input images.

Contrast measures the coarseness of texture in an image. Correlation finds the linear dependency of gray levels on neighboring pixels. Energy measures the image textural uniformity and Homogeneity calculates the distribution of elements in an image. Mean of Contrast, Correlation, Energy and Homogeneity are calculated.

Standard deviations of these parameters are found. The histograms show that textural features for individual diseases are different from each other. Hence, these results are a positive indication that different diseases can be classified into different groups. Figure 1 shows the proposed methodology of our work.

Our research paper is arranged into following sections. Section II gives us views of related research on textural extraction using GLCM technique and various classification methods for dermatological diseases. Section III explains briefly explains about textural analysis and GLCM. In Section V, we introduce methodology of the proposed work. Section VI, explains system experimental results and their discussions. Section VII, concludes the work done and explains future work.

II. LITERATURE SURVEY

This section provides an overview GLCM technique for feature extraction and classification techniques.

Sertan Kaya et al. [5] paper proposed a method to do quantification of sharpness of lesion patterns in addition to boundary of lesion. Lesion border detection method based on density detects skin lesion. The method was tested and validated on various dermoscoy images and the results indicated that proposed method was efficient on detection of malignancy.

In paper [7], the authors have developed an efficient segmentation system by merging skin detection, color segmentation, and morphological imaging. Also, a set of features which can efficiently presents the color and texture variation for healthy skin, mild eczema and severe eczema are extracted. The proposed system is a first prototype which shows that an automatic eczema detection and severity measurement system is possible. The authors have concluded that the system accuracy could be better in accuracy if calibrated images were used in the dataset of images.

Teck Yan Tan et al. [9] proposed a system for the recognition of malignant and benign skin tissues of dermatological images. The Genetic Algorithm was also applied to identify the most discriminative feature subsets to improve classification accuracy. The proposed work has been evaluated with 100 images from the dataset and authors concluded that their work achieved an average accuracy of 92% and 84% for respective classification of skin lesions.

The paper [10] addressed Self Organization Maps (SOM) technique for clinical and pathological findings and investigation of cluster of conditions. The proposed algorithm was implemented for six types of Erythemato-Squamos with SOM skin diseases separately and together. The proposed method tried to improve the SOM classification performance using networks and various artificial intelligence techniques together in the classification of the point where the SOM network failed.

The paper [2] gives a study on various segmentation approaches that can be applied for melanoma detection using image processing. Region merging, adaptive thresholding, etc. are discussed in this paper. A comparative study of these segmentation methods is also performed based on the parameters accuracy, sensitivity and specificity. The authors have concluded that multilevel thresholding has the highest accuracy and specificity and maximum sensitivity is obtained for iterative stochastic region merging.

R.B Aswin et.al [3] gave a system which could automatically separate and detect various types of skin cancers. GLCM algorithm is used for extraction features of skin lesions. The authors have concluded that the accuracy of their method is 81.43%.

In paper [4], the authors have presented an important segmentation approach using GLCM method. The authors have discussed implementation of skin tissue segmentation. Experimental results demonstrated that based on texture descriptors GLCM algorithm extracted the skin lesions.

P.B.Sangamithraa et.al [1] have developed a system that first segments the region of interest (lung) and which then analyses separately the obtained area for nodule detection in order to examine the disease. The segmentation of the CT images has been carried out by using K- Means clustering method. To the clustered result, EK-Mean clustering was applied. For classification, Back Propagation Network was used resulting in image to be classified as normal image or the tumor image which gave an accuracy of more than ninety percentage.

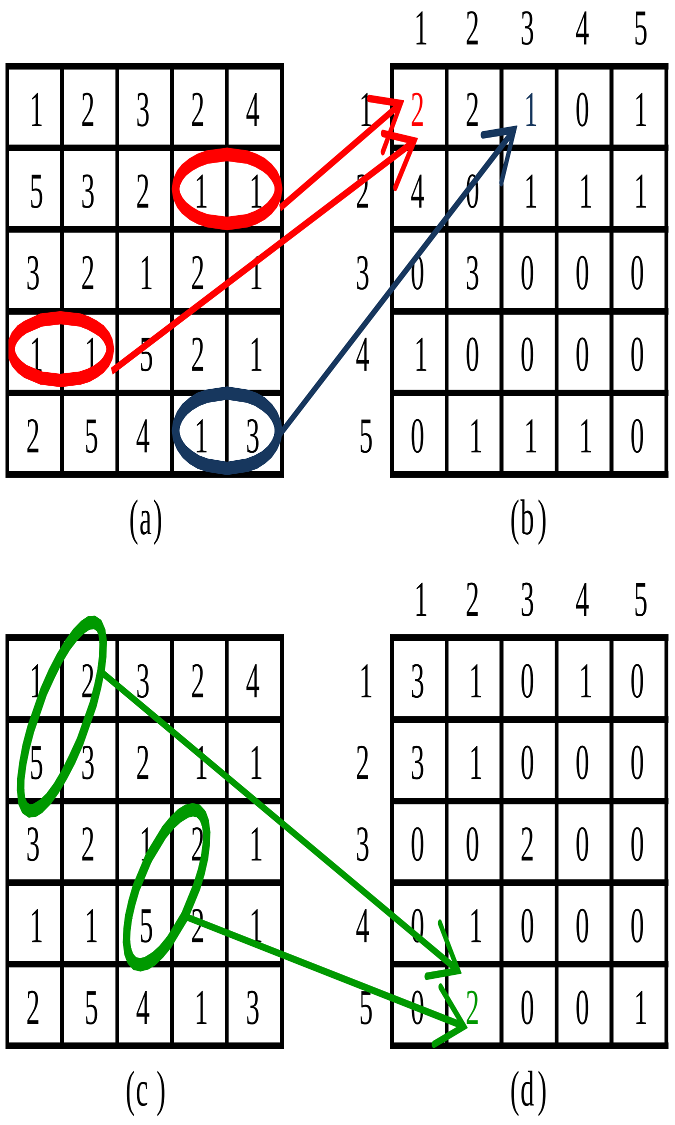
III. TEXTURE ANALYSIS

Combination of repeated patterns with a regular frequency is known as texture. It can be seen in various images, ranging from MRI images, dermatological images, CT scan images etc. Texture of images is useful in a variety of applications and is an important field of study these days. One important application of image texture is the in the field of dermatology.

To analyze the texture in digital images is difficult in terms of mathematics because texture cannot be standardized quantitatively due to huge amount of data. Analysis of image techniques has played an important role in several medical, geographical applications [18] and various other fields. Certain features from filtered images are computed using signal processing methods. The filters commonly include spatial domain filters, Gabor filters, etc.

IV. GRAY LEVEL CO-OCCURRENCE MATRIX

Image texture is one of the important parameter used in identification of various objects or the region of interests in an image. The textural features based on gray level spatial dependencies have applications in image classification. GLCM these days is one of the most widely used texture measurement methods. The Gray Level Co occurrence Matrix is a second order Statistical method. The GLCM is a square matrix. It is a two dimensional array in which rows and columns represent various possible values of image pixels. Figure 1 shows Concept of GLCM Matrix.



(a)

1	2	3	2	4
5	3	2	1	1
3	2	1	2	1
1	1	5	2	1
2	5	4	1	3

(b)

	1	2	3	4	5
1	2	2	1	0	1
2	4	0	1	1	1
3	0	3	0	0	0
4	1	0	0	0	0
5	0	1	1	1	0

(c)

1	2	3	2	4
5	3	2	1	1
3	2	1	2	1
1	1	5	2	1
2	5	4	1	3

(d)

	1	2	3	4	5
1	3	1	0	1	0
2	3	1	0	0	0
3	0	0	2	0	0
4	0	1	0	0	0
5	0	2	0	0	1

Figure 1.(a) 5x5 image matrix (b) GLCM with 0º orientation (c) 5x5 image matrix (d) GLCM with 45º orientation

In Figure 1(a) shows a 5x5 image matrix. Element pair (1, 1) in Figure 1(b) contains the value 2 because there are 2 horizontally adjacent pixels having the value 1 and 1 in image matrix shown in Figure 1(a) encircled in red. Element pair (1, 3) in Figure 1(b) contains the value 1 because there is only one horizontally adjacent pixel having the value 1 and 3 in image matrix shown in Figure 1(b) encircled in red. This is shown for GLCM matrix having 0º orientation. Element pair (5, 2) in Figure 1(d) contains the value 2 because there are two diagonal pixels having the value 5 and 2 in image matrix shown in Figure 1(c) encircled in red. This is shown for GLCM matrix having 45 º orientations. All the values in the matrix are filled using GLCM.

Feature extraction technique is used for extracting various textural features. The colored image is converted into gray scale image which is given as input to the GLCM method.

Correlation, Energy, Contrast and homogeneity are retrieved from GLCM based matrix. Contrast measures the coarseness of texture in an image. Energy measures the image textural uniformity and Homogeneity measures the distribution of elements in an image.

GLCM technique introduced by Haralick is most widely used in image processing to study the gray level intensities present in the image. Application of GLCM includes detection of malignant masses in breast tissue images [24].

Another application of GLCM algorithm is detection of Skin Cancer [19].GLCM Textural Features are used for Brain Tumor Classification [17] and are also used to Enhance Low Contrast Regions of Regenerated Satellite Image Texture[22]. The main application of GLCM is for seismic facies description [18] in which GLCM-based attributes are used in combination with neural networks. In our paper, an application for texture analysis using the GLCM is discussed. Textural features extracted from matrix based on GLCM are Contrast, Correlation Energy, and Homogeneity. Also, mean and standard deviation of each parameter is calculated.

V. METHODOLOGY

In our work, texture analyses for 3 Pyoderma diseases (Boil, Carbuncle and Impetigo Contagiosa) are done using the Gray Level Co-occurrence Matrix (GLCM).

Four different image features (Contrast, Energy, Correlation and Homogeneity) are extracted for three color channels from the given diseased dermatological images. The image texture presents visual patterns that have properties of homogeneity that do not result from the presence of only a single color or intensity.

Regularity and the coarseness of a texture of an image is quantified by the correlation function. Energy returns the sum of squared elements in the GLCM.Homogeneity returns a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal. To calculate the various features, following formulas shown in equations (i) (ii), (iii) and (iv), respectively [22] are used.

$$Contrast = \sum_{i,j} |i-j|^2 \, p(i,j) \qquad \text{(i)}$$

$$Correlation = \sum_{i,j} \frac{(i-\mu i)(j-\mu)p(i,j)}{\sigma_i \ \sigma_j} \qquad \text{(ii)}$$

$$Energy = \sum_{i,j} p(i,j)^2 \qquad \text{(iii)}$$

$$Homogeneity = \sum_{i,j} \frac{1}{1-(i-j)^2} p(i,j) \qquad \text{(iv)}$$

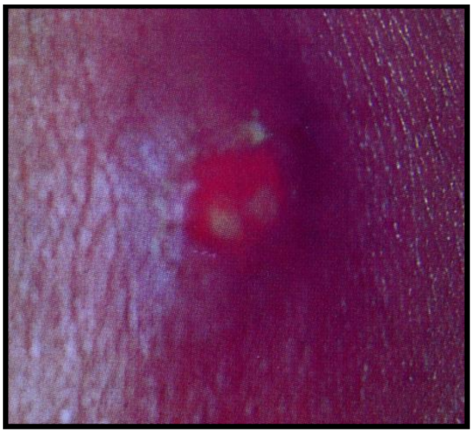
The textural features are tested using input image as shown in Figure 2. (a), Figure 2. (b) and Figure 2. (c) that act as a test images for the experiment in our work. In our work, a very simple algorithm has been developed for three different diseased dermatological images. Textural features (Contrast, Correlation, Energy and Homogeneity) are extracted from the selected 1-dimensional image for three color channels.

The above parameters are independent of orientation and size of image. Mean of Contrast, Energy, Correlation and Homogeneity are calculated. Also, standard deviations of these parameters are found. The histograms show that textural features for individual diseases are different from each other.

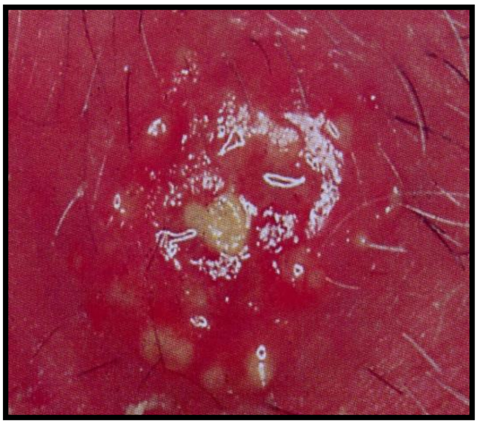
The input images shown used as input in this experiment are taken from Bhutani's Color Atlas of Dermatology. The original images have been resized to 256 pixels x 256 pixels in .jpeg format. The images are separated into 3 basic color components, thus, generating 3 new images having red, green and blue channels respectively.

For each color component image we segmented the entire image into 8 x 8 blocks. To each block of red component image, Gray Level Co-occurrence Matrix is applied and the four features as discussed are obtained for Red component of image.

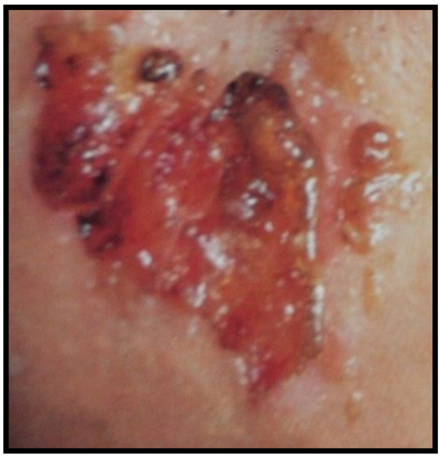
The same process is done for green and blue components of image. The features are found using the above formulas. Mean and standard deviation are calculated for the proposed features in overall processed image. Figure 3 shows complete methodology of the proposed work.



(a)



(b)



(c)

Figure 2. Three Pyoderma diseases taken as test images. (a) Boil (b) Carbuncle (c) Impetigo Contagiosa

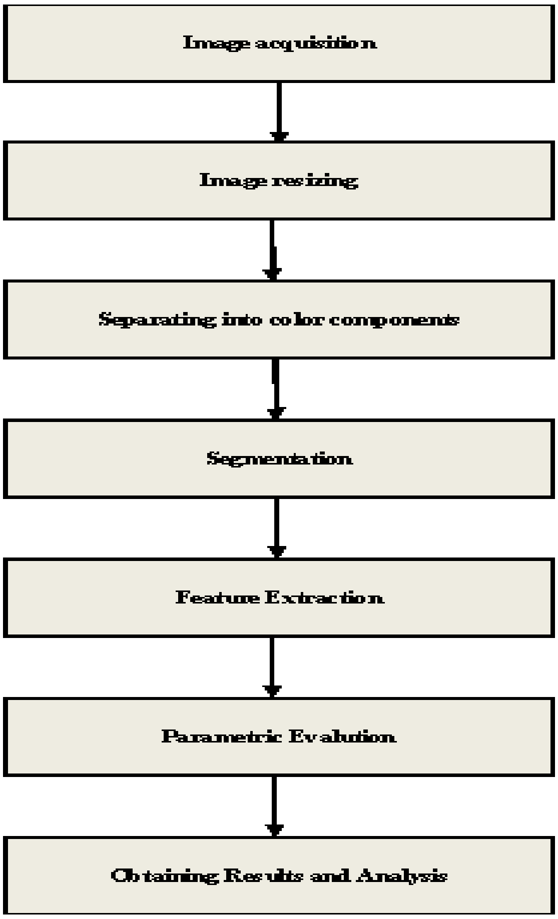


Figure 3. Block Diagram of Proposed Work

VI. RESULTS & DISCUSSIONS

The experiments are done to evaluate four different features which are contrast, energy, correlation and homogeneity from the color components of 3 diseased dermatological images. Here, the category of dermatological images chosen is Pyoderma. Table 1 shows the various values of mean correlation, mean contrast, mean energy and mean homogeneity for the three gray diseased (Boil, Carbuncle, and Impetigo

Contagiosa) dermatological images taken as input. Table 2 shows the standard deviation (S.D.) shows standard deviation for the proposed features in gray scaled input images.

Table 1. Calculated values of mean contrast, mean correlation, mean energy and mean homogeneity for proposed features in gray scaled diseased dermatological images

Diseases	Contrast	Correlation	Energy	Homogeneity
Boil	0.22	0.35	0.57	0.89
Carbuncle	0.36	0.25	0.6	0.87
Impetigo Contagiosa	0.15	0.45	0.67	0.92

Table 2. Standard Deviation for the proposed features in gray scaled diseased dermatological images

Parameters	Diseases		
	Boil	Carbuncle	Impetigo Contagiosa
Contrast	0.17	0.55	0.15
Co-relation	0.36	0.39	0.41
Energy	0.27	0.29	0.28
Homogeneity	0.07	0.11	0.07

Table 3 represents the calculated values of standard deviation and mean correlation for the proposed features in processed images having diseases (Boil, Carbuncle, and Impetigo Contagiosa) Table 4 represents the values of standard deviation and the contrast, i.e., mean calculated for the proposed features in the diseased images. Table 5 represents calculated values of mean energy and standard deviation and table 6 shows the homogeneity which is calculated as mean and standard deviation for the proposed features in Boil, Carbuncle, and Impetigo Contagiosa.

Table 3. Mean contrast and standard deviation calculated for the proposed features in processed images of Pyoderma disease

Diseases	Mean Contrast			Standard Deviation		
	Red	Green	Blue	Red	Green	Blue
Boil	0.19	0.26	0.19	0.18	0.18	0.16
Carbuncle	0.12	0.53	0.5	0.22	0.75	0.77
Impetigo Contagiosa	0.12	0.19	0.2	0.16	0.22	0.21

Table 4. Mean correlation and standard deviation calculated for the proposed features in processed images

Diseases	Mean Co-relation			Standard Deviation		
	Red	Green	Blue	Red	Green	Blue
Boil	0.29	0.29	0.35	0.4	0.31	0.41
Carbuncle	0.47	0.19	0.15	0.5	0.31	0.32
Impetigo Contagiosa	0.58	0.48	0.45	0.43	0.4	0.39

Table 5. Mean energy and standard deviation calculated for the proposed features in processed images of Pyoderma disease

Diseases	Mean Energy			Standard Deviation		
	Red	Green	Blue	Red	Green	Blue
Boil	0.67	0.53	0.64	0.28	0.24	0.29

Carbuncle	0.82	0.49	0.53	0.25	0.28	0.28
Impetigo Contagiosa	0.77	0.65	0.64	0.28	0.3	0.29

Table 6. Mean homogeneity and standard deviation calculated for the proposed features in processed images of Pyoderma disease

Diseases	Mean Homogeneity			Standard Deviation		
	Red	Green	Blue	Red	Green	Blue
Boil	0.91	0.88	0.91	0.09	0.07	0.08
Carbuncle	0.95	0.84	0.85	0.08	0.11	0.12
Impetigo Contagiosa	0.94	0.91	0.91	0.07	0.09	0.09

Finally, the graphs for color components of input and processed output image are studied. Figure 4 shows the histogram for contrast, correlation, energy and homogeneity of gray scale images for Boil, Carbuncle and Impetigo Contagiosa diseases. Also, the standard deviation for them is shown in the same histogram while Figures. 5(a), 5(b), 5(c) and 5(d) shows the mean contrast and standard deviation obtained for the individual color components of the processed images for the three selected diseases from Pyoderma category.

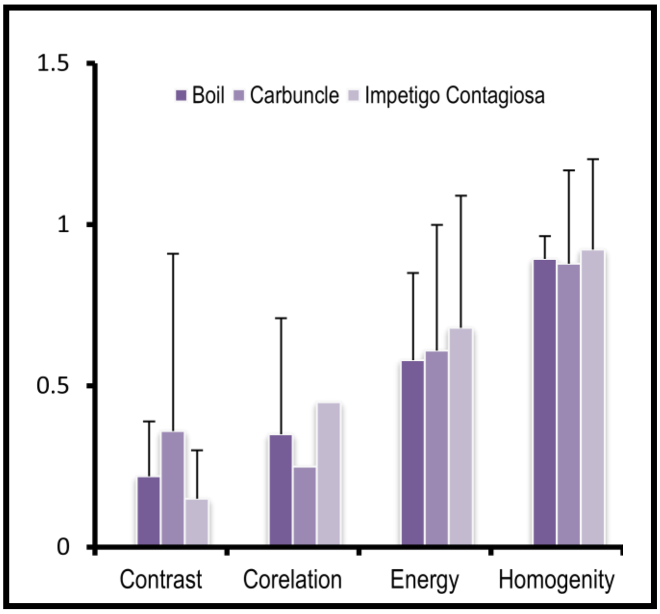
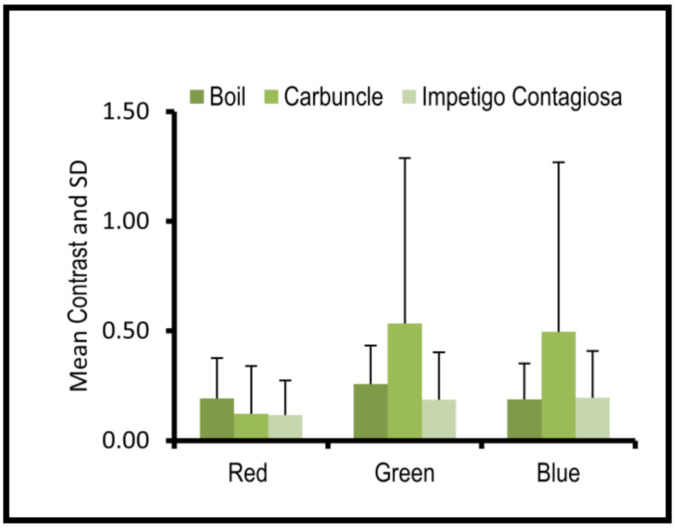
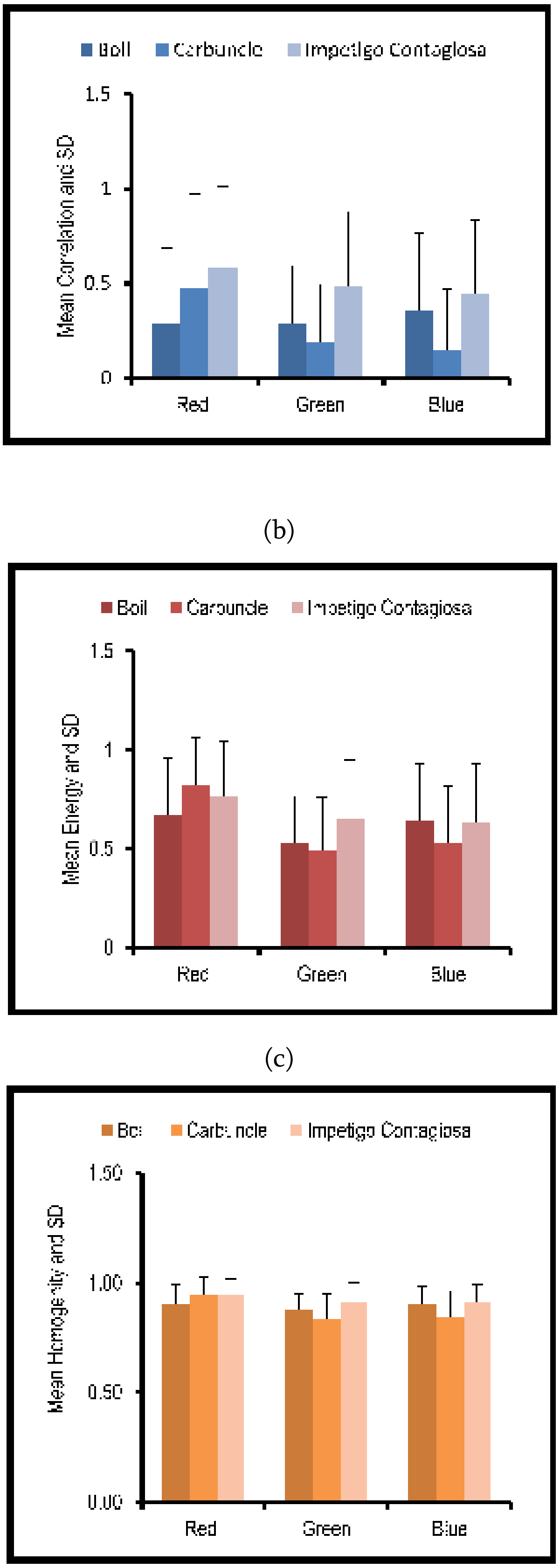


Figure 4 . Histogram for contrast, correlation, energy and homogeneity of gray scale images for Boil, Carbuncle and Impetigo Contagiosa diseases



(a)



(b)

(c)

(d)

Figure 5. Histogram obtained for 3 types of Pyoderma diseases (Boil, Carbuncle and Impetigo Contagiosa) (a) mean contrast and standard deviation (b) mean correlation and standard deviation (c) mean energy and standard deviation (d) mean homogeneity and standard deviation of processed dermatological images

VII. CONCLUSION AND FUTURE WORK

From the above histograms, it is concluded that texture analysis for 3 Pyoderma diseases (Boil, Carbuncle, and Impetigo Contagiosa) is done effectively using GLCM. Figure 6 shows the histogram for contrast, correlation, energy and homogeneity of gray scale images for Boil, Carbuncle and Impetigo Contagiosa diseases. Also, the standard deviation for them is shown in the same histogram Four different image features are extracted for three color channels from the given input dermatological images. Mean of Contrast, Correlation, Energy and Homogeneity are calculated. Also, standard deviations of these parameters are found. The histograms show that textural features for individual diseases are different from each other. Hence, it shows promising results that different diseases can be classified and identified into separate categories using machine learning algorithms. In our future work, we will use Gaussian Mixture Model for classification and identification of various categories of dermatological diseases.

VIII. REFERENCES

[1]. P.Sangamithraa and S. Govindaraju, "Lung Tumour Detection and Classification using EK-Mean-Clustering," WISPNET conference", in 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 2016, pp. 2201-2206.

[2]. A. Santy and R. Joseph, "Segmentation methods for computer aided melanoma detection", in

2015 Global Conference on Communication Technologies (GCCT), 2015, pp. 490-493.

[3]. R. Aswin, J. Jaleel and S. Salim, "Hybrid Genetic Algorithm - Artificial Neural Network Classifier for Skin Cancer Detection", in International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) ,IEEE, 2014, pp. 1304-1309.

[4]. M. Hassan, M. Hossny, S. Nahavandi and A. Yazdabadi, "Skin lesion segmentation using Gray Level Co-occurance Matrix", in IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, Hungary, 2016, pp. 820-825.

[5]. S. Kaya,M. Bayraktar and S. Kockara, "Abrupt skin lesion border cutoff measurement for malignancy detection in dermoscopy images," 13th Annual MCBIOS conference Memphis, IEEE, 2016.

[6]. F. Nachbar, W. Stolz and P. Bilek," The ABCD rule of dermatoscopy High prospective value in the diagnosis of doubtful melanocytic skin lesions," AcadDermatol, pp. 551-559,1994.

[7]. Md. Alam, T.T Khan Munia and K. Tavakolian," Automatic Detection and Severity measurement of Eczema Using Image Processing ," IEEE, pp.1365-1368, 2016.

[8]. S. Simonthomas and N. Thulasi, "Automated Diagnosis of Glaucoma using Haralick Texture Features," ICICES, IEEE, pp.2014

[9]. T. Tan, Li Zhang and Ming Jiang, "An Intelligent Decision Support System for Skin Cancer Detection from Dermoscopic Images," 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, IEEE, pp. 2194-2199, 2016.

[10]. U. Fidan, N.Ozkan and I. Calikusu, "Clustering and Classification of Dermatologic Data with Self Organization Map (SOM) Method,"IEEE, 2016.

[11]. R. Yasir, Md. Ashiqur and N. Ahmed ,"Dermatological Disease Detection using Image Processing and Artificial Neural Network," 8th International Conference on Electrical and Computer Engineering, IEEE,pp.687-690, 2014.

[12]. V. Bannihatti, S. Kumar and V. Saboo, "Dermatological Disease Detection Using Image Processing and Machine Learning," IEEE, pp.88-93, 2016.

[13]. K. K Jyothilakshmi and J. B. Jeeva," Detection of Malignant Skin Diseases Based on the Lesion Segmentation," International Conference on Communication and Signal Processing, IEEE, pp.382-386, 2014.

[14]. R. Maurya, S. Singh, Ashish K. Maurya and Ajeet Kumar, "GLCM and Multi Class Support Vector Machine based Automated Skin Cancer Classification," IEEE, pp.444-447, 2014.

[15]. W. Paja,"Medical Diagnosis Support and Accuracy Improvement by Application of Total Scoring from Feature Selection Approach," Proceedings of the Federated Conference on Computer Science and Information Systems, IEEE, pp. 281–286, 2015.

[16]. J. Premaladha and K. S. Ravichandran," Novel Approaches for Diagnosing Melanoma Skin Lesions Through Supervised and Deep Learning Algorithms," Springer, 2016.

[17]. M. Arabi, "Performance evaluation of GLCM and pixel intensity matrix for skin texture analysis", Perspectives in Science, Elsevier, April, Vol.8, pp.203-206, 2016.

[18]. S. Cheng-Yun, Pre-stack-texture-based reservoir characteristics and seismic facies analysis", Applied Geophysics, Vol.13(1), pp. 69-79,March, 2016.

[19]. M.K. Soumya, "Cervical Cancer Detection and Classification Using Texture Analysis", Biomedical & Pharmacology Journal, Vol. 9(2), pp. 663-671, April, 2016.

[20]. X. Shi, "Research on the Multiple Feature Fusion Image Retrieval Algorithm based on

Texture Feature and Rough Set Theory",5th International Conference on Advanced Materials and Computer Science, pp.288-292, 2016.

[21]. W. Zha, "Texture Attribute Analysis of GPR Data for Archaeological Prospection", Pure Appl. Geophys,Springer, Vol.173, pp.2737–2751, 2016.

[22]. Y.Dong, J.Pan, "Medical Image Categorization based on Gaussian Mixture Model", in Proc. of IEEE International Conference on BioMedical Engineering and Informatics, pp. 128-131, 2008.

[23]. S. Parveen, "Texture Analysis Using Local Ternary Pattern For Face Anti-spoofing", Vol.28(2), pp.965-971, 2016.

[24]. M. Nasser,"The Impact of Pixel Resolution, Integration Scale, Preprocessing and Feature Normalization on Texture Analysis for Mass Classification in Mammograms", 2016.

[25]. J. Luis,"Pattern recognition of lower member skin ulcers in medical images with Machine Learning Algorithms", pp.50-53, 2016.

[26]. S. Jain,"Brain Cancer Classification Using GLCM Based Feature Extraction in Artificial Neural Network", International Journal of Computer Science & Engineering Technology, Vol.9, pp.966-970, 2013.

[27]. L. Nanni,"Different Approaches for Extracting Information from the Co-Occurrence Matrix", Vol. 8, pp. 1-9, 2013.

[28]. Lavania, "Image Enhancement using Filtering Techniques International Journal on Computer Science and Engineering", Vol. 4, No. 1, pp. 14-20, 2012.

[29]. A.R.Rivera,"Local directional number pattern for face analysis: face and expression recognition,"IEEE Transactions on Image Processing, Vol. 22, No. 5, pp. 1740–1752, 2013.

[30]. J. Melendez,"Improving mass candidate detection in mammograms via feature maxima propagation and local feature selection," Medical Physics, Vol. 41, No. 8, 2014.

[31]. P. Mohanaiah "Image Texture Feature Extraction Using GLCM Approach", Vol.3, International Journal of Scientific and Research Publications, 2013.

[32]. C. Mosquera-Lopez, "Computer-Aided Prostate Cancer Diagnosis from Digitized Histopathology: A Review on Texture-Based Systems", IEEE Reviews In Biomedical Engineering, Vol.8, P.98-113, 2015.

[33]. D. Gutman, N.C. Codella, E. Celebi, B.Helba and A. Halpern, "Skin Lesion Analysis toward Melanoma Detection: A Challenge", at the International Symposium on Biomedical Imaging (ISBI) hosted by the International Skin Imaging Collaboration (ISIC), 2016.

[34]. E. Flores and J.Scharcanski, "Segmentation of melanocytic skin lesions using feature learning and dictionaries , Expert Systems with Applications," Vol.56, pp.300-309, 2016.

[35]. K. Goswami, "Fast algorithm for the High Efficiency Video Coding (HEVC) encoder using texture analysis", Information Sciences, pp. 72–90, Elsevier,2016.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Enhanced EEG-Based Emotion Detection Technique using Deep Belief Network and Wavelet Transform

Sahar Jodat, Khosrow Amirizadeh

ABSTRACT

Today's, the role of emotion in communication , brain-computer interface, brain diseases and mental states, car driver monitoring and recommendation systems is proven. Therefore, automatic emotions detection has become one of the most challenging issue. Until now, numerous studies have been addressed different technique on improving automatic emotion detection.In this study, to achieve bether validation in classification of emotion by EEG signals, we combined wavelet transform with deep belief network. For, non-stationary and time-varying are the most important properties of EEG signals, we decided to use discrete wavelet transform (sym8) for extracting features such as power, then applied deep belief network as a classifier to classify emotions according to two-dimensional arousal-valence model. To examine the effectiveness of the method, we used DEAP database and mapped different emotions on two different classes of valence and arousal. Final results show an acceptable enhancement with the accuracy of 75.52% and 81.03% for valence and arousal, respectively.

Keywords: EEG signals, Discrete Wavelet Transform, Deep Belief Network, two-dimensional arousal-valence model, DEAP

I. INTRODUCTION

Emotions express mental state of the mind and thought process that can be perceived conscious or unconscious in different situations. Introducing different methods for processing signals, easy usage of ellectrodes in collecting data, various classification methods and real-world applications of computer and human interaction for normal people have provided the possibility of different emotion detection by intelligent devices. Generally, there is three different approaches in this case. In the first method, emotions are classified according to analysis of facial expressions or speech [2-4]. The second method take the periphery physiological signals, such as electrocardiogram (ECG), skin conductance (SC), respiration and pulse into account in classification[5-7]. In the last method, brain signals captured from central nervous system such as electroencephalograph (EEG), electrocorticography (ECOG) and functional magnetic resonance imaging (fMRI) become the seat of researchers attention [8-10]. As modern equipment such as electrodes provide collecting EEG signals easily, among all of mentioned methods recognition by EEG signals is a trustable method as these signals contain information of central neural system related to brain activities and have short time answering in detection. EEG signals reflect brain activities and can be acquired by electrodes according to 10-20 system.

In 1949, international standard of 10-20 [11] was introduced to determine the place of electrodes on the scalp. This method provides the possibility of comparing the results of the recording and processing of brain signals of diffierent people at any time, illustrated in Figure 2.

One of the problems with the classification of emotions and their naming is that the distinction between the boundaries of different emotions is not clear, since different individuals express their feelings differently, modeling emotions seems to be difficult. To tackle this problem, researchers have used two different methods for emotion modeling. The first method of emotion modeling is to consider them as separate and discrete senses.The second method is to consider feelings in a multidimensional space. The discrete model is considered as a complete set for describing emotions, such as happiness, sadness, surprise, anger, fear, disgust and the rest of the emotions are derived from the basic emotions. But the main problem with this type of model is that how many and which of the emotions are chosen as the main prototype. For example, Weiner considered only happiness and sadness as basic emotions [12], whereas Kemper suggested fear, anger, depression and satisfactions to be basic [13]. To overcome these problems, multiple dimensional or scale to categorize emotions become popular. For example, Russell describe two dimensional (2D) model for specified emotions by their position [1]. Two dimensional model described by two axes of valence and arousal. Valence represents Positive or negative emotional state of the individuals or, in other words, the rate of pleasure or unpleasantness (horizontal vector). Arousal refers the degree of excitement that a person feels generally, changes from calm to excitement (vertical vector), shown in Figure 3. Different emotions can be labeled in various positions in 2D models. In addition to the 2D model, in some models, the third dimension, is also considered, which represents the degree of dominance and its range from weak to strong.

As automatic emotion detection system can be used in real world and real-time applications, improving the accuracy is an important issue in this case. Until now, a large number of classifiers have been used in this field. For example, support vector machine (SVM), neural network(NN), k-nearest neighbor (KNN), deep belief network (DBN) and so on. Considering this point into account that, deep learning algorithm is capable to represent and classify a set of data in hence of their hierarchical structure and provides comprehensive presentation in comparison with shallow structures, we selected DBN among other classifiers. Then, chose discrete wavelet transform, for extracting statistical features like, power. In this study we used the data of DEAP1 dataset for testing the accuracy of proposed model [16]. In this dataset (is explained in the next section completely), two kinds of data exist, raw data and processed data. We applied processed data that down-sampled (to 128 Hz) and EOG artefacts are removed [17]. The data are labeled in 4 categories, valence, arousal, dominance and liking. We used just valence and arousal dimensions for categorize emotions. We achieved the accuracy of 75.52% for valence and 81.03% for arousal. The final results show acceptable progress compared with other experiments that are explained in section 4. Different steps in our experiment is illustrated in Figure 1.

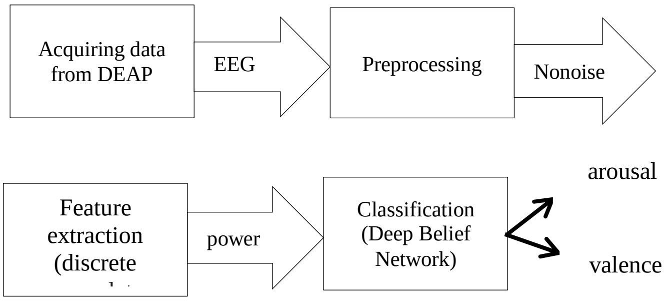


Figure 1. Different steps in emotion detection from EEG

The paper has divided into 5 sections. In section 2, we explained basic contents of applied classification method. Then we analyzed our different part of methodology in section 3. Section 4 is allocated to final results and comparison them with other methods. In last section summarized our work and represent our suggestion for future works.

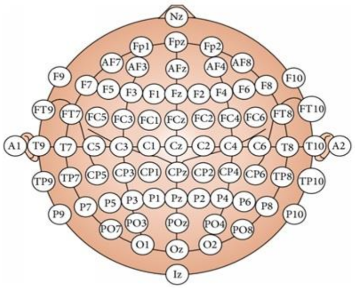

Nz
Fp1 Fpz Fp2
AF7 AF3 AFz AF4 AF8
F9 F7 F5 F3 F1 Fz F2 F4 F6 F8 F10
FT9 FT7 FC5 FC3 FC1 FCz FC2 FC4 FC6 FT8 FT10
A1 T9 T7 C5 C3 C1 Cz C2 C4 C6 T8 T10 A2
TP9 TP7 CP5 CP3 CP1 CPz CP2 CP4 CP6 TP8 TP10
P9 P7 P5 P3 P1 Pz P2 P4 P6 P8 P10
PO7 PO3 POz PO4 PO8
O1 Oz O2
Iz

Figure 2. 10-20 international standard of electrodes placement [11]

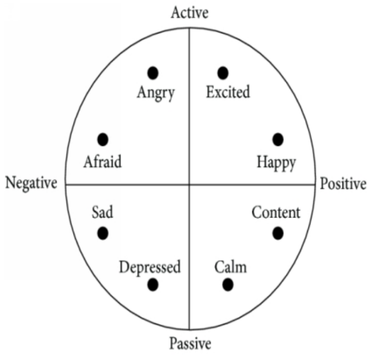


Figure 3. Two-dimensional emotion model [1]

II. DEEP BELIEF NETWORK (DBN)

Neural networks are the base of deep learning. These networks are formed of an input layer, one hidden layer and an output layer to model the human nerves system. Unfortunately, in implementing complicated model, with large number of nodes neural networks do not have goof performance. For solve this problem, deep neural networks are replaced. Generally, deep neural networks consist of a series of shallow networks (such as single-layer neural networks) that enable networks to learn and extract nonlinear hierarchical features. This feature has led to more automatic recognition methods towards deep learning methods. Although, using deep neural network is beneficial, but it has a main problem. Training all layers at once is difficult, with random initialization of weights, they do not converge to the correct answer and it increase the probability of get stuck in local minimum [19]. To overcome the difficulties, Hinton et al. proposed deep belief network. In fact, deep belief network is made up of few layers of the restricted Boltzmann machine (RBMs) [20]. RBMs stacked together with shared layers create a DBN and trained layer by layer in a greedy way [21], [22]. The process of training a deep belief network has two phases .The first step is the unsupervised pre-training, in which unlabelled data is used for training. The training starts from the lowest layer of the network (the first layer) and features are derived from raw input data. Then the training takes moves up to higher level (between the hidden nodes of the first layer and the second layer). The training of the hidden nodes of the first layer is a new input for obtaining the features in the second layer's hidden nodes. Greedy training continues to reach the topmost layers of hidden nodes. Finally, a productive model with weights between layers trained by using input data features. The greedy layer-to-layer training method will calculate the weights and biases of different layers. Fine-tuning weights and supervised learning are performed in the second phase of training at the upper layer. In this phase a new label layer will be added to the upper layer of the deep belief network and removes all links in the top-down direction. Now, the DBN becomes a feed-forward neural network, shown in Figure 4(b).Then the backward algorithm is used to learn the weights and biases that are trained based on labels. The goal of learning is to reduce the classification error from labeled examples. Weights and biases are initialized in the non-supervision training phase, except those that are randomly assigned in the upper layer. Figure 4 illustraites the structure of a DBN with three hidden layers.

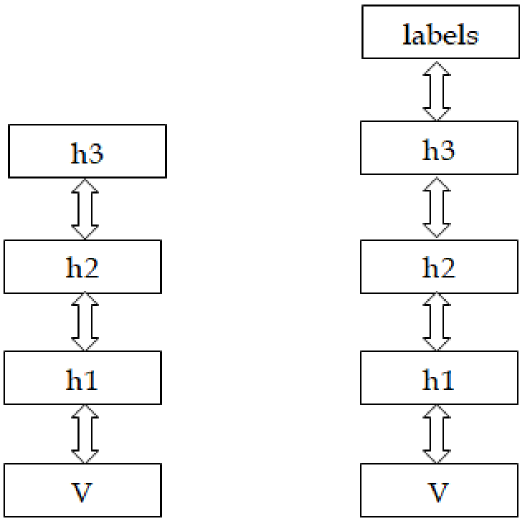


Figure 4. The structure of a DBN with three hidden layers: (a) The pre-training stage with un-labeled data and (b) The fine-tuning stage using the new layer added at the top of the network that is trained by labeled data[19].

2.1 Restricted Blotzmann Machine (RBM)

Restricted Boltzmann machine is an energy based generative model with two binary layers (visible and hidden). In other words, restricted Boltzmann machine is a two-part, weighted, non-directional, symmetrical graphical model that takes random decisions about the status of the nodes, whether on or off. A graphical model of an RBM is shown in Figure 5.

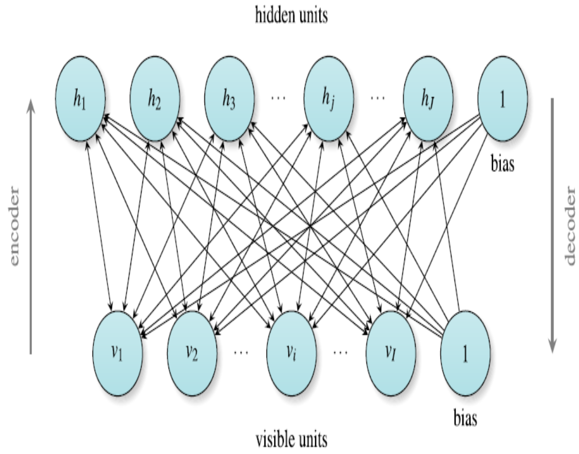


Figure 5. Restricted Boltzman machine [24]

The energy function for connecting visible layer to hidden layer obtain from the following equation [24] :

$$E(v\,,h) = -\sum_{i=1}^{I}\sum_{j=1}^{J} w_{ij}v_ih_j - \sum_{i=1}^{I} a_iv_i - \sum_{j=1}^{J} b_jh_j \qquad (1)$$

Where w_{ij} is the weight between visible unit i and hidden units j, a_i and b_j refer to biases in visible (v) and hidden (h) layer. The probability of given configuration is the normalized energy function [24] :

$$p(v,h) = \frac{e^{-E(v,h)}}{\sum_{u,g} e^{-E(u,g)}} \qquad (2)$$

Since there is no direct connection between hidden units in an RBM, these units considered independent according to visible units. The binary state h_j of each hidden units j will be equal one (activated) with the below conditional probabilities :

$$P(h_j = 1|v) = g(b_j + \sum_i v_iw_{ij}) \qquad (3)$$

Where g(x) is the logistic sigmoid function $g(x) = \frac{1}{1+e^{-x}}$.

Similarly, there is no direct connection between visible units in an RBM, with having a hidden vector, it will be easy to calculate unbiased the state of a visible unit (activated).

$$P(v_i = 1|h) = g(a_i + \sum_j h_jw_{ij}) \qquad (4)$$

Restricted Boltzmann machines are trained to maximize the product of probabilities of a set of training examples X:

$$\mathrm{argmax}_W \prod_{x\in X} P(x) \qquad (5)$$

or equivalently to maximize the log likelihood

$$\mathrm{argmax}_W \sum_{x\in X} \log P(x) \qquad (\ 6)$$

Unfortunately, calculating the gradient of the log likelihood is so difficult. Therefore, [22] proposed contrastive divergence (CD) by doing k iterations of Gibbs sampling to approximate it. Also, using CD method, enable an RBM to update weights according to Equation 7.

$$\Delta w_{ij} = \varepsilon(< v_i h_j >^0 - < v_i h_j >^k) \qquad (7)$$

where $<\cdot>^m$ is the average in a contrastive divergence iteration m and ε is the learning rate.

2.2 Contrastive divergence

To solve the problem of calculating the log-likelihood gradient, Hinton proposed a contrastive divergence method in 2002. In this method, the state of visible units is initialized to training data. Then the binary state of the hidden units is calculated according to Equation 3. After the binary state of the hidden units is computed, the values of v_i will be update based on Equation 4. At the end, again, the probability of activation the hidden units is computed and the value of $<\cdot>^m$ will be calculated from the final values obtained from the hidden and visible units.

2.3 Softmax classifier

The softmax classifier is used to estimate the probability of output values in a deep belief network. The method of this type of classification is to learn all the parameters of weights and biases using the featured learned from the last hidden layer. In the case of binary classification (k = 2), the softmax regression hypothesis output $h_\theta(x)$ is obtained from the following Equation :

$$h_\theta(x) = \frac{1}{e^{\theta_1^T x} + e^{\theta_2^T x^{(i)}}} \begin{bmatrix} e^{\theta_1^T x} \\ e^{\theta_2^T x} \end{bmatrix} \qquad (8)$$

Softmax classifier can be generalized to be multiclass classification .The hypothesis will output a vector of k estimated probabilities, shown as follows:

$$h_\theta(x) = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \begin{bmatrix} e^{\theta_j^T x^{(i)}} \\ e^{\theta_j^T x^{(i)}} \\ \cdot \\ \cdot \\ \cdot \\ e^{\theta_j^T x^{(i)}} \end{bmatrix} \qquad (9)$$

The softmax layer needs to learn the weight and bias parameters with supervised learning approach by minimizing its cost function, shown as follows:

$$\text{cost} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y_i = j\}\log\frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_j^T x^{(i)}}} + \frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=1}^{n}\theta_{ij}^2 \qquad (10)$$

where m is number of hidden units, n is number of inputs, k is number of classes, y is ground truth, and θ is weight of hidden nodes.

III. METHODOLOGY

3.1 DEAP Dataset

DEAP [16] is a a multi-model database designed to provide signals for emotion detection. In this database, video clips are used as visual stimulus for stimulation different emotions. This database contains a set of brain signals that are used to analyze emotions. The way of collecting information in this database is that in this way, 40 pre-selected video clips, each one for one minute, are displayed as emotional stimulus for 32 participants between 19 and 37 years old, of which 50% is female, and EEG signals and other fuzzy signal signals, such as ECG, EMG GSR and BVP, are collected from 40 channels during video viewing.

The ordering of videos is based on the code number of the test, not on the order in which it is displayed, which means that the first video clip is the same for each participant. It should be noted that the electrodes are arranged according to the standard 10-20. After the end of each clip, participants rate it according to arousal, valence, liking or not, level of dominance and familiarity.

The format of the data is 40*40*8064 that represent the concept of video/trial*channel*data, similarly the format of the labels are 40*4 (valence, arousal, dominance, liking).

Self-assessment manikins (SAM) [26], as shown in Figure 6, were used to visualize the scales. The scales between 1 and 9 for 2 different levels of valence and arousal are mapped in the order below. For valence dimension the numbers between 1 and 3 represents negative emotions, numbers between 4 to 6 represent neutral feelings and numbers between 7 to 9 represent positive emotions, while in the dimension of the arousal the numbers between 1 and 3, represent the inactive emotions, the numbers Between 4 to 6 neutral feelings and numbers 7 to 9 represent active emotions

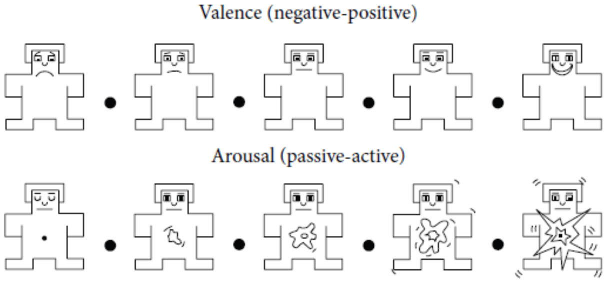


Figure 6. An example of a self-assessment. The first line above indicates the feelings of valence and the second line is allocated to arousal [26]

3.2 Channel selection

In order to reduce the number of the EEG channels as much as possible and implement, an emotion recognition method that would result in a more user-friendly environment in the future, the signals were acquired from fifteen positions only, according to the 10–20 system. According to essay [27] we chose the EEG signals recorded at positions AF3, F7, C3, T7, CP5, Pz, AF4, F8 FC6, FC2, CZ, C4, T8, CP2, O2. IN addition, the left frontal area is involved in the experience of positive emotions (high values of valence), such as joy or happiness (the experience of positive affect facilitates and maintains approach behaviors), whereas the right frontal region is involved in the experience of negative emotions (lower valence values), such as fear or disgust (the experience of negative affect facilitates and maintains withdrawal behaviors) [28].

3.3 Feature extraxtion

A wavelet transform is a variable-length window technique that uses a time-scale domain. A wavelet is a definite function with a mean of zero and with a limited period, and expansion is carried out based on transformation and scale. A wavelet transform with a multi-resolution analysis feature is suitable for analyzing the signal at different time and frequency bands. In fact, a wavelet is a mathematical transformation function that divides the signal into different frequency bands. Wavelet transform is the representation of a function by mother wavelets. ($\psi_{a,b}$, the mother wavelet).

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi(\frac{t-b}{a}) \tag{11}$$

Where a,b ∈ R (a>0), R is the wavelet space. Parameter 'a' is the scaling factor and 'b' is the shifting factor. The only limitation for choosing a prototype function as mother wavelet is to satisfy the admissibility condition.

The time-frequency representation is performed by repeatedly filtering the signal with a pair of filters namely high pass filter (H(n)) and low pass filter (L(n)), that cut the frequency domain in the middle. Specifically, the discrete wavelet transform decomposes the signal into an approximation coefficients (CA) and detailed coefficients(CD). The approximation coefficient is subsequently divided into new approximation and detailed coefficients. This process is carried out iteratively producing a set of approximation coefficients and detail coefficients at different levels or scales [29].

In this work, the multiresolution analysis of wavelet functions, namely sym8 was used to decompose the EEG signals into five different frequency bands delta(0-4)Hz, theta(4-8)Hz, alpha(8-12)Hz, beta(12-30)Hz and gamma(>30)Hz that the characteristics of each band can be utilized to estimate subject's cognition and emotion states. This wavelet functions was chosen due to it is near optimal time-frequency localization properties. Moreover, the waveforms of this wavelet was similar to the waveforms to be detected in the EEG signal. In order to analyze the characteristic natures of different EEG patterns, we derived linear feature (power). This feature was derived from the five frequency bands of EEG and was concatenated to form a feature vector. Table 1.

Table 1. Statistical feature used for emotion recognition and it's description

Feature	Formula	Description
Power	$\mathrm{P_j} = \frac{1}{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{N}}(\mathrm{d_j(k)^2})$ $\mathrm{d_j(k)}$ is the detail wavelet coefficient	Measure the squares of amplitude of EEG signals

J = decomposition level; k = No. of wavelet coefficient, varies from 1 to N

3.4 Classification

In this study, proposed DBN contains a layer as an input, three hidden layers with two softmax classifiers in output layers, one for valence and another for arousal. Here the data is divided in two parts 90 % of data is used for train the system and remaining 10% data is used for testing the data. The DBN uses unsupervised pre-training technique with greedy layer-wise training, starting from the input layer to the softmax layer. The first hidden layer is trained on the inputs' features extracted from data to learn the primary features of first hidden layer on these input features. Subsequently, the algorithm performs forward propagation by using the input features into this trained hidden layers to obtain the primary feature activations. The features, deriving from feedforward propagation of the 1st hidden layer, must be used to perform unsupervised pretraining in the second hidden layer. The algorithm computes its features in the same procedure from the learned features from the previous hidden layers. The weight and bias parameters of the softmax layer are trained by using a supervised learning approach. The output features of the last hidden layer are used as the input features of both softmax layers. We used a set of self-assessment emotion states (valence and arousal) of subjects as a ground truth. These softmax layers can be trained as the parameters concurrently. After the network finishes learning weight and bias parameters in both softmax classifiers, the algorithm has to perform fine-tuning of all weight and bias parameters in the whole network simultaneously. However, we are not able to use the same network parameters for two classifiers. We need to save the learned parameter outcomes of unsupervised pretraining and load the parameters for fine-tuning process of another softmax classifier. The fine-tuning process treats all layers of a stacked hidden layers and softmax layer as a single model and improves all the weights of all layers in the network by using backpropagation technique with supervised approach. The backpropagation process is used to learn the network weights and biases based on labeled training examples to minimize the classification errors. For evaluating the accuracy of the proposed system cross validation with six repetition has been done. We implemented the proposed model with DeeBNet toolbox [23], and for setting primary parameters performed like Hinton essay [30].

IV. RESULTS AND DISCUSSION

The proposed model is tested to classify EEG signals from DEAP dataset. The signals are collected from 15 channels out of 40. Whole data are divided into test and training set and the cross validation is used to validate the performance of classification results. Finally, the average of all (32) participants' classification accuracies was investigated. We categorized each emotional valence and arousal into two-level class according to the SAM ratings values on a scale of 1–9 [33].Table 2. The results for valence and arousal is shown in Table 3.

Table 2. SAM rating for each emotion class- the conditions for categorizing the emotional class levels

	Two-level class	
	High	Low
SAM rating (S_R)	$S_R \geq 5$	$S_R < 5$

Table 3. Arousal and Valence classification accuracy (%)

Subjects	Valence	Arousal
S01	70.83	70.83
S04	91.67	83.33
S14	66.67	87.5
S15	79.17	75
S16	83.33	79.17
S24	83.33	95.83
S25	79.17	87.5
S26	62.5	66.67
S27	66.67	79.17
S28	87.5	79.17
S31	79.17	87.8
S32	62.5	83.33

According to the Table 3 the best result is for subject 04 with 91.67% in valence and 95.83% for subject 24 in arousal. However, the average of final results for 32 participants were 75.52% and 81.03%, in a row. In continues, to show that the suggested model has better accuracy, at first we compared the final results with the results in essay [18]. In this paper, the validity of valence and arousal were analyzed for 10 subjects separately. In the experiment, six statistical features such as mean, standard deviations, means of the absolute values of the first differences of the raw signals, means of the absolute values of the first differences of the normalized signals, means of the absolute values of the second differences of the raw signals, means of the absolute values of the second differences of the normalized signals are calculated. In addition, fractal dimension (FD) values are calculated and support vector machine classifier with polynomial kernel is used for classification. Researchers tested the performance of their proposed method by using the data of 10 subjects from DEAP database. In classification according to arousal dimensional they used the combination of arousal-dominance or high and low dominance states.

They propose a novel subject-dependent valence level recognition algorithm and apply it to recognize up to 16 emotions where 4 levels of valence are identified with each of the four arousal-dominance combinations, and to recognize up to 9 levels of valence states with controlled dominance level (high or low). In the proposed emotions recognition algorithm, first, four classes of combinations of high/low dominance and high/low arousal levels or two classes of high/low dominance are recognized.

The resulting accuracy using SVM for four arousal-dominance combinations is shown in Table.4. As we can see from the table, the best and worst accuracy obtained in recognition of four arousal-dominance combinations are 80.50% for subject 13 and 46.67% for subject 14, while in our study, the highest and lowest accuracy was 95.83% for subject 24 and 66.67% for subject 25. In addition, the average accuracy across all subjects was 63.04% which contrast highly with 81.03% for our proposed model.

Table 4. Arousal-Dominance recognition accuracy (%)

Subjects	Arousal
S01	63.25
S05	53.08
S07	74.17
S10	65.21
S13	80.50
S14	46.67
S16	67.12
S19	67.29
S20	58.41
S22	49.72
Avg	63.04

However, in the same essay researchers evaluated the performance of valence by six statistical features (mentioned above) and fractal dimension, received the accuracy approximately 50%. Table5 According to the information of the Table 5, fractal dimension features give better accuracy compared with other features in classification with SVM. The average value of accuracy for this feature are around 50%. According to Table 3, in our model the lowest value of accuracy is 62.5% that even is higher than the average in Table5. In our method, the rest of value of accuracy are more than 70%.

Table 5. Valence recognition accuracy (%), $\overline{\Delta\mu_x}$ (mean), $\overline{\Delta\sigma_x}$ (standard deviations), $\overline{\Delta\delta_x}$ (means of the absolute values of the first differences of the raw signals), $\overline{\overline{\Delta\delta_x}}$ (means of the absolute values of the first differences of the normalized), $\overline{\Delta\gamma_x}$(means of the absolute values of the second differences of the raw signals), $\overline{\overline{\Delta\gamma_x}}$(means of the absolute values of the second differences of the normalized Signals)

subjct	$\overline{\Delta\mu_x}$	$\overline{\Delta\sigma_x}$	$\overline{\Delta\delta_x}$	$\overline{\overline{\Delta\delta_x}}$	$\overline{\Delta\gamma_x}$	$\overline{\overline{\Delta\gamma_x}}$	Fractal dimension
S01	50.00	45.16	63.71	53.23	55.65	50.8	45.16
S05	38.31	46.77	57.66	64.92	50.40	56.8	48.79
S07	49.19	58.06	50.00	47.58	45.97	50.8	56.45
S10	50.00	51.61	55.65	54.84	54.03	45.9	65.32
S13	42.74	35.48	50.00	42.74	53.23	47.5	34.68
S14	42.74	47.58	40.32	54.03	44.35	57.2	54.03
S16	48.39	47.58	37.10	54.42	42.74	49.1	54.84
S19	48.39	54.84	55.65	51.61	50.40	54.4	50.00
S20	41.94	43.55	46.77	41.13	45.97	44.3	52.42
S22	49.19	49.19	27.42	47.58	28.23	48.3	53.23
Avg	46.09	47.98	48.43	51.01	47.10	50.5	51.49

Finally, for more comparison, we explained three different other methods then illustrated the final outcome in Table 6.

In paper [31], two and three categories of the DEAP database are used to classify valence and arousal. AR regression coefficients have been calculated as features. Feature selection is done using sequential forward feature selection (SFS) to decrease the complexity of computing and redundancy of features. Then, the three KNN, LDA and QDA clasifiers are used for categorization and the final results are compared. The best results are between %72.33 and %74.2 for the classification of the two valence and arousal categories and 61.1 and 65.16 for the classification of the three classes.

In essay [32] researchers integrated singular value decomposition (SVD) and Deep Belief Network (DBN) to gain better results. For achieving their goal, they used signals of DEAP database, then extracted information of channels F3 and F4. They applied empirical mode decomposition (EMD) method for decomposing EEG signals to into a set of intrinsic mode functions (IMFs). Then, effective components of IMFs were selected by using SVD. Extracted features that reduced by standard deviation (SD) method were considered as an input for DBN. Finally, classifier mapped emotion to three classes of valence and arousal. The results show the accuracy of 57.25% and 56.70% in a row.

In paper [33], emotions are classified according to the model of arousal-valence by using Fast Fourier transform analysis to extract features and Pearson correlation coefficient method to feature selection. Then researchers used a probabilistic classifier based on Bayes theorem with supervised learning using a perceptron convergence algorithm. To verify the proposed methodology, they used an open database, DEAP. They achieved the average accuracy of the valence and arousal, 70.9% and 70.1%, respectively.

Table 6. The accuracy of valence and arousal with different method of feature extraction and classifier (%)

Method	Valence	Arousal	Reference
SFS & KNN	72.33%	74.20%	[31]
SVD & DBN	57.25%	56.70%	[32]
FFT & Bayes	70.9%	70.1%	[33]
Proposed model	75.52%	81.03%	

According to the Table 6, in the case of valence, our system has the highest value 75.52%, which was closely followed by [31], [32]. In fact there is a few difference with other two methods. But, in the case of arousal this difference become more, that show accuracy has improved noticeably.

V. CONCLUSION

Deep belief network as a classifier is capable of discovering unknown features coherences of input signals that is crucial for the learning task to represent such a complicated model. The DBN provides hierarchical feature learning approach. When learning algorithms process more data, they provide better performance. The key advantage of self-taught learning and unsupervised feature learning is that the algorithm can learn from unlabeled data, and then it can learn from massive amount of information. Consequently, DBN algorithm is suitable for problems where there are a plenty of sets of unlabeled data and a handful amount of sets of labeled data. According to this, We developed a DBN based on restricted Boltzmann machine for classifying emotions. Discrete wavelet transform was used to extract linear features such as, power from EEG signals of 15 channels. Extracted features were considered as an input vector of DBN. An open access DEAP database employed for evaluating the efficiency of model. Finally, we obtained the accuracy rate of 75.52%, and 81.03%, for valence and arousal, respectively. It is shown that the proposed method has better performance in comparison with mentioned methods in section 4.

In future work, The performance of DBNs on the raw data from more than 15 channels in the dataset, up to all the 40 channels, should be investigated. Also, development methods for selecting channels is necessary to improve the performance of the algorithm.

Secondly, we will develop the model from two dimensions to four dimensions . We will investigate dominance and liking, too. Thirdly, For real application, the accuracy should be further improved. The DBN has some parameters that could be effectively improved to get better result. In the future, we will work on improving DBN structure and use other features and feature extraction methods.

VI. REFERENCES

[1]. Russell, James (1980). "A circumplex model of affect". Journal of Personality and Social Psychology. 39: 1161–1178.

[2]. V. Petrushin, Emotion in speech: recognition and application to call centers, in: Proceedings of the Artificial Networks in Engineering Conference, 1999, pp. 7–10.

[3]. M. Black, Y. Yacoob, Recognizing facial expressions in image sequences using local parameterized models of image motion, Int. J. Comput. Vis. 25 (1997) 23–48.

[4]. K. Anderson, P. McOwan, A real-time automated system for the recognition of human facial expressions, IEEE Trans. Syst. Man Cybern. B Cybern. 36 (2006) 96–105.

[5]. J. Wagner, J. Kim, From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification, in: Proceedings of IEEE International Conference on Multimedia and Expo, 2005, pp. 940–943.

[6]. K. Kim, S. Bang, S. Kim, Emotion recognition system using short-term monitoring of physiological signals, Med. Biol. Eng. Comput. 42 (2004) 419–427.

[7]. J. Brosschot, J. Thayer, Heart rate response is longer after negative emotions than after positive emotions, Int. J. Psychophysiol. 50 (2003) 181–187.

[8]. J.Coan,J.Allen,Frontal EEG asymmetry as a moderatoran dmediator of emotion, Biol.Psychol.67(2004)7–50.

[9]. P.Petrantonakis, L.Hadjileontiadis, A novel emotion elicitation indexusing frontal brain asymmetry for enhanced EEG-based emotion recognition, IEEE Trans. Inf.Technol. Biomed.15(2011)737–746.

[10]. X.Li,B.Hu,T.Zhu,J.Yan,F.Zheng,Towards affective learning with an EEG feedback approach, in: Proceedings of the 1st ACM International Workshopon Multimedia Technologies for Distance Learning, 2009,pp.33–38.

[11]. Klem, G. H., Lüders, H. O., Jasper, H. H., & Elger, C. (1999). The ten-twenty electrode system of the International Federation. Electroencephalogr Clin Neurophysiol, 52(3), 3-6.

[12]. B. Weiner, Attribution, emotion, and action, Handbook of Motivation and Cognition: Foundations of Social Behavior 1 (1986) 281–312.

[13]. T. Kemper, A Social Interactional Theory of Emotions, Wiley, New York, 1978z

[14]. R.Davidson,G.Schwartz,C.Saron,J.Bennett,D.Goleman,Frontalversus parietalEEGasymmetryduringpositiveandnegativeaffect,Psychophysiology 16(1979)202–203

[15]. P.Ekman,R.Davidson,TheNatureofEmotion:FundamentalQuestions,Oxford UniversityPress,1994

[16]. http://www.eecs.qmul.ac.uk/mmv/datasets/deap/

[17]. Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., ... & Patras, I. (2012). Deap: A database for emotion analysis; using physiological signals. IEEE Transactions on Affective Computing, 3(1), 18-31.

[18]. Liu, Y., & Sourina, O. (2013). Real-time fractal-based valence level recognition from EEG. In Transactions on Computational Science XVIII (pp. 101-120). Springer, Berlin, Heidelberg.

[19]. Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," Advances in neural information processing systems, vol. 19, pp. 153, 2007.

[20]. P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," Parallel distributed processing:

explorations in the microstructure of cognition, Cambridge, MA, USA: MIT Press, 1986, pp. 194-281.

[21]. G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, pp. 504-507, 2006.

[22]. G. E. Hinton, "Training products of experts by minimizing contrastive divergence," Neural computation, vol. 14, pp. 1771-1800, 2002.

[23]. divergence," N Keyvanrad, M. A., & Homayounpour, M. M. (2014). A brief survey on deep belief networks and introducing a new object oriented toolbox (DeeBNet). arXiv preprint arXiv:1408.3264.

[24]. Lopes, N., & Ribeiro, B. (2015). Deep Belief Networks (DBNs). In Machine Learning for Adaptive Many-Core Machines-A Practical eural computation, vol. 14, pp. 1771-1800, 2002.

[25]. Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., ... & Patras, I. (2012). Deap: A database for emotion analysis; using physiological signals. IEEE Transactions on Affective Computing, 3(1), 18-31.

[26]. Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry, 25(1), 49-59.

[27]. Kshirsagar, P., & Akojwar, S. (2016). Classification of Human Emotions using EEG Signals. International Journal of Computer Science, Volume146, (7).

[28]. Petrantonakis, P. C., & Hadjileontiadis, L. J. (2011). A novel emotion elicitation index using frontal brain asymmetry for enhanced EEG-based emotion recognition. IEEE Transactions on information technology in biomedicine, 15(5), 737-746.

[29]. Parameswariah, C., & Cox, M. (2002). Frequency characteristics of wavelets. IEEE Transactions on Power Delivery, 17(3), 800-804.

[30]. Hinton, G. (2010). A practical guide to training restricted Boltzmann machines. Momentum, 9(1), 926.

[31]. Hatamikia, S., Maghooli, K., & Nasrabadi, A. M. (2014). The emotion recognition system based on autoregressive model and sequential forward feature selection of electroencephalogram signals. Journal of medical signals and sensors, 4(3), 194.

[32]. Hosseini M, Pouyan A, Ferdosi S,Mashayekhi H,(2016),"Emotion Recognition of EEG data using Deep Belief Network and Empirical Mode Decomposition",submitted on journal of computational of Neuroscience

[33]. Yoon, H. J., & Chung, S. Y. (2013). EEG-based emotion estimation using Bayesian weighted-log-posterior function and perceptron convergence algorithm. Computers in biology and medicine, 43(12), 2230-2237.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Analysis of different Heart Rate Measuring Sensors with Arduino and Smartphone

Swapnil G. Deshpande[1], Dr. Vilas M. Thakare[2], Dr. Pradeep K. Butey[3]
[1]Computer Science, Arts, Commerce & Science College, Amravati, Maharashtra, India
[2]PG Deaprtment of Computer Science, S.G.B.A.U. Amravati, Maharashtra, India
[3]Computer Science, Kamla Nehru, Mahavidyalaya Nagpur, Maharashtra, India

ABSTRACT

This paper discuss about Arduino based heart rate measuring techniques using easy pulse sensor and pulse sensor. Heart rate monitoring and counting is performed using different software tools. The proposed framework collects input from pulse sensor by placing the patients' finger on the sensor. Then it is processed using Arduino to count the number of pulses and display the output in the smartphone via Bluetooth application. By using this framework physical presence of doctor is not needed at the time of measuring the heart rate. Such a system could be used in hospitals, for home-care, and people suffering from heart diseases.

Keywords : Heart Rate Sensor, Heart Rate Measurement (HRM), photophelthysmography (PPG), Arduino, Pulse Sensor, Bluetooth, Smartphone.

I. INTRODUCTION

Heart rate is important organ of human being. Heart functioning is indicated by using heart rate. It helps in finding the causes of symptoms, such as an irregular or rapid heartbeat, vertigo, weakness, and chest pain or breathe problem. High heart rate can cause cardiac problem. Therefore it is important to constantly monitor the heart beat rate [1]. Accurate heart rate detection is important in terms of our daily healthcare and exercise monitoring by classifying PPG signals obtained from wearable devices [2].

A simple and low-cost alternative method to estimate the heart rate is the use of the PPG signals [3]. Smartphones are used in combination with other sensing devices to capture heart rates. Pulse sensor is commonly used to measure heart rate.

The proposed design and development of a Heart Rate Measuring device measures the heart rate efficiently in a short time and with less expense without using time consuming and expensive clinical pulse detection systems [4].

The heart rate rises slowly during exercises and returns slowly to the rest value after exercise. When the pulse returns to normal is an indication of the fitness of the person. Heart rate below the normal condition is known as bradycardia, while above is known as tachycardia.

Heart rate can be measured by placing the thumb over the subject's arterial pulsation, and feeling, timing and counting the pulses usually in a 30 second period. Heart rate (bpm) is calculated by multiplying the obtained number by 2. ECG is most frequent technique to measure heart rate. But it is very expensive. PPG sensor and pulse sensor is cheaper and useful instrument in knowing the pulse of the patient.

Heart rate may differ from person to person. As age changed, the regularity of pulse will be changed.

Smartphone market is growing rapidly in mobile phone sector, with Apple's iPhone, Android and Microsoft's Windows Phone. A substantial rise in smartphone applications would be those that are employed in the fields of health care and medicine.

The heart rate of different sensors on android smartphone is studied. Through this study, methodology helps to determine the best sensor for monitoring heart rate. To perform these operations the system uses two heart beat sensor, i.e PPG Sensor and pulse Sensor.

II. PREVIOUS WORK

Bandana Mallick et.al. monitored heart rate which is capable to monitor the heart beat rate of patient [5].

Y.S.Harish Kumar et. al. determine human heart rate, especially for heart patients who need to monitor their heart rate, it being an important indicator for prognosis and diagnosis, and also share it with their physician anytime to seek medical advice when needed [6].

Salomi S. Thomas et. al. measure body temperature and heart rate using arduino. This device will allow one to measure their mean arterial pressure (MAP) in one minute and the accurate body temperature will be displayed on the Android. The system can be used to measure physiological parameters, such as Heart rate and Pulse rate [7].

Sonal Chakole et. al. focuses on health monitoring System using sensors and can help people by providing healthcare service [8].

Prasad Kumari Nisha et. al. implemented heart rate monitoring system using low cost arduino board and other easily available resources[9].

III. EXISTING TECHNOLOGY

There are many technologies that have been developed to estimate heart rate, which could affect someone’s heart rate such as motion, emotions, and stress. These technologies are very expensive.

IV. PROPOSED FRAMEWORK

This framework uses Easy pulse sensor and pulse sensor for extracting statistical parameters from the processed signals. The process signal analyzed to determine how changes would affect the heart rate of an individual.

The goal of this framework is to design home-care systems that is low in cost, consumes low power and provide reliable heart rate readings.

V. PROPOSED SYSTEM

Physical activities as well as physiological signals of the heart monitoring patient could be easily monitored with the help of wearable sensors. The whole activity can be monitored remotely by doctors, nurses, or caretakers.

The framework consists of Arduino UNO microcontroller system, transmission system and Android based application. The framework calculates heart rate (BPM) on the portable device in real time and shows it on Android based smartphone. The cost of proposed framework is affordable as compared to other developed devices due to use of Arduino, smartphone and Android device [10].

Heart rates are recorded through the sensor nodes and transmit to the smart phone via Bluetooth. For measuring the heart rate (beats per minute) of a person, technique of PPG Sensor and pulse Sensors is used. The sensor should be placed in those areas of body where the blood is having a higher concentration. Android device is connected with Arduino microcontroller via wireless serial BLE connection.

This framework is designed for monitoring and measuring heart beats by using PPG Sensor and pulse Sensors of 10 human subjects of different age groups with smart phone and comparing result, and then identifies the best sensor.

VI. HARDWARE REQUIREMENT

Arduino Microcontroller

Arduino is an open source platform. It can use for building digital devices and interactive objects. Arduino boards are available commercially in preassembled form, or as do-it-yourself kits. Arduino board are equipped with sets of digital and analog input/output (I/O) .

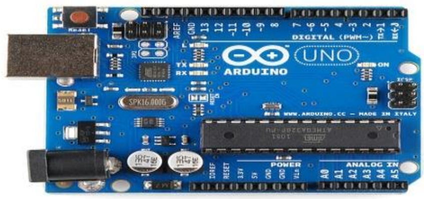


Figure 1. Arduino Uno board

Bluetooth Module HC-05:

Bluetooth Module HC-05 is intended for transparent wireless serial connection setup. It is used in a Master or Slave configuration for making it a great solution for wireless communication. It is easy to use. The HC-05 Bluetooth Module has 6 pins. ENABLE, Vcc, GND,TXD , RXD, STATE.

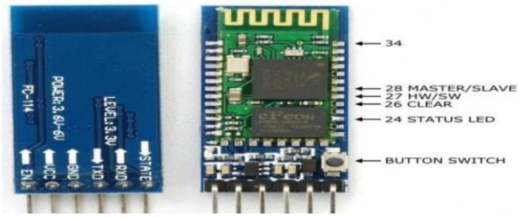


Figure 2. HC- 05 Bluetooth Module

Button Switch:

Button switch is used to switch the module into AT command mode. To enable AT command mode, press the button switch for a second. If user want to change the parameter of this module when the module is not paired with any other BT device then AT command is used. If the module is connected to any other bluetooth device, it starts to communicate with that device and fails to work in AT command mode.

Jumper Cables:

A group of electrical wire with a connector at each end is called jump wire or jumper. Jumper cables are used to interconnect the components of a breadboard with other equipment or components, without soldering.

Figure 3. Jumper Cables

Easy Pulse Sensor based on PPG Sensor

Easy pulse sensor is based in the principle of PPG sensor. PPG sensor is designed to measure the heart beat when a finger is placed on it. It is directly connected to microcontroller to measure the Beats Per Minute (BPM) rate. PPG works on the principle of light modulation.

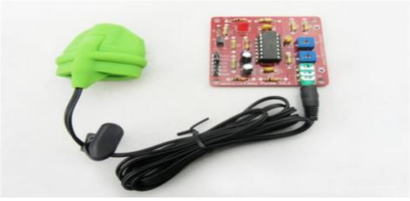


Figure 4. Easy Pulse Sensor

Photoplethysmography (PPG) is a non-invasive method for measuring the variation in blood volume in tissues using a light source and a detector. This technique is used to calculate the heart rate.

The Figure 4 shows how PPG sensor extract the pulse signal from the fingertip. A subject's finger is illuminated by an infrared light-emitting diode. Depending on the tissue blood volume more or less light is absorbed. The intensity of reflected light varies with the pulsing of the blood with heart beat.

The sensor consists of a red LED and light detector. The LED needs to be super bright as the maximum light must pass spread in finger and detected by detector. With each heart pulse the detector signal varies. This variation is converted to electrical pulse [11].

Pulse sensor:

Pulses can be recorded by holding a finger to your neck or wrist and counting the beats with watch. Pulse sensor fits over a fingertip and uses the amount of infrared light reflected by the blood circulating inside the body. Figure 5 shows front side and backside of pulse sensor.

The pulse sensor is a well designed plug-and-play heart-rate sensor for Arduino.

The small, round shape of the Pulse Sensor makes it convenient for obtaining the heart rate signal from subjects' fingers.

The sensor consists of an infrared emitter and detector build up side by-side and pressed closely against the skin. When the heart pumps, blood pressure gets rising, so amount of infrared light from the emitter gets reflected back to the detector. The detector passes more current when it receives more light.

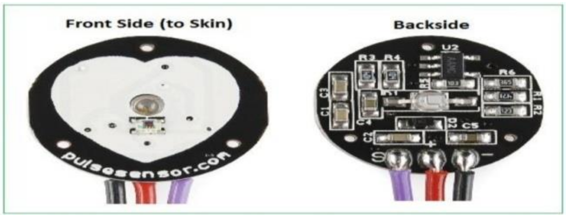


Figure 5. Pulse Sensor

VII. SOFTWARE REQUIREMENTS

Android Programming:

The Android operating system has come in market in late of 2007. It is an Open Handset Alliance. The idea of an open source OS for embedded systems was not new, but Google helped to push Android to the forefront in just a few years.

Many wireless communication protocols have one or more Android phones available. Other embedded system, such as tablets, notebooks, televisions, set-top boxes, and even automobiles, also have accept the Android OS. Android applications are written in Java that is sometimes known as the Dalvik virtual machine.

ARDUINO Programming- A Proposed Algorithm:

Arduino programming is user friendly, more compact, and less complex, which is used to perform several tedious and repetitive tasks [12].

Arduino has low power consumption, low cost, small size, etc. so that real time monitoring is possible & patient can be treated on time with the system & is helpful in worst condition.

Arduino Uno is a microcontroller board based on the ATmega328P (datasheet). It has 14 digital input/output pins, 6 analog inputs, a 16 MHz quartz crystal, a USB connection, a power jack, an ICSP header and a reset button. It contains everything needed to support the microcontroller; simply connect it to a computer with a USB cable or power it with a AC-to-DC adapter or battery to get started [8].

Ardunio is used because it can sense the environment by receiving input from a variety of sensors and can affect its surroundings by controlling lights and motors. Arduino coding is needed for sensing heart rate by using arduino software.

Data Flow Diagram:

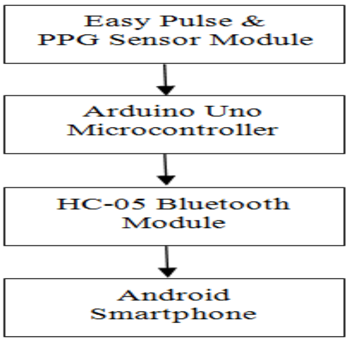


Figure 6. Basic Workflow of framework

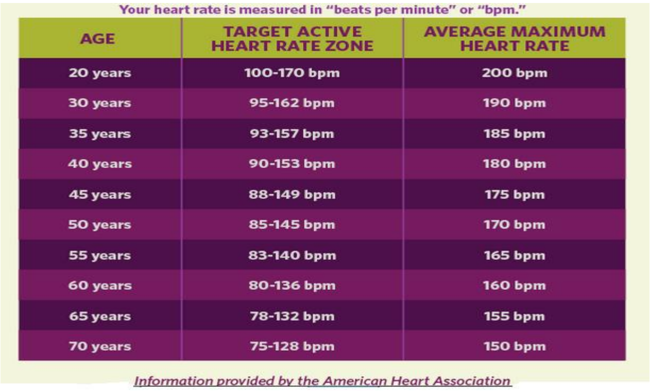
This framework first connect PPG and pulse sensor with the Arduino Uno Microcontroller. Then send the data on Android mobile via Bluetooth. Connect Bluetooth module to Arduino Uno Microcontroller. After making the connection with Arduino, upload the Arduino sketch into the Arduino. After uploading the sketch the sensors activated. After activating the sensor, attach the Pulse sensor and PPG sensor to user finger.

Connect the android smartphone to catch the heart rate. Start the app and connect it to HC-05 Bluetooth module. After successful connection, the human heart rate will display on to the user smartphone in the unit of Beats per Minutes (BPM).

VIII. RESULT & PERFORMANCE

Different age group of patient can be tested in this paper. Following Heart Rate variations can be measured using BPM (Beats Per Minute):

Table 1. Normal Heart Rate Chart:



Your heart rate is measured in "beats per minute" or "bpm."

AGE	TARGET ACTIVE HEART RATE ZONE	AVERAGE MAXIMUM HEART RATE
20 years	100-170 bpm	200 bpm
30 years	95-162 bpm	190 bpm
35 years	93-157 bpm	185 bpm
40 years	90-153 bpm	180 bpm
45 years	88-149 bpm	175 bpm
50 years	85-145 bpm	170 bpm
55 years	83-140 bpm	165 bpm
60 years	80-136 bpm	160 bpm
65 years	78-132 bpm	155 bpm
70 years	75-128 bpm	150 bpm

Information provided by the American Heart Association

Table 2. Output Result of age group 20 to 30:

Name	Age	Direct Measure	Easy Pulse Sensor(BPM)	Pulse Sensor(BPM)
Pooja	26	98	94	85
Sakshi	24	129	121	111
Ankita	20	100	98	92
Amol	30	119	116	106
Anand	21	102	90	95

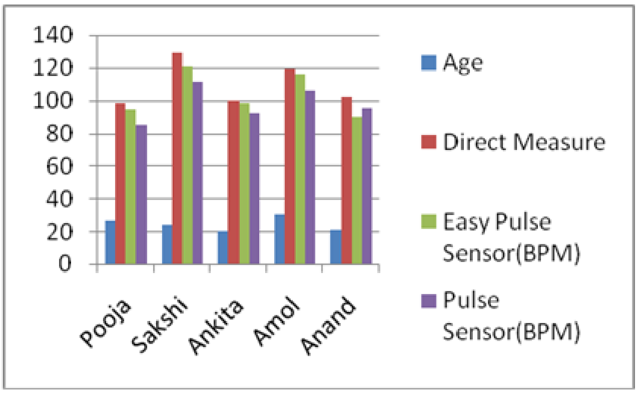


Figure 7. Output Result of age group 20 to 30: Easy pulse sensor performs better than pulse sensor

Table 3. Output Result of age group 30 to 40:

Name	Age	Direct Measure	Easy Pulse Sensor(BPM)	Pulse Sensor(BPM)
Sonali	37	108	104	100
Mohini	31	109	110	111
Sanika	35	120	113	106
Sagar	39	112	116	96
Prasad	34	98	96	91

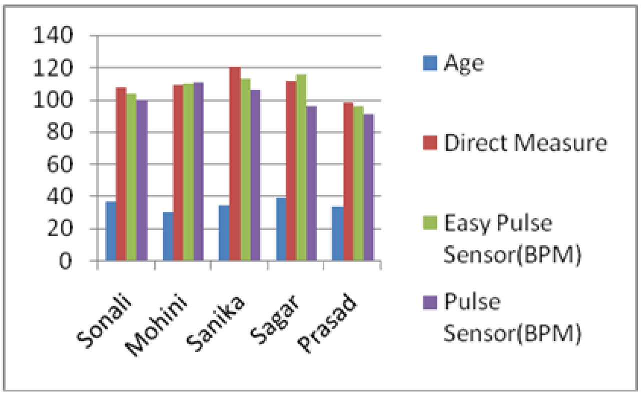


Figure 8. Output Result of age group 30 to 40: Easy pulse sensor performs better than pulse sensor

Table 4. Output Result of age group greater than 50:

Name	Age	Direct Measure	Easy Pulse Sensor(BPM)	Pulse Sensor(BPM)
Amit	55	94	88	84
Govind	75	111	105	101
Srikant	62	120	103	109
Rasika	70	102	97	100
Sita	66	96	99	89

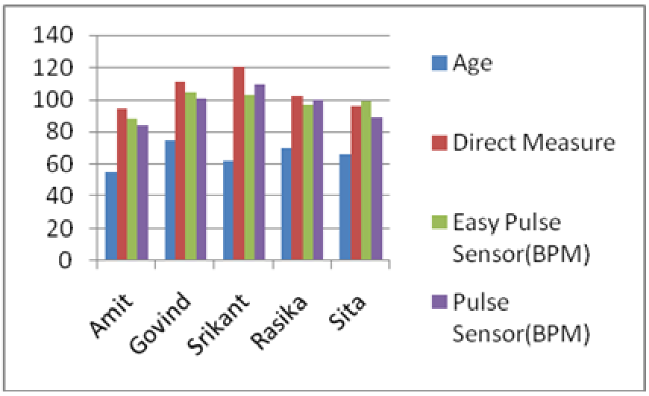


Figure 9. Output Result of age group greater than 50: Easy pulse sensor performs better than pulse sensor

IX. CONCLUSION

This paper determines the heart beat rate per minute of patient. If critical situation is occur then sends alert message to the mobile phone. As the designed system is portable, cost effective and easy to use, this will reach easily to rural people.

With the help of developed application, following points are observe -

- ✓ Easy Pulse sensor is easy to use and handle while pulse sensor not.
- ✓ Easy Pulse sensor gives more accurate result while the pulse sensor not gives accurate result.
- ✓ Sometimes, pulse sensor not gives result.
- ✓ Easy Pulse sensor is costly than the finger tip sensor.

After studying all the details about the sensors and its result, we conclude that the Easy Pulse sensor is best for measuring heart rates.

X. FUTURE WORK

- ✓ In future, the design can be extended by using WIFI or GSM or GPS for long distance communication.
- ✓ A portable heart rate monitoring system can be designed using Arduino.

- ✓ Continuous wearing of sensors was uncomfortable and irritating to users so used inbuilt sensor of smartphone.

XI. REFERENCES

[1]. Md. Tarikul Islam Papon, Ishtiyaque Ahmad, Nazmus Saquib, and Ashikur Rahman, "Non-invasive Heart Rate Measuring Smartphone Applications using On-board Cameras: A Short Survey", IEEE International Conference on Networking Systems and Security (NSysS), DOI: 10.1109/NSysS.2015.7043533, 2015.

[2]. Rifat Zaman, Chae Ho Cho, Yeesock Kim, and Jo Woon Chong, " A Novel Heart rate Monitoring Using a Smartphone", IEEE Healthcare Innovation Point-Of-Care Technologies Conference (HI-POCT), PP.5 - 8, DOI: 10.1109/HIC.2016.7797683, 2016.

[3]. Dragos Daniel Taralunga, Irina Nicolae, Bogdan Hurezeanu, Mihaela Ungureanu, Rodica Strungaru, "Non-contact Heart Rate Estimation from a Video Sequence", International Conference and Exposition on Electrical and Power Engineering (EPE 2016), 20-22 October 2016.

[4]. M.M.A. Hashem, Rushdi Shams, Md. Abdul Kader, and Md. Abu Sayed, "Design and Development of a Heart Rate Measuring Device using Fingertip", International Conference on Computer and Communication Engineering (ICCCE 2010), 11-13 May 2010.

[5]. Bandana Mallick, Ajit Kumar Patro, "Heart Rate Monitoring System Using Finger Tip Through Arduino And Processing Software", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 5, Issue 1, January 2016 .

[6]. Y.S.Harish Kumar, K. Naghabhushanam, "The Smartphone Accessory Heart Rate Monitor", International Journal of Advance technology and Innovative Research, Vol.06,Issue.08, Pages:810-815, October-2014.

[7]. Salomi S. Thomas, Mr. Amar Saraswat, Anurag Shashwat, "Sensing Heart beat and Body Temperature Digitally using Arduino", IEEE International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), DOI: 10.1109/SCOPES.2016.7955737, 2016.

[8]. Mrs. Sonal Chakole, Ruchita R.Jibhkate, Anju V.Choudhari, Shrutika R.Gawali, Pragati R.Tule, "A Healthcare Monitoring System Using Wifi Module", International Research Journal of Engineering and Technology (IRJET), Volume: 04,Issue: 03, Mar -2017.

[9]. Prasad Kumari Nisha , Yadav Vinita, "Heart Rate Monitoring and Data Transmission via Bluetooth", International Journal of Innovative and Emerging Research in Engineering Volume 2, Issue 2, 2015.

[10]. Md. Asaduzzaman Miah, Mir Hussain Kabir, Md. Siddiqur Rahman Tanveer and M. A. H. Akhand, "Continuous Heart Rate and Body Temperature Monitoring System using Arduino UNO and Android Device", International Conference on Electrical Information and Communication Technology (EICT), 2015.

[11]. Yunyoung Nam, "Monitoring of Heart and Breathing Rates Using Dual Cameras on a Smartphone ", PLOS ONE, March 2016.

[12]. Rajalakhshmi.S S.Nikilla," Real Time Health Monitoring System using Arduino", South Asian Journal of Engineering and Technology Vol.2, Issue18 , PP. 52–60, 2016.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Role of Agent Oriented Programming over Object Oriented Programming - A Part of Artificial Intelligence

Pranav Ratta
Assistant Professor, Department of Computer Science, Institute of Management Sciences Jammu, Jammu and Kashmir India

ABSTRACT

The first question is what is meant by an 'Agent'. The answer is, Software with Mental State. Agent Oriented Programming (AOP) sits one level of abstraction above Object Oriented Programming (OOP). In this paper, firstly we discuss what AOP is. Then we discuss the brief history of AOP after that we analyze how the change occurs due to Agent Oriented Programming over the Object Oriented Programming.

Keywords : Agent Oriented Programming, Object Oriented Programming

I. INTRODUCTION

Artificial Intelligence is one of the newest fields of intellectual research; its foundation began thousands of years ago where human fantasy of having intelligent and thinking machines appears in myths or stories. In recent years Agent oriented programming is one of the most important area of development, and still an area of considerable research. Wikipedia traces OOP back to the 1960s, while AOP came about from research into artificial intelligence by one Yoav Shoham in the 1990s [1].

Evolution of Programming Languages: Monolithic programming, Modular programming, Object-oriented programming, Agent programming . This paper discusses both the similarities and differences between objects and agents than u will decide how agent oriented programming have impact over object oriented programming.

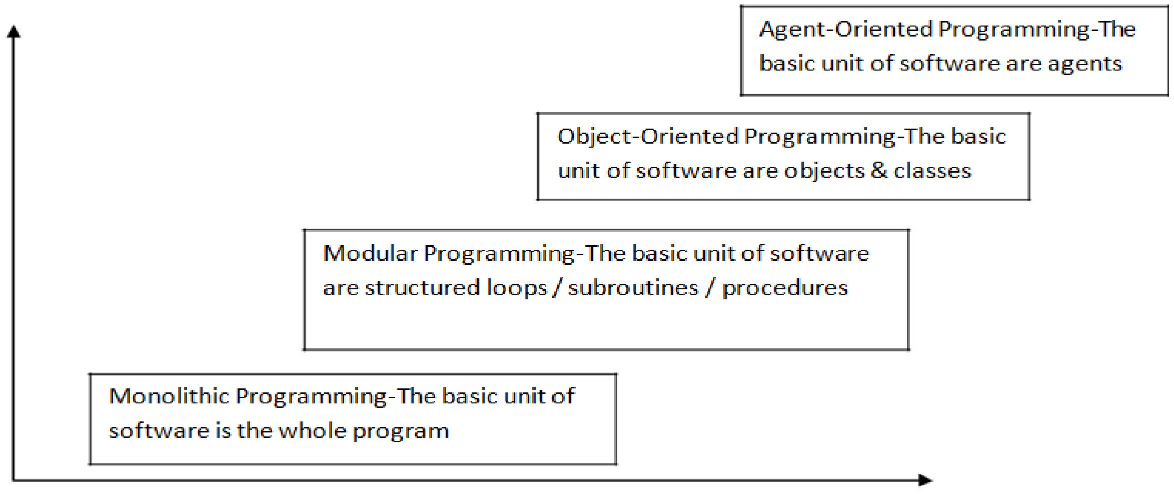

Agent-Oriented Programming-The basic unit of software are agents
Object-Oriented Programming-The basic unit of software are objects & classes
Modular Programming-The basic unit of software are structured loops / subroutines / procedures
Monolithic Programming-The basic unit of software is the whole program

Figure 1

II. WHAT IS AGENT?

Agents means entity that functions autonomously in an environment. Agents provide a very effective way of building applications for dynamic and complex environments. develop software components as if they have beliefs and goals, act to achieve these goals, and are able to interact with their environment and other agents.[2]

Intelligent agent: Set of independent software Components linked with other applications. The main objective of an intelligent agent is to store the user preferences dynamically related to an application and implement the same when user accesses the same application. Intelligent Agent is an autonomous entity which learn from its environment and use their knowledge to acts upon an environment and directs its activity towards achieving goals. [3]

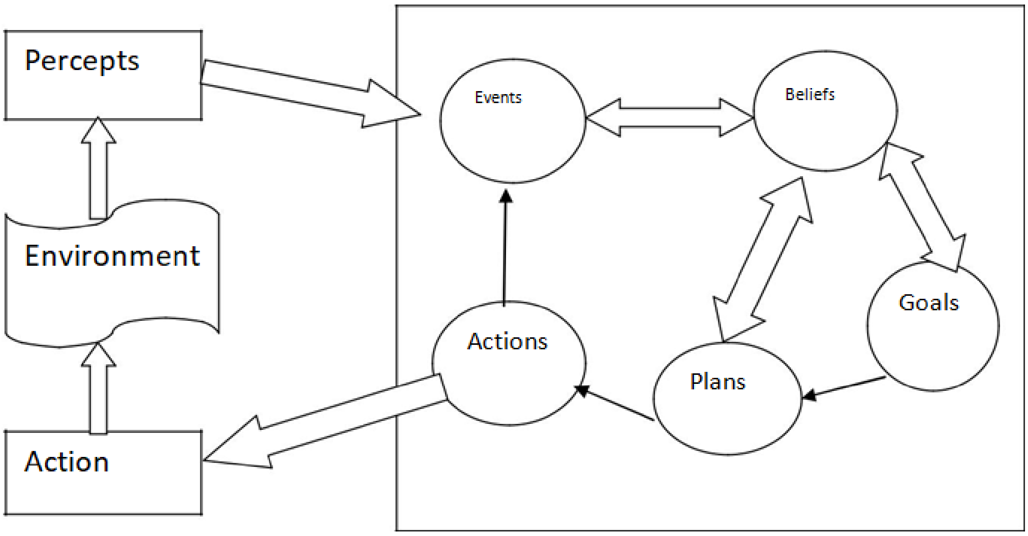


Figure 2

III. OVERVIEW OF OBJECT AND AGENT ORIENTED PROGRAMMING

Object-oriented programming is a programming model organized around objects rather than actions. Conventional procedural programming normally takes input data, processes it, and produces output data. The primary challenge of programming is how to write the logic. Object-oriented programming focuses on the objects that you want to manipulate, their relationships, and the logic required to manipulate them. The concepts of a type, class, interface, and object are closely related but it is important to understand the difference between these four terms. A type is a name that identifies specific members of a class, which can include methods, properties, data members, and events. A class defines the implementation of a type and its class members. An abstract class is essentially a type with an incomplete implementation. An interface also defines a type that identifies certain class members (properties, methods, or events) that a class must implement. An object is an instance of a class whose type can be represented as any class or interface that contributes members defined in the object's class hierarchy.[4]

AOP is a more recent development, and still an area of considerable research and standardization. The objective of Agent Oriented (AO) Technology is to build systems applicable to real world that can observe and act on changes in the environment. Such systems must be able to behave rationally and autonomously in completion of their designated tasks. Shoham suggests that AOP system needs each of three elements to be complete

A formal language with clear syntax for describing the mental state. This would likely include structure for stating beliefs, passing messages etc.

A programming language in which to define agents. The semantics of this language should be closely related to formal language.

A method for converting neutral applications into agents. This kind of tool will allow an agent to communicate with a non-agent .

Brief history of Agent oriented Programming

Shoham's first attempt at an AOP language was the AGENT-0 system in 1990. The key component of Agent-0 is speech act. A more refined implementation was developed by Thomas is PLACA in1993 (AGENT-0 extension with plans).The inability of Agent-0 is planning. PLACA used for Planning. AgentSpeak was developed in 1996 by Anand Rao.Golog was agent oriented language introduced in 1996 for Action theories, logical specification.

Some more Agent oriented programming languages

- ✓ 1997: 3APL (Hindriks et al.) used for Practical reasoning rules
- ✓ 2000: JACK (Busetta, Howden, Ronnquist, Hodgson) used for Capabilities and it is Java-based
- ✓ 2000: GOAL (Hindriks et al.)
- ✓ 2000: CLAIM (Amal El FallahSeghrouchni)
- ✓ 2002: Jason (Bordini, Hubner; implementation of AgentSpeak)
- ✓ 2003: Jadex (Braubach, Pokahr, Lamersdorf)
- ✓ 2008: 2APL (successor of 3APL) [2]

Aop versues Oop

Extension of OOP where objects become agents by redefining both their internal state and their communication protocol in intentional terms.

Agents have quality of volition that is using AI techniques intelligent agents judge their results and modify their behavior and their own internal structure to improve their perceived fitness Normal objects contain arbitrary values in their slots and communicate with messages.

AOP agents contain beliefs, commitments, choices, and the like and communicate with each other via a constrained set of speech type acts such as inform, request, promise, decline the state of the agent is called its mental state OO focused on defining interfaces for objects coupling where one objects needs to invoke a specific method with specific arguments on the other object thereby coupling the two in code.

This same method invocation does occur in agents with one major difference, there effectively just one method with each agent and one argument. All the semantics of the invocation are bundled into that one argument just like in human communication where one language is used to initiate complex cooperative behavior.

Agents may communicate using an ACL or ICL where objects communicate with a fixed method of interfaces Objects are abstractions of things like invoices.

Agents are abstractions of intelligent beings they are essentially anthropomorphic not intelligent in the human sense only modeling an anthropomorphic architecture with beliefs, desires, etc

IV. CONCLUSION

In this paper, I discussed agent-oriented programming over Object-Oriented Programming , how Agent oriented programming is better ,where the tasks are in charge of autonomous computational entities, which interact and cooperate within a shared environment. Agents have th ability to learn , it can add subtract features dynamically. .I conclude this paper with remark that In order to strss and investigate, a full value of agent oriented approach, we need programming languages which work for agent development .

V. REFERENCES

[1]. McCorduck, Artificial Intelligence in myth, 2004, pp. 4-5.

[2]. Leuven October 2011, Agent - Oriented Programming

[3]. Kaiserslautern, 2004 ,Nick M. M., Building and Running Long-Lived Experience-Based Systems, PhD thesis, Dept. of Computer Science, University of Kaiserslautern,

[4]. Ashish Bishnoi , Artificial Intelligence: Analysis of various Agent Programming Languages

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Machine Learning : Concept and Applications in Smart City Projects in India

Moonis Ali, Faisal Rasheed Lone, Ali Hussain
Department of Computer Science & Engineering, University of Kashmir, North Campus Delina, Baramulla, J&K, India

ABSTRACT

Machine learning algorithms automatically learn from the inputs that are fed to them. This not only saves time but is also very cost effective as it saves enormous resources required for programming them. As more and more progress is made in this field, the applications are becoming ever increasing. The present paper explains the various types & techniques of Machine learning and sheds light on the applications of the field with the focus on its use in the concept of a smart city. While a smart city is an urban development vision which integrates various information & communication technologies with the Internet of Things(IoT) in a secure way, the notion of smart city described here has taken care of the conceptualization & variation of the idea of 'smart city' in the Indian context. Some suggestions have also been made for laying an effective groundwork for seamless integration of Machine learning with the smart city concept. At the end some challenges have been identified while discussing the future scope and implementation of these fields.

Keywords : Machine learning, smart city, Internet of Things.

I. INTRODUCTION

In a field as vast and robust as Computer Science the pace of evolution and innovation is only gaining momentum with each passing milestone. One of the key areas that has promoted to the enormous growth in the past decade or two is the field of Machine Learning (ML). The scope and field of Machine Learning is tremendous and the range of applications that it offers in various fields, from identifying/predicting cases of cancer causing genes in health care to daily traffic forecast analysis in transportation to sustainability, environment to leveraging data driven decision making in policy planning and governance. The extent of this field is broad and expansive. While previously ML was mostly used to identify user search patterns for optimized search engine results or for creation of chess playing bot, ML nowadays has evolved in itself as an independent & capable leader in the field of computer science in terms of the huge array of solutions that it offers for the betterment of the world in general and human race in particular.

The formal definition for Machine Learning that is introduced in this paper has endeavored to formulate a simpler, yet concise & informative explanation of the field.

" Machine learning is a subfield of computer science that empowers computers to act, [1] learn and make decisions like humans, by feeding them data and information in the form of observations and or real-world interactions without the need of explicit programming them. Machine learning systems automatically learn from the inputs that are fed to them. This is a better alternative to manually constructing them because it saves us a lot of time and resources. In the last decade, the use of machine learning has grown rapidly throughout computer science and beyond. Use of Machine learning has encompassed various facets ranging from spam and firewall blockers, web search, credit scoring finance, stock trading, drug designing, forensics, etc.

II. TYPES OF MACHINE LEARNING ALGORITHMS

Machine Learning algorithms can be classified into categories depending on the algorithm and the objectives it wishes to achieve. The following categorization of Machine Learning algorithms based on their ability to learn tasks is given as under:

- ✓ Supervised Learning
- ✓ Unsupervised learning
- ✓ Semi supervised Learning
- ✓ Reinforced Learning

A) Supervised learning

This Machine Learning technique is implemented when an algorithm learns from example data and associated target responses that can consist of numeric values or string labels, such as classes or tags, in order to later predict the correct response when posed with new examples. The supervised approach is indeed similar to human learning under the supervision of a teacher. The teacher provides good examples for the student to memorize, and the student then derives general rules from these specific examples [2] . Supervised learning is fairly common in classification problems because the goal is often to get the computer to learn a classification system that we have created. Digit recognition, once again, is a common example of classification learning. More generally, classification learning is appropriate for any problem where deducing a classification is useful and the classification is easy to determine [3] .Some of the characteristic of supervised learning are as follows:

- ✓ Supervised learning algorithms try to model relationships and dependencies between the target prediction output and the input features.
- ✓ Using supervised learning algorithms we can predict the output values for new data based on those relationships which it learned from the previous data sets.[4].

Some of the common algorithms in use are:

- ✓ Nearest Neighbour
- ✓ Naive Bayes
- ✓ Decision Trees
- ✓ Linear Regression
- ✓ Support Vector Machines (SVM)

B) Unsupervised Learning

An Unsupervised Machine learning technique is implemented when an algorithm learns from plain examples without any associated response, leaving it to the algorithm to determine the data patterns on its own. This type of algorithm tends to restructure the data into something else, such as new features that may represent a class or a new series of uncorrelated values. They are quite useful in providing humans with insights into the meaning of data and new useful inputs to supervised machine learning algorithms [2]. Unsupervised algorithms are the family of machine learning algorithms which are mainly used in pattern detection and descriptive modeling. However, there are no output categories or labels here based on which the algorithm can try to model relationships.

- ✓ These algorithms try to use techniques on the input data to mine for rules, detect patterns, and summarize and group the data points which help in deriving meaningful insights and describe the data better to the users [4].

Common unsupervised algorithms in use are:

- ✓ K-means clustering algorithms
- ✓ Association rule learning algorithms.

C) Semi Supervised Learning

The Semi-supervised ML technique is a class of supervised learning tasks and techniques that make use of unlabeled data for training – typically a small portion of labeled data with a large amount of unlabeled data (together). The Semi-supervised sub-class of learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy. The acquisition of labeled data for a learning problem often requires a skilled human agent (e.g. to transcribe an audio segment) or a physical experiment (e.g. determining the 3D structure of a protein or determining whether there is oil at a particular location) [5]. The cost that is associated with the labeling process may render a completely labeled training set infeasible, while as the acquisition of unlabeled data is relatively inexpensive. In situations as these, semi-supervised learning can be of great applied value.The characteristic property of unsupervised learning is:

- ✓ These methods exploit the gaps within the group memberships of an unlabeled dataset, which are unknown; this data carries important information about the group parameters [4].

D) Reinforced Learning

Reinforced Learning algorithm aims at using observations gathered from the interaction with the environment to take actions that would maximize the reward or minimize the risk. Reinforcement learning algorithm (called the agent) continuously learns from the environment in an iterative fashion. In the process, the agent learns from its experiences of the environment until it explores the full range of possible states. Reinforcement learning allows machines and software agents to automatically determine the ideal behavior within a specific context, in order to maximize its performance. Simple reward feedback is required for the agent to learn its behavior. this is known as the reinforcement signal. In order to produce intelligent programs (also called agents), reinforced learning goes through the following steps:

- ✓ The input state is observed by the agent.
- ✓ Decision making function is used to make the agent perform an action.
- ✓ After the action is performed, the agent receives reward or reinforcement from the environment.
- ✓ The state-action pair information about the reward is stored.[4]

Common reinforcement algorithms in use are:

- ✓ Q-Learning
- ✓ Temporal Difference (TD)
- ✓ Deep Adversarial Networks

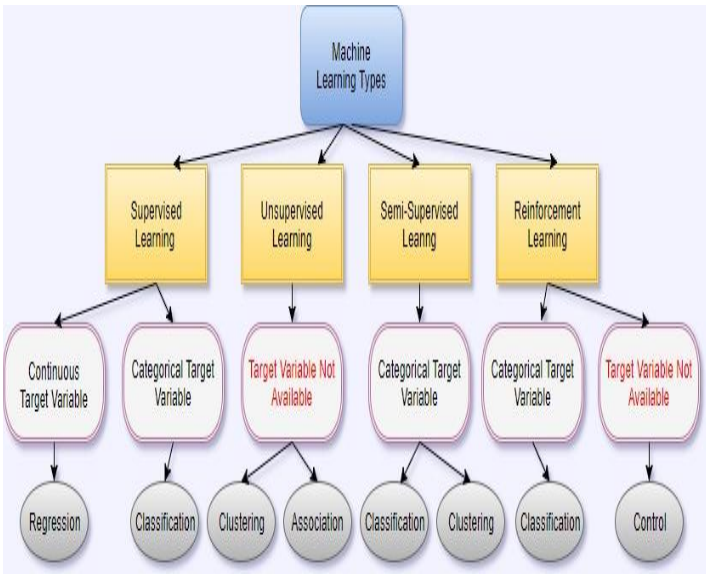


Figure 1. Categorization of various Machine learning techniques based on their ability to learn tasks

III. SMART CITIES IN INDIAN SCENARIO

In a general context, a smart city is considered as an urban development vision to integrate information & communication technology (ICT) with the Internet of things (IoT) in a secure fashion in order to manage a city's assets and make optimal use of it. These assets may include the entire local: information systems, educational institutes, libraries, transportation systems, hospitals, power plants, water supply networks, waste management, law enforcement, and other community services. A smart city is promoted to use urban informatics and technology to improve the efficiency of services. ICT allows city officials to interact directly with the community and the city infrastructure and to monitor what is happening in the city, how the city is evolving, and how to enable a better quality of life. Through the use of sensors in the IoT (infrastructure) integrated with real-time monitoring systems, data is collected from citizens and devices – then processed and analyzed. The information and knowledge gathered are keys to tackling inefficiency [6]

For a country like India, the connotation of a smart city would be different than, say a country in Scandinavian Europe. In the approach of the Smart Cities Mission (India), the objective is to promote cities that provide core infrastructure and give a decent quality of life to its citizens, a clean and sustainable environment and application of 'Smart' Solutions. The focus is on sustainable and inclusive development and the idea is to look at compact areas, create a replicable model which will act like a light house to other aspiring cities.[7]

IV. IMPLEMENTATION OF MACHINE LEARNING VIZ A VIZ SMART CITY

The Machine Learning scenario in the 21st century is potent enough to augment the infrastructure in the information and communication technology (ICT) and Internet of things (IoT) systems and, if integrated within the existing command-control structure of a smart city, ML can yield tremendous results. The most extensive use cases for ML in the smart city concept are found in management and redressal of traffic congestion, management of utilities, and implementation in demand based smart water system and wastewater disposal mechanisms.

The case for implementing various Machine Learning techniques can be gauged by looking at the comprehensive research already performed in the form of predictive and analytical models for traffic, utilities and waste management. The following cases shed extensive light on why ML may be incorporated in the smart city architecture of developing nations like India.

A. Management of Traffic

The case study by Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia (2015) in Research Analytics, provides ample data about how ML can be implemented actively for dealing with the problems of traffic jams, identifying spots of traffic congestion or forming a responsive mechanism for dealing with traffic related incidents by live classification of social media reportage.

The case study focuses on Predicting Near Future Traffic Jams & Identifying Hot Spots of Congestion and Incident Acknowledgement which have their applications relevant to any smart city.

Predicting Near Future Traffic Jams and Hot Spots of Congestion:

If an accident or congestion occurs on a major road, it is likely that the traffic in the vicinity of the road will be affected. An early prediction of the nearby congested roads to avoid is very significant in deciding the optimal route for the drivers so that they may avoid the area and emergency vehicles make also make smart route choices based on the prediction to reach the site. Figure 2. illustrates the frequent pattern for congestion propagation around Olympic Park in Melbourne.

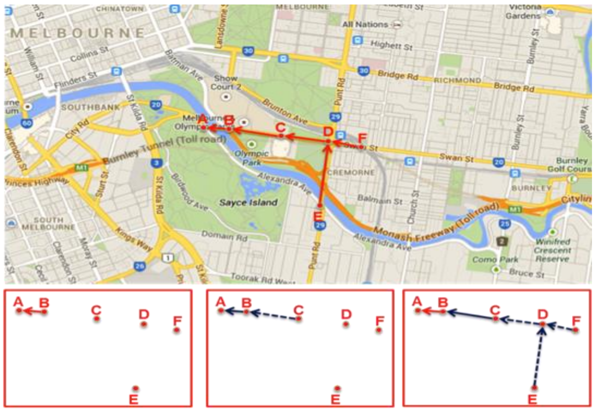


Figure 2. Congestion propagation pattern around Olympic Park, Melbourne [8]

The introduction of an algorithm based on the combination of association rule mining and dynamic Bayesian network to construct causality trees from congestions and estimate their propagation probabilities based on temporal and spatial information. Frequent sub-structures of these causality trees reveal not only recurring interactions among spatio-temporal congestions, but potential bottlenecks or flaws in the designs of existing traffic networks. [8].

Responsive Incident Acknowledgement System for Engineers:

For incident management purpose, the live system developed in the study displays the current location and status of all incidents from multiple sources including social media. Figure 3 presents the web service and interface for the real-time view of current incidents in Victoria. Heat map and cluster map are also generated.

This system utilizes advanced web technologies and efficient machine learning (ML) algorithms such as Support Vector Machines and Conditional Random Fields to classify the tweets relevant to traffic incidents.[8]

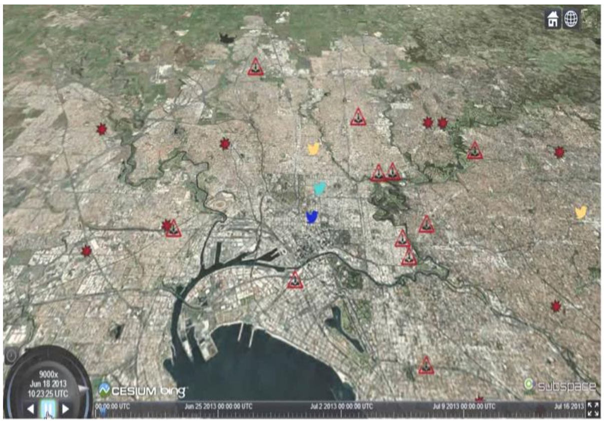


Figure 3. Incident detection and visualization [8]

Smart Waste Classification and Management:

As shown by the study performed by Kedia P [9], a supervised learning KNN (K-Nearest Neighbours) algorithm is used to identify and recognize patterns about the days on which particular waste- bins get filled. The KNN algorithm is then applied to determine the days when the waste-bins need to be picked up for unloading. Subsequently, using the results from identification of waste-bins, weekly plans can be drawn for waste-bins requiring servicing on daily basis as contrary to those who don't need it every day. The amount of time and cost (both in terms of labor and fuel) which is potentially saved is huge. Once the need based routines for servicing of different waste-bins (from different areas) on daily basis are established, ML can again be used for the classification of waste and waste can be categorized as household, industrial and medical waste. Depending on this second level of classification of waste using K-means or Density Estimation algorithms, a smart policy can be developed for the reuse and recycle of the waste. Thus another servicing routine can be established wherein specialized waste-trucks can service specific types of waste depending on the categorization of the areas and the occurrences of the waste thereof. Application of Machine learning algorithms across various fields of a smart city is summarized in Table 1.

Table 1. Application of machine learning algorithms across various fields of a smart city

Area/Field	Type of ML technique	Model / Method	Algorithm Used	Author
Waste management	Supervised	Instance Based Learning	KNN (K-nearest neighbour)	Parag Kedia et al
	Unsupervised	Clustering Methods	a. K-means b. Apriori - algorithm	
Traffic congestion	Supervised Learning	Bayesian Methods	Dynamic Bayesian network	Case Study by CSIRO, Canberra
Emergency Service (Accidents)	Supervised	Classification	Support Vector Machine	CSIRO, Canberra

V. PROPOSITIONS FOR A SMART CITY

While formulating policies for smart city planning:

- ✓ Adequate arrangements should be made for accommodation and implementation of ML techniques.
- ✓ Measures should be taken in the infrastructure-design phase to ensure uninterrupted streamlining & broadcasting of real time data to the command-control centre where ML methods are carried out.
- ✓ The ICT (Information & Communication Technology) & IoT framework of a smart city should be robust enough to generate & transmit large amounts of relevant data required for training of the ML algorithms for increased accuracy.

VI. CONCLUSIONS AND FUTURE COURSE

Machine learning techniques if integrated within the collective ICT and IOT infrastructure of a smart city project can have a tremendous uplifting effect on the quality of life of people and the environment. The disbursal of services and utility mechanisms of a smart city can be optimized and made more efficient using the ML techniques providing the citizens with a more robust and response driven emergency acknowledgement and utility model. Overtime the use cases for implementation of ML within a smart city can only increase thus leaving the door open for more innovative & efficient models, be it for public transportation sector, energy sector (for smart grids) or for environmental engineering and ecosystem conservation of a city. With Machine Learning already being implemented in various countries for urban planning and decision-driven policy making along with the arrival of hybrid and swarm intelligence algorithms, the future scope & applications of this field appear to be humongous.

However, some challenges such as:

- ✓ Cohesive integration of Machine learning and IoT infrastructure on a macro scale based on the identification of citizen needs, &
- ✓ The protection of huge amounts of data (both public and private), need to be addressed.

VII. REFERENCES

[1]. Pedro Domingos "A Few Useful Things to Know about Machine Learning.", Communications of the ACM, 55 (10), 78-87, 2012.

[2]. http://www.dummies.com/programming/big-data/data-science/3-types-machine-learning.

[3]. http://www.aihorizon.com/essays/generalai/supervised_unsupervised_machine_learning.htm

[4]. https://medium.com/towards-data-science/types-of-machine-learning-algorithms-you-should-know-953a08248861

[5]. https://en.wikipedia.org/wiki/Semi-supervised_learning

[6]. Sam Musa. "Smart City Roadmap". http://www.academia.edu/21181336/Smart_City_Roadmap

[7]. Smart Cities Mission Document , Ministry of Housing and Urban Affairs, Government of India. http://smartcities.gov.in/content/innerpage/what-is-smart-city.php

[8]. CSIRO , (2015). https://research.csiro.au/data61/advanced-data-analytics-in-transport-machine-learning-perspective/

[9]. Kedia Parag, "BIG DATA ANALYTICS FOR EFFICIENT WASTE MANAGEMENT",IJRET, eISSN: 2319-1163 | pISSN: 2321-7308

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

A Digital Video Copyright Protection Scheme using Colored Watermark Embedding Algorithm : CWEA

Jabir Ali, Satya Prakash Ghrera
Computer Science & Engineering Jaypee University of Information Technology Solan, Himachal Pradesh, India

ABSTRACT

Digital Watermarking is an important and finest method of protecting the copyright of the digital media. In this paper, a secure and robust digital video watermarking algorithm is proposed, that is Color Watermark Embedding Algorithm (CWEA). This algorithm has two important parts. First, YCbCr color format is used to insert the variable size watermark. Second, embedding of detail coefficients of LUMINANCE (Y-luminance) of watermark into the detail coefficients of CHROMINANCE (Cb and Cr- chrominance) of identical frames (I-Frames) of digital video. Watermark data is inserted into the detail coefficients in an adaptive manner based on the energy of high frequency. We have performed number of tests for many video-frame manipulations and attacks. All these tests are also performed on CWEA and it provides good results. In this paper, non-blind and semi-blind watermarking systems are used where non-blind watermarking mechanism has been proved to be robust, imperceptible and efficient to protect the copyright of H.264 and MPEG-4 coded video within the video retrieval system.

Keywords : CWEA, DWT, LUMINANCE, CHROMINANCE, I-Frame, Detail Coefficient, H.264, MPEG-4.

I. INTRODUCTION

The future growth of the domestic digital copyright protection products basically depends on Real Network and Microsoft Windows Media. Large numbers of these products follow only the protection of copyright of electronic publications, magazines, journals and static images. But still, the necessity of such a platform exists which can apply the copyright protection on digital videos professionally.

A. Video Watermarking

Nowadays, a large amount of multimedia data has been exchanged over the internet and many internet users are sharing their images, videos and audios. Security provided to protect shared and transferred data over internet is not enough and many people are not aware of this security issue and control access techniques. So there are various approaches for data security such as steganography, fingerprinting, copyright protection [1, 2], and so on. In this paper, we have considered the Copyright Protection technique of data security in Digital Videos. For this, we are inserting a Digital Watermark or digital pattern (an image) inside digital video frames. Some important aspects of watermark systems include Robustness (quality of the watermark should not be degrade due to any attack, whether the attack is intentional or unintentional), Imperceptibility (the data embedded inside the video frames that should not be visible), Capacity (the number of bits that can be hidden), Security (to control the illegal use of data) and computational complexity of the embedding and detection process [4-6].

B. Types of Watermarking

For detection process of the watermark, Data hiding techniques can be divided into three categories: Blind, Semi-blind, and non-blind. In this paper, non-blind and semi-blind watermarking systems are used [7-9].

Blind systems do not require the original host data (image, video, Audio etc.) to extract the watermark. For the sake of security, some additional information (e.g., a secret "key") may be needed in order to detect or decrypt the watermark. The key can be suppressed if additional security is not needed. Meanwhile, non-blind systems need to have the original host data at the decoder in order to decode the watermark. A semi-blind watermarking system can be imagined as a communication system with side information. In such types of systems, the watermark or some information about the original host data (but not the entire host data) is required to extract the watermark sequence. This classification of watermarking system has shown in Table 1.

In this paper, digital video copyright protection using DWT is proposed. Extraction of I-frame, watermark preprocessing, watermark embedding and extraction, piracy tracking and various others are implemented in this paper. In this paper, work is done to build a professional platform which coalesces watermark embedding and extraction with source video tracking. This platform embeds a watermark in video for copyright protection to satisfy the client's requirements and then, embedded watermark is extracted from the original source video to authenticate the copyright. It also compares extracted watermark with the original watermark to verify whether the product is authorized or not.

In the literature review, we have studied that an attacker may be crack, damage or detect watermark with the help of some possible algorithms [10]. But in CWEA, it is very difficult to detect the original pattern of inserted watermark. In this paper, we have improved the robustness and security of inserted digital pattern or watermark and transparency of watermarked media. TABLE 1 has shown the classification of watermark system.

Table 1. Classification of watermark system.

Criteri	Class	**Brief Description**
Domain Type	Pixel	Manipulate the Pixels values to embed the watermark
	Transform	Modified the coefficients of Transform Domain to embed the watermark. These are the some popular Transform:- Discrete Cosine Transform (DCT) Discrete Wavelet Transform (DWT) Discrete Fourier Transform (DFT) Principle Component Analysis (PCA)
Watermark Type	PRNS	Detecting the presence or absence of a watermark statistically. A PRN Sequence is generated by feeding the generator with a secret seed.
	Visual	The visual quality of embedded watermark is evaluated.
Information Type	Non-blind	Both the original image and Secret key required.
	Semi-blind	Watermark and Secret key are required.
	Blin d	Only secret key required.

C. Possible Attacks

Attacks create disturbances in original content. There is a list of some possible attacks in Digital video Watermarking technique.

1) **Adaptive Noise:** An Additive noise typically forces to increase the threshold values at which the correlation detection process works.
2) **Filtering:** There are two types of filtering Low pass and high pass filtering. Low pass filtering does not introduce the considerable degradation in watermarked image or watermarked signals but it can affect the performance.
3) **Cropping:** This is a very common attack to analyze the pattern inserted in the multimedia document and in this type of attack an attacker or intruder just select the small portion of the watermarked object and try to find the pattern that is inserted inside that particular multimedia document.
4) **Compression:** Sometimes compression is known as the unintentional attack which appears very often in multimedia applications [11].

Rotation and Scaling, Multiple Watermarking and Statistical Averaging are also the types of attacks which can be performed on digital video.

In Figure 1. we have shown the general diagram of video watermarking. Here, we have a colored watermark and original video file and applying some watermarking algorithm on both, and get the final watermarked video. After applying de-watermarking algorithm on watermarked video and get the extracted watermark and original video.

In this paper, a new digital video watermarking algorithm Color Watermark Embedding Algorithm (CWEA) is proposed. CWEA has two important parts. First, YCbCr color format is used to insert the variable size watermark. Second, Embedding of detail coefficients of LUMINANCE (Y) of watermark into the detail coefficients of CHROMINANCE (Cb & Cr) of identical frames (I-Frames) of digital video. Data is inserted into the detail coefficients in an adaptive manner based on the energy of high frequency.

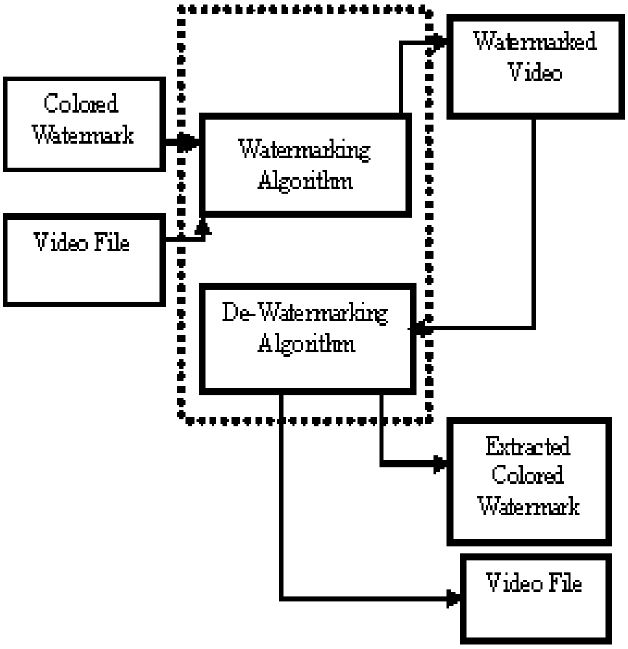


Figure 1. Block Diagram of Video Watermarking

Rest of the paper is organized as follows: In Section II, proposed scheme is illustrated which explains the embedding and extraction of watermark from video sequence. Section III shows some experimental results and evaluates the performance of the proposed technique. At the end, conclusions are drawn in Section IV and provide some future work directions.

II. PROPOSED WATERMARKING TECHNIQUE

In proposed method, A YCbCr color space model [12] is used to embed the colored watermark signal (256 x 256) in a video file to increase the imperceptibility of watermarked video and detected watermark. For embedding the colored watermark we take apart that image as in Luminance and Chrominance components. Now Only Luminance components have to be embedded inside the video file to get the high PSNR and low MSE.

Where, luminance and chrominance contains the information about the brightness and colors respectively.

For the result analysis, we have taken two video sequences, namely, ‘car_race.mp4’ and ‘wakna_road.mp4’ video sequence. After applying scene changed algorithm on these video sequences, we obtain 77 and 97 scene changed frames respectively. Before embedding the watermark inside the original media (original video), we have to preprocess the watermark and input video.

Watermark Pre-process

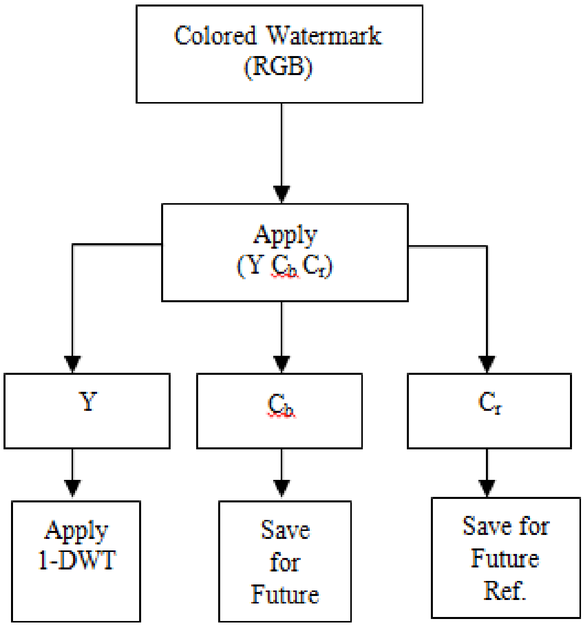


Figure 2. Watermark preprocesses to generate Encrypted watermark

For the process of watermark embedding, at first we have to take separately the colored watermark into YCbCr color space model. Where, Y (luminance) and Cb & Cr (chrominance) components contains the information about brightness and its color respectively.

With the aim of high-quality copyright protection, the information of watermark, which is we are embedding in original source video should be as small as possible. If the information of watermark embedding is too short, it will diminish the visual effect of watermark embedded video. As a result, the interval time should not be too small. In general, the interval time is less than 1s.

MPEG-4 video coding standard has the sequence of GOPs (group of pictures). The length of GOP entails at least one I-frame image but in the main, each GOP has 12 frames of images. These 12 frames of images includes: one I-frame, three P-frames and eight B-frames, managed as IBBPBBPBBPBB. Where I is an identical frame, P is predictive frame (that has the information of previous frame), B is Bidirectional frame (that has the information of previous frame and future frame or next frame.)

In Figure 2 we have original watermark juit.jpg and applied DWT to find out the detail and approximation coefficients. In the next step applied bit plane slicing on the watermark to convert it into 8-bit plane and for the next step place the bit plane side by side and finally decompose the image with a secret key to make it encrypted watermark. Fig. 3 is showing the preprocess action on a video file to find out those coefficients where we have to embed the watermark bits. So here we have taken a video file and applied Scene changed detection algorithm [13, 14].

After getting scene changed frames applied YCbCr color space model on each frame and then apply 2-Level DWT only on Cb and Cr components.

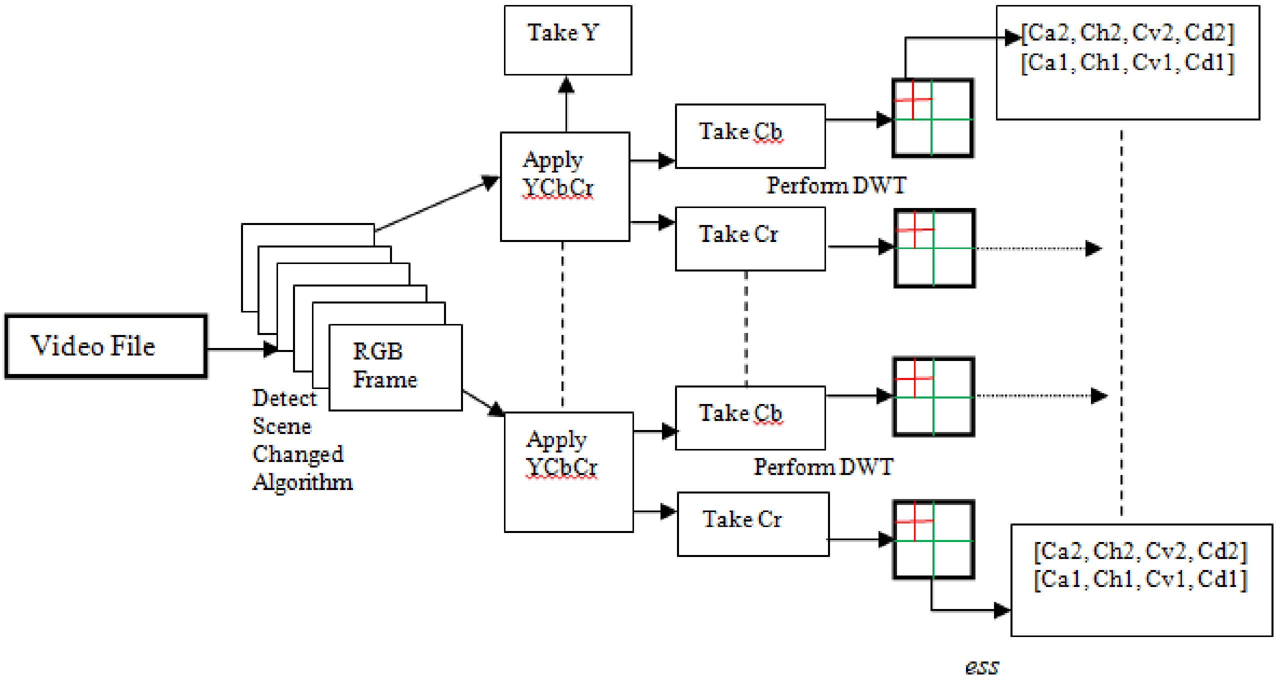


Figure 3. Video Pre-process

In Figure 3 we applied a scene changed detection algorithm [13, 14] on input video sequence and get the non-overlapping GOP (group of pictures). Each GOP has at least 1 I frame. Select that I frame with the help of identical frame selection scheme. Which said that identical frames are self dependent frames means they are not dependent on the previous (P-frames) and bidirectional (B-frames) frames.

After getting the Identical frames (I-frames) applied YCbCr color space model on each I frame. Here in the video preprocess we are not considering Luminance (Y) components and taking only Chrominance component for embedding the watermark information because luminance components having the maximum information of the frame so these components are very sensitive for embedding the watermark information.

After getting the chrominance components Cb and Cr apply 2-level DWT scheme on Cb component only. Finally we get the target frames where we have to embed the watermark information.

Watermark embedding

1. Apply a scene changed detection algorithm [6] on the original video sequence (O_{video}) and then divide the each scene into non-overlapping group of pictures. Each group of pictures has an Identical frame (I). Select all I frames from input video for embedding the watermark.

$$WmI_i = k \times (Lf_2) \; + q \times (Wm_2)$$

Where

WmI_i is watermarked I frame.

Lf_2 is low frequency approximation of original frame. Wm_2 is low frequency approximation of watermark image. k & q are scaling factors.

2. Take each I-frame and apply YCbCr color format on each frame.

Where Y= 0.299R + 0.587G + 0.114B

Cb = 0.564 (B – Y)

$Cr = 0.713 (R - Y)$

$Cb + Cr + Cg = 1$

$Cg = 1 - (Cb + Cr)$

3. Apply 2-level DWT on CHROMINANCE (Cb & Cr-chrominance) of each frame and store LUMINANCE (Y-luminance) for future reference.

   $DWT (Cb) = [Cai,Chi,Cvi,Cdi]_o$

   $DWT (Cr) = [Cai,Chi,Cvi,Cdi]_o$

   Where i=1, 2.

4. Let W = Digital Watermark Color Image. Take YCbCr of RGB-frame.
5. Apply 2-level DWT on LUMINANCE component and get detail coefficients of LUMINANCE .

   $DWT (Y) = [Cai, Chi, Cvi, Cdi]_w$

   Where i= 1, 2.
6. To embed the watermark in each I-frame, add detail coefficients of LUMINANCE of watermark with the detail coefficients of CHROMINANCE of I-frame of original video.

   Now, for Cb

   $[Mod\ Chi]_{Cb} = [Chi]_o + [Chi]_w$

   $[Mod\ Cvi]_{Cb} = [Cvi]_o + [Cvi]_w$

   Now, for Cr

   $[Mod\ Chi]_{Cr} = [Chi]_o + [Chi]_w$

   $[Mod\ Cvi]_{Cr} = [Cvi]_o + [Cvi]_w$

   Where i= 1, 2.

7. Now, mod Chi and mod Cv_i are the modified coefficients of CHROMINANCE of watermark inserted identical frame.
8. Take IDWT of watermark inserted CHROMINANCE components of identical frames. Finally, get the modified CHROMINANCE (mod Cb and mod Cr) of I-frame.
9. Take LUMINANCE (Y) from step-2 and add this with the modified CHROMINANCE (mod Cb and mod Cr) and get watermark inserted I-frame. Convert YCbCr format to RGB.
10. Combine all the watermark inserted I-frame with the remaining frames and get watermarked video for transmission/braodcasting.

Watermark detecting

1. Apply a scene changed detection algorithm [6] on the watermarked video sequence (W_{video}) and then divide the each scene into non-overlapping group of pictures. Each group of pictures has an Identical frame (I). Select all I frames from input watermarked video for detecting the watermark.
2. Take the watermarked frame (W_f) (Identical frame) and original identical frames (I_i). Apply YCbCr color format on both.
3. Apply DWT of CHROMINANCE of both watermarked frame and identical frames.
4. Subtract detail coefficients of CHROMINANCE of watermarked frame from the detail coefficients of original I-frame which are mod Ch_i, mod Cv_i, Ch_i and Cv_i respectively.

   Now, for Luminance

   $[NewCh_i]_{dw} = [mod\ Ch_i]_{ew} - [Ch_i]_o$

   $[NewCv_i]_{dw} = [mod\ Cv_i]_{ew} - [Cv_i]_o$

   Where i= 1, 2.
5. Take IDWT of detail coefficients of detected LUMINANCE and add this detected LUMINANCE with the original CHROMINANCE and get the YCbCr format of detected watermark.
6. Calculate cross correlation between new values of detected watermark and the original watermark.
7. If correlation = high

Then, Stop the execution. Detected watermark is similar to original watermark.
else
Take both detail coefficients together and repeat from step 3 (initially i=1 and in repetition process the value of i=2).
Else if

Take 2-level detail coefficients and repeat from step 3 until the detected watermark will get similarity with original watermark.

else
Watermark not found.

Figure 4 shows the procedure of embedding the watermark on to the video. Video sequence taken and applied scene change detection algorithm to get the identical frames. After extracting the scene changed frame, YCbCr color space model was introduced on each identical frame.

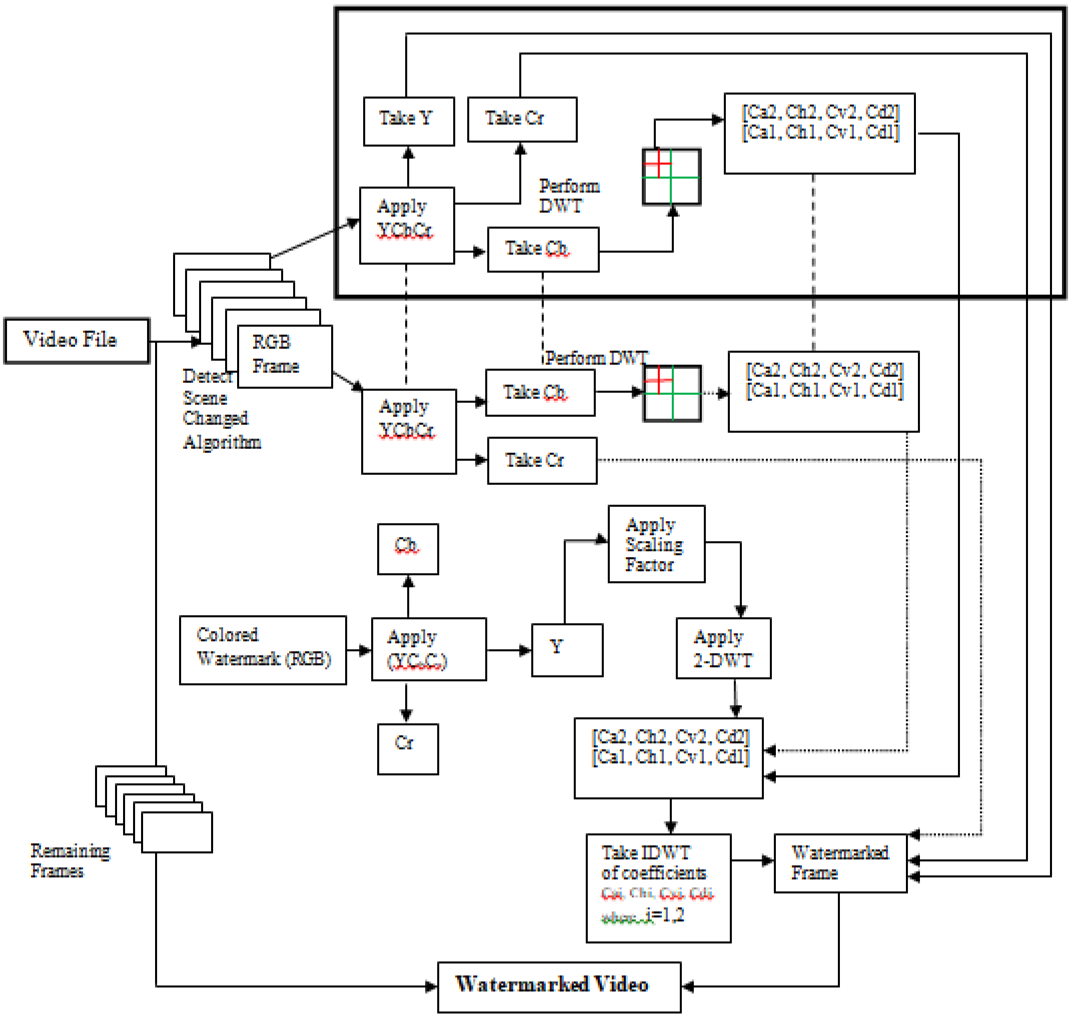


Figure 4. Watermark embedding process

In this color space model Y represent Luminance and CbCr represents chrominance, where Luminance depicts the information of brightness and CbCr about color of the frame. Blue chroma (Cb) was taken and applied with 2-level DWT (Discrete Wavelet Transform).

For the next process of embedding the watermark in to the video, colored watermark was taken and applied YCbCr. Now by HVS (Human Visual System) [15] indicated that Luma (Y) Component has the maximum information of the watermark. So, firstly scaling factor was applied on the Matrices of Luma to reduce the size of the matrices and then applied a 2-level DWT [16-18]. Then the coefficients of blue chroma of original frames were merged with the coefficients of Luma of watermark. Lastly, inverse of Discrete Wavelet Transform was applied and added to the luma and Cr of the original frame with the merged coefficients.

III. EXPERIMENTAL RESULTS AND ANALYSIS

To implement this technique we have used original video 'wakna road' and 'car race' at the dimension of 320 x 240 and 640 x 360 respectively and the size of original watermark image is 256 x 256. Figure 5, 6 & 7 shows original video frames (wakna road, car race and foreman respectively) and their YCbCr components. Figure 8 shows original watermark image and its YCbCr components. The value of scaling factor k is varied from 0.2 to 0.6 for embedding the watermark into the original video frame and q will be 0.6 to 1.0 for all frames. Where 'k' is used for luminance (Y) of original watermark and 'q' is used for chroma (Cb) of original video frame.

When 'k' is 0.3 and 'q' is 0.6 we get the adequate result for embedding and extracting process also.

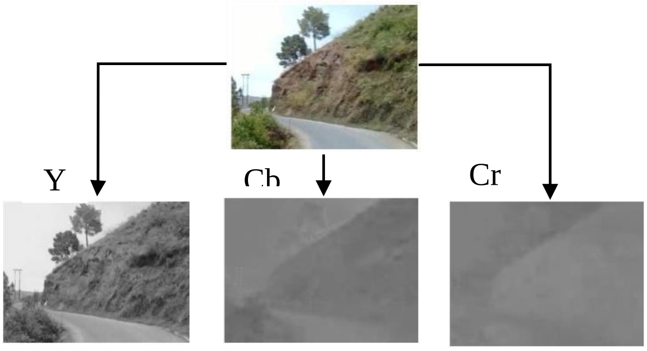


Figure 5. Wakna road original video frame and its YCbCr color components

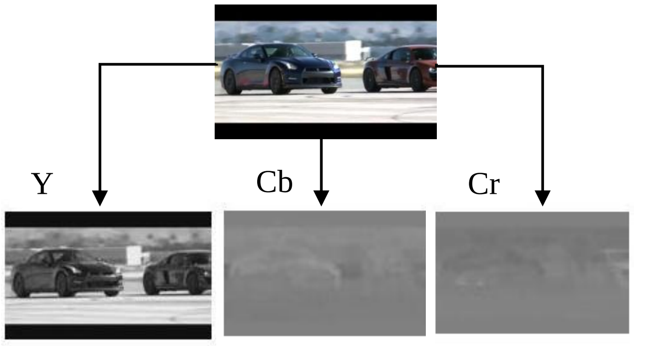


Figure 6. Car race original video frame and its YCbCr color components

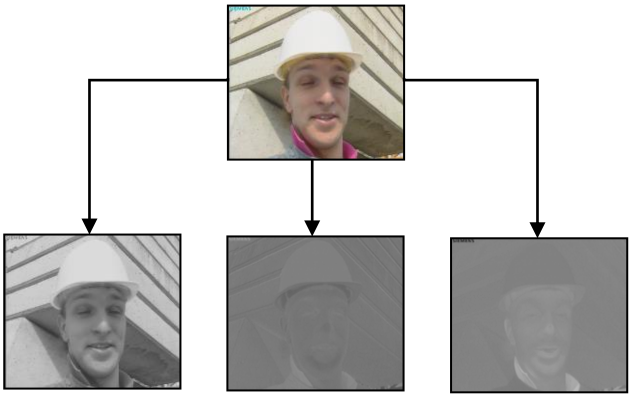


Figure 7. Foreman original video frame and its YCbCr color components

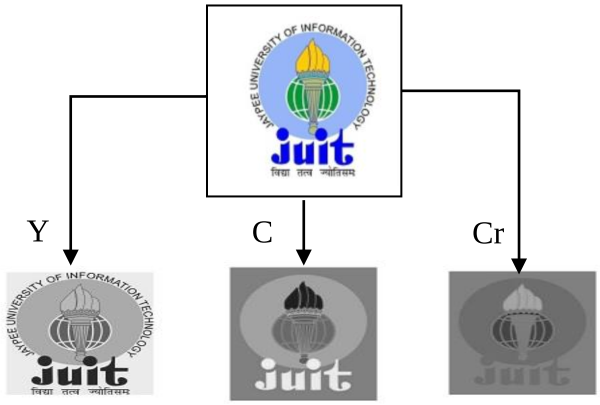


Figure 8. Original watermark and its YCbCr color components.

Embedding watermark into I-frame is more suitable and robust because I frame is independent and it has its own information only. The watermark embedding algorithm based on I-frame, entrenches watermark only in I-frame but simultaneously it guarantees the knowledge of embedded watermark to each GOP.

In Figure 9 we have explain the module of 2-level 'haar' discrete wavelet transform. In this figure we have 4 images, in which the lower right corner is having de-composition image and this is showing low frequency and high frequency parts of original watermark. We have chosen the 2-level low frequency image as the watermark because as we can see in this figure this part is having the maximum information of the original image.

In Figure 10. We have shown the procedure of embedding and extraction of the watermark. Here in this figure we have taken an original video frame and luma part of original watermark. After embedding the watermark we add some noise in watermarked frame. After extraction of watermark we add the original Cb&Cr of original watermark with the extracted watermark and we get a good quality watermark with the PSNR of 46.22.

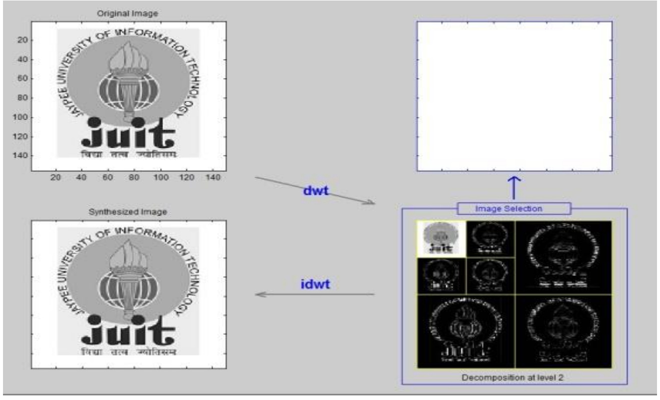


Figure 9. 2-Level DWT of Luma (Y) of original watermark.

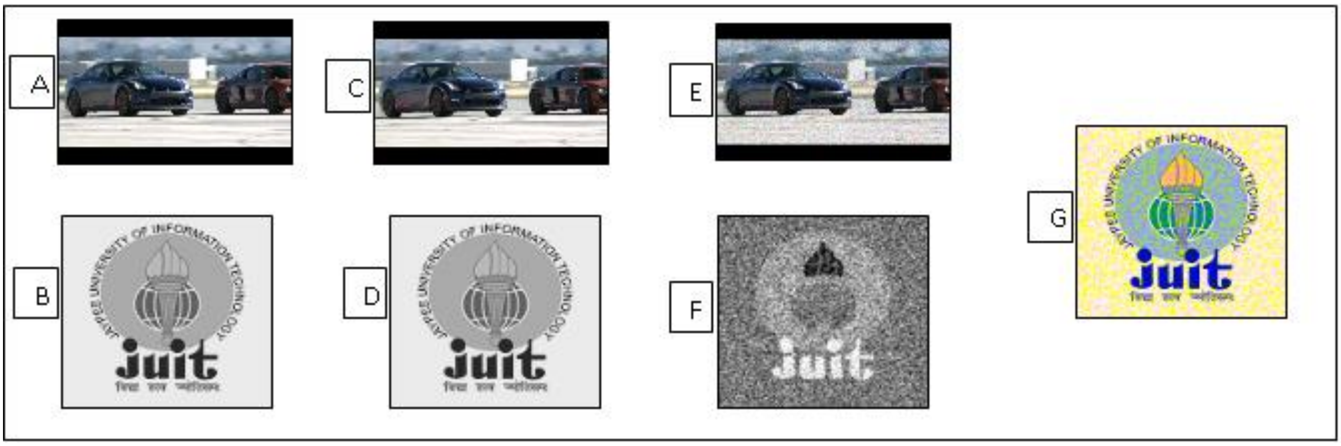


Figure 10. Watermark extraction

(A)Original Car race frame (B) Original luma of colored watermark (C) Watermarked Car race frame (D) Extracted Watermark (E)Noisy watermarked car race frame (F) Extracted watermark from noisy frame (G)Added (Cb & Cr) of original watermark with extracted watermark

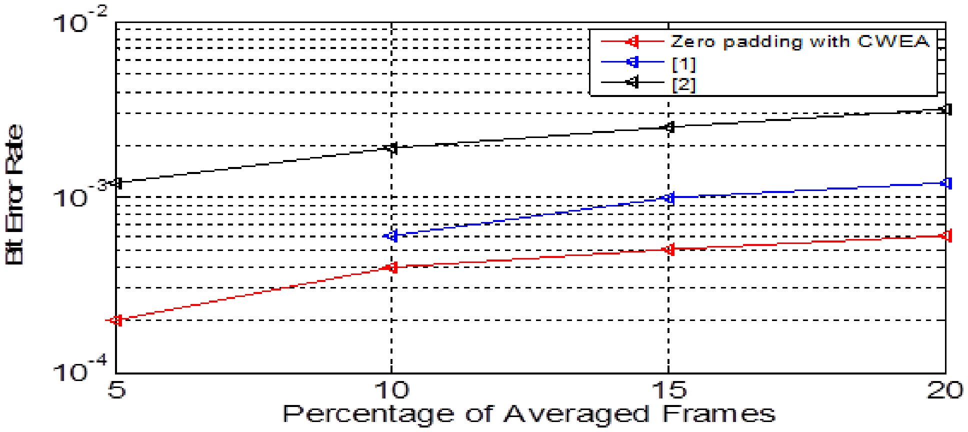


Figure 11. BER (log scale) under temporal attack (Frame Averaging)

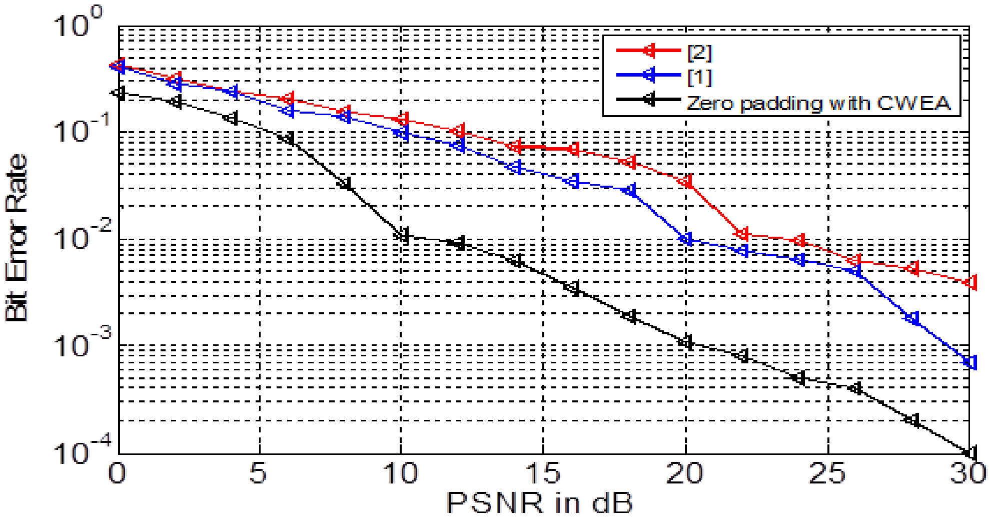


Figure 12. BER (log scale) under spatial attack (Uniform noise)

In figure 11 we have shown the results against temporal attack (Frame Averaging). Where BER (log scale) on x-axis and percentage of averaged frames on y-axis are showing the result graph and a red indicator is showing our results. In figure 12 we have shown the results against spatial attack (Uniform Noise). Where BER (log scale) on x-axis and PSNR on y-axis are showing the result graph and a black indicator is showing our results. In figure 13 we have shown the averaged PSNR of car race video after embedding the watermark in all I-frames.

IV. CONCLUSION

The proposed approach is more proficient because the quality of extracted watermark is better than A.K. Verma et. al. in “Robust Temporal Video Watermarking using YCbCr Color space in Wavelet Domain” in terms of PSNR and BER. We have taken Cb of the video frame for embedding the watermark because, as per HVS (Human Visual System), we cannot identify the changes in Cb (Blue chroma) because of its low resolution. Compression attack will not degrade the quality of the embedded watermark because we embed the same watermark at every scene changed frame so at the time of extraction we can extract that watermark from the couple of frames and finally collect the entire extracted watermark and find out the best possible pattern on the basis of image collaboration technique. Secondly, we added the (Cb & Cr) of original watermark with extracted watermark image that is giving a best quality of extracted watermark. Thirdly, it is hard to know the spot where it is inserted, because it is inside the blue chroma (Cb). Another advantage of this technique is that the quality of the watermarked video also will not degrade because we have embedded the watermark inside the blue chroma (Cb) that has a very low sensitivity. Video watermarking is an essential need of copyright protection and a lot of research is still going on to find out the new methods for security and privacy of the multimedia contents. Current methods for video copyright protection techniques are extended form of image watermarking and there is a great scope of innovation. Research can be carried out to establish new strategies for digital video copyright protection.

V. REFERENCES

[1]. C. Langlaar, et al, "Watermarking Digital Image and Video Data", in IEEE Signal Processing Magazine, September 2000.

[2]. I. J. Cox, et al, "Digital Watermarking and Steganography," in 2nd ed. San Mateo, CA, USA: Morgan Kauffman, 2008.

[3]. T. Al-Khatib, et al, "A Robust Video Watermarking Algorithm," in Journal of Computer Science, 4(11):6~9.1280, 2004.

[4]. Wang, Y., et al, "A Blind MPEG-2 video watermarking robust against geometric attacks: a set of approaches in DCT domain," in IEEE Trans. Image Process, 15(6), 1536–1543, 2006.

[5]. Lin, et al, "An embedded watermark technique in video for copyright protection," in Proc. Int. Conf. Pattern Recog. 4, 795–798, 2006.

[6]. J. Shieh, et al, "A semi-blind digital watermarking scheme based on singular value decomposition," In Computer Standard International, vol. 28, no. 4, pp. 428–440, Apr. 2006.

[7]. H. Khalilian, et al, "Multiplicative video watermarking with semi- blind maximum likelihood decoding for copyright protection," in Proc. IEEE Pacific Rim Conf. Commun., Comput. Signal Process, pp. 125–130, Aug. 2011.

[8]. S. Gandhe, U. Potdar, and K. Talele, "Dual watermarking in video using discrete wavelet transform," in Second International Conference on Machine Vision, ICMV ’09, pp. 216 –219, Dec. 2009

[9]. A. Murshid, et al, "Blind compressed video watermarking in DCT domain robust against global geometric attacks. Proceeding of the International Cryptology Workshop and Conference, 2008.

[10]. S. Bhattacharya, T. Chattopadhyay, and A. Pal, "A survey on different video watermarking techniques and comparative analysis with reference to h. 264/avc," in IEEE Tenth International Symposium on Consumer Electronics, ISCE’06. IEEE, pp. 1–6, 2006.

[11]. J. Y. Park, J. H. Lim, G. S. Kim, C. S. Won, "Invertible Semi-fragile Watermarking Algorithm Distinguishing MPEG-2 Compression from Malicious Manipulation", in International Conference on Consumer Electronics, pp. 18-19, June 2002.

[12]. A. K. Verma, M. Singhal, C. Patvardhan, "Robust Temporal Video Watermarking Using YCbCrColor Space in Wavelet Domain", In 3rd IEEE International Advance Computing Conference (IACC), 978-1-4673-4529, Dec. 2013.

[13]. B. Yeo, et al, "Rapid scene change detection on compressed video," in IEEE Trans. Circuits Syst. Video Technol., vol. 5, no. 6, pp. 533–544, Dec. 1995.

[14]. B Shahraray "Scene Change Detection and Content-Based Sampling of Video Sequences," in SPIE 2419 (Digital Video Compression: Algorithms and Technologies): 2–13, 1995.

[15]. Y.Yang, M. Yang, S. Huang, "Multifocus Image Fusion Based on Extreme Learning Machine and HumanVisual System" in IEEE Access Volume:PP, Issue: 99Pages: 1-1, 2017

[16]. E. Ganic, Eskicioglu, "Robust DWT-SVD domain image watermarking: embedding data in all frequencies," In Proceedings of the ACM Multimedia and Security Workshop, pp. 166–174, 2004.

[17]. X. Niu, S. Sun, "A new wavelet-based digital watermarking for video," In Proceedings of the IEEE Digital Signal Processing Workshop, pp. 1–6. Texas, 2000.

[18]. F. Hartung and B. Girod, "Watermarking of uncompressed and compressed video," in Signal Process., vol. 66, no. 3, pp. 283–301, May 1998.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

A Robust and Secure Method of Copyright Protection for Digital Videos using Split Watermark Embedding Algorithm : SWEA

Jabir Ali, Prof. S.P. Ghrera
Computer Science & Engineering Jaypee University of Information Technology Solan, Himachal Pradesh, India

ABSTRACT

In the current scenario, Copyright Protection of the digital media is a serious issue and watermarking is an important method of protecting the intellectual property and copyright of the digital media. In this paper, we proposed a robust and secure algorithm SWEA (Split Watermark Embedding Algorithm) for digital videos. SWEA has three important parts, first one is splitting of original watermark into small pieces, second is embedding of the small pieces of watermark into the detail coefficients of identical frames (I Frames)of digital video and the third is extracted these small pieces and merge them in an efficient manner. Watermark data is inserted into the detail coefficients in an adaptive manner based on the energy of high frequency. The proposed algorithm has undergone various attacks, such as compression, uniform noise, Gaussian noise frame repetition and frame averaging attacks. The proposed algorithm, sustain all the above attacks and offers improved performance compared with the other methods from the literature.

Keywords : SWEA, DWT, I-Frame, Detail Coefficient, H.264, MPEG-4., SCD

I. INTRODUCTION

The future growth of the domestic digital copyright protection products basically depends on real network and Microsoft windows media. A huge amount of multimedia data has been exchanged with the internet but the security is provided to protect the data is not enough. There are various approaches for data security such as steganography [1], fingerprinting and copyright protection etc. In this paper we have introduced a new copyright protection technique i.e. SWEA. For this, a block of digital watermark or digital pattern (text, image, audio or video) is inserted inside Digital Video Frames. In this paper, digital video copyright protection using DWT [1]-[5] is proposed. Extraction of I-frame, watermark pre-processing, watermark embedding and extraction, piracy tracking are implemented. In the case of SWEA, it is very hard to detect the original pattern of inserted watermark. In this paper, the robustness and security of inserting a digital pattern or watermark and transparency of watermarked media is improved using SWEA.

In Figure 1 we are showing the block diagram of our algorithm. Here we have applied SWEA (Split watermark embedding algorithm) algorithm on original watermark and SCD (Scene changed detection) algorithm on original video. Now to extract the watermark, apply the de-watermarking algorithm on watermarked video.

The rest of the paper is organized as follows: In Section II, the Proposed Scheme is illustrated which explains the embedding and extraction of the watermark from video sequence. Section III shows some experimental results and evaluates the performance of the proposed technique. At the end, conclusions are drawn in Section IV and provide some future work directions.

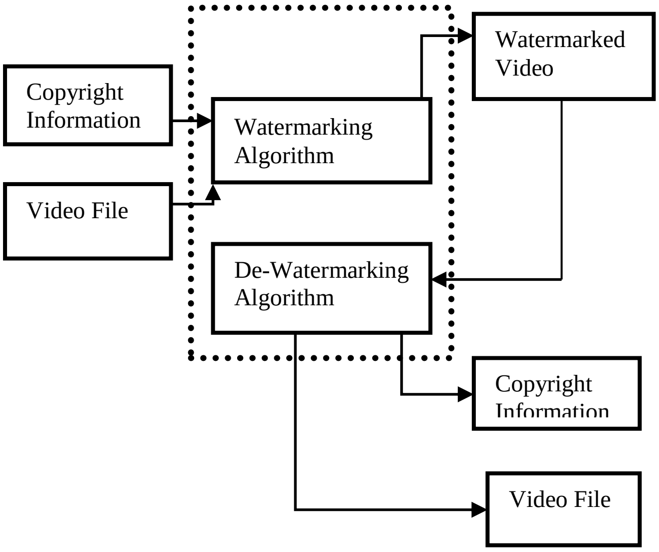


Figure 1. Block Diagram of Video Watermarking

II. PROPOSED WATERMARKING TECHNIQUE

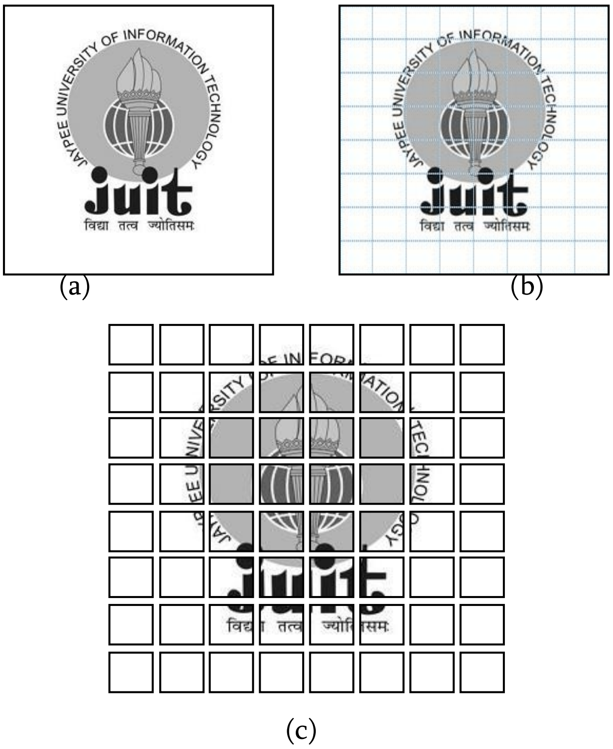
A gray level image (256 x 256) is used as a watermark signal. For embedding the watermark we have taken 'Foreman' video sequence. After applying scene changed algorithm [6] on these video sequences, 77 scenes changed frames are obtained. Before embedding the watermark inside the original video, preprocess the watermark and input video.

2.1. Watermark Pre-Process

For the process of watermark embedding, scale the watermark to a particular size with the help of equation 1. Fig. 1 shows an original watermark juit.jpg (256 x 256).

$$\left(4^n \leq m;\ n > 0\right) \qquad (1)$$

Where m is the total no. of scene changed frames and 4^n is the total number of split watermark in which n is an integer. In this case, the watermark will be divided into 4^n smallimages using SWEA that are shown in Figure 1.



Figuer 2. Watermark Pre-process
(a) Original Watermark. (b) Split watermark in 8x8 using proposed algorithm. (c) 64 small pieces of original watermark.

The MPEG-4 video coding standard has the sequence of GOPs (group of pictures) and each GOP has 12 frames of images which are managed as IBBPBBPBBPBB where I, P and B areidentical frame, predictive frame and Bidirectional frame respectively.

2.2. Video Pre-Process

In Figure 2, a scene changed detection algorithm is applied on input video sequence and get the non-overlapping GOP [7], [8]. Select the I-frame with the help of the identical frame selection scheme. After getting the I-frames, apply 2-level DWT and embed the watermark information.

2.3. Watermark Embedding Algorithm

1. Apply a scene changed detection algorithm [6] on the original video sequence (O_{video}) and then divide the each scene into non-overlapping

group of pictures. Each group of pictures has an Identical frame (I). Select all the I-frame from input video for embedding the watermark.

$$WmI_i = k\times\left(Lf_2\right)\ + q\times\left(Wm_2\right)$$

Where WmI_i is watermarked I frame, Lf_2 is low frequency approximation of original frame, Wm_2 is low frequency approximation of watermark image, k & q are scaling factors.

2. Let W = Digital Watermark Image (256 x 256). Generate small blocks of digital watermark using the proposed algorithm in which $4^n \le$ m. Where 4^n is the total number of small blocks of the watermark (W1, W2...... W $_{(4^n)}$). N is the number of rows or columns of split watermark images and m is the total number of scenes changed identical frames.

3. Size of watermark block (x,y) = $\left[\left(\sqrt{4^n}\right)*\left(\sqrt{4^n}\right)\right]$

4. while m ≥ 4^n, I_i = W_{4^n} I_i, I_i is Original identical frame (where small blocks of watermark has to be embedded), W_{4^n} I_i = watermarked identical frame, i = 1,2,3……………m, 4^n = total number of watermark pieces and n > 0.

5. Take 2 level 2 dimensional DWT of I_i

```
for i = 1 : m
 [Ca_i, Ch_i, Cv_i, Cd_i ] = DWT2(I_i, "haar")
 j=i+1
[Ca_j, Ch_j, Cv_j, Cd_j ] = DWT2(Ca_i, "haar")
```

6. Insert first block of watermark into detail coefficients of step 5 which are Ch_i, Cv_i, Ch_j, Cv_j. Now , mod Ch_i,mod Cv_i, mod Ch_j, mod Cv_j are the modified coefficients of watermark inserted identical frame.

Take IDWT of watermark inserted identical frames. Finally, get the watermark inserted frame. Let Watermark inserted frames = W_f, where f=1,2,3……………………2^n.

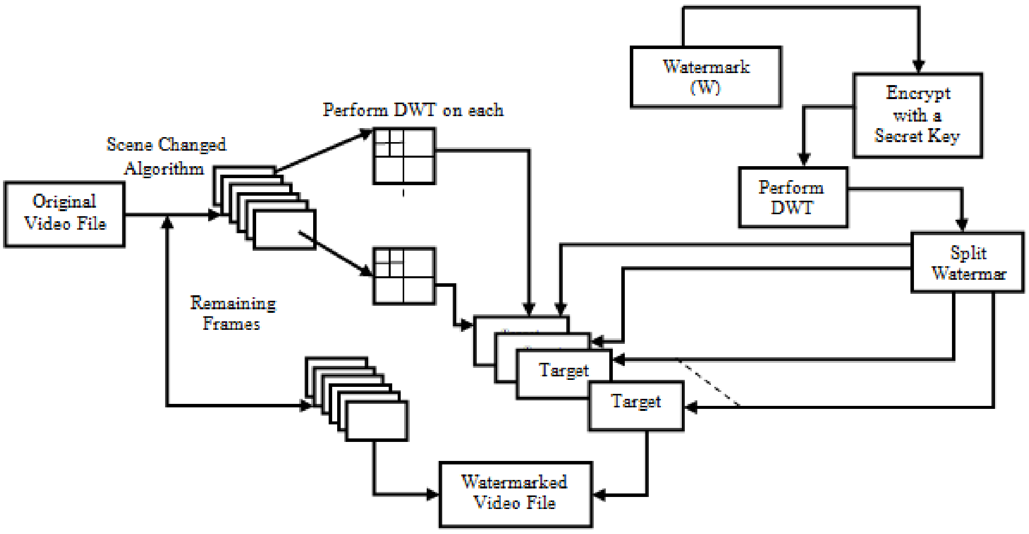


Figure 3. Video Pre-process

2.4. Watermark Detection Algorithm

1. Apply a scene changed detection algorithm [6] on the watermarked video sequence (Wvideo) and then divide the each scene into non-overlapping GOPs. Select all I frames from input watermarked video. Take the watermarked frame (Wf) (I-frame) and original identical frames (Ii). Apply DWT on both watermarked image and identical frames.

2. Subtract first level detail coefficients of the watermarked frame from the first level detail coefficients of original I frame which are mod Chi, mod Cvi, Chi and Cvi respectively.

Now, NewChi = mod Chi - Chi
NewCvi = mod Cvi – Cvi

3. Calculate cross correlation between the new values of detail coefficients (NewChi and NewCvi) and the detail coefficients of original watermark.

4. If correlation = high
Then, Stop the execution. Detected watermark is similar to original watermark.
else

Take both detail coefficients together and repeat from step 3.
else if
Take 2-level detail coefficients and repeat from step 3 until the detected watermark will get similarity with original watermark.
else
Watermark not found.

2.5. Merging small pieces of watermark images

Propose watermark split and detection algorithm on the watermarked I-frames, collect all the small pieces of watermark picture and embed one after another. Vertical and horizontal axes of final matrix picture have the same rule. Now Scan xy where x = number of rows and y = number of columns. These pictures are stored in the program folder with the same name. With the help of 'for' loop all grayscale images are used. Total no of scene changed frames are 77 for foreman video.

Now with the SWEA- : 43 ≤ 77 for n =3. Then, the watermark is divided into 4n smallblocks are 64.
Now, size of watermark block for foreman video = 8 x 8.

Now applied watermark detection algorithm on the watermarked I frames and collect all the small pieces of watermark picture. When one picture stops, a new picture always starts. It means that these pictures are not having over-end points. Vertical and horizontal axes of final matrix picture have the same rule.

Now Scan xy
Where x = number of rows
y = number of columns

We have the pictures of an object with the help of a proposed watermark split algorithm. These pictures are stored in the program folder with the same name given by the proposed watermark split algorithm.

This algorithm creates a new empty image frame. With the help of 'for' loop all grayscale images are used. In Matlab, 'imread' function is used to read images.

These empty image frames are composed from the read images & finally, the output image is created.

In Figure 5 we have smaller pieces of split watermark and in Figure 6 we have shown the process of merging the small pieces of split watermark. In the next part it has been shown in the form of a matrix.

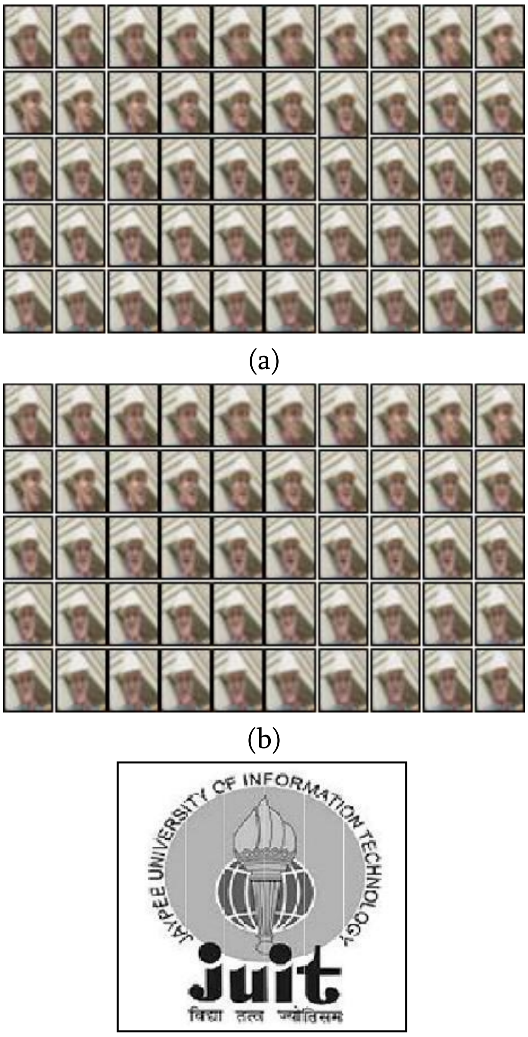
$$\begin{pmatrix} \text{img}[1] & \dots & \text{img}[n] \\ \vdots & \ddots & \vdots \\ \text{img}[n] & \cdots & \text{img}[n] \end{pmatrix} = \begin{pmatrix} [n[k]\, m[l]] & \dots & n[k]m[l] \\ \vdots & \ddots & \vdots \\ n[k]m[l] & \cdots & n[k]m[l] \end{pmatrix}$$

$$= \left[\sum_{i=1}^{n} n_i[k] \sum_{i=1}^{n} m_i[k] \right]$$

In our example we have taken a watermark image that has the dimension of 256 x 256 and we obtain total no of scene changed frames 77 and 97 for Foreman video and Car race video respectively. Now with the help of the proposed watermark split algorithm generate small blocks of digital watermark as follows for Foreman video.

III. EXPERIMENTAL RESULTS

To implement SWEA, original video 'Foreman' at the dimension of 352 x 288 and the size of the original watermark image is 256 x 256 are used. Fig. 3(a) shows original video frames and 3(b) watermarked frames using SWEA, 3(c) extracted watermark pieces and 3(d) extracted watermark. TABLE I shows the results, when the size of inserting a watermark is small, the PSNR of watermarked frame will reach too high.



(a)

(b)

(c)

Figure 4. SWEA algorithm on Foreman Video. (a) Original Frames of Foreman video. (b) Watermarked frames of Foreman video. (c) Extracted watermark.

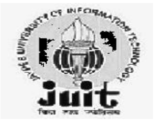
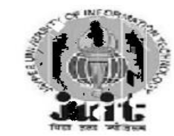
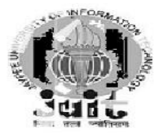
	Gamma correction 0.5	Gamma correction 2	Gamma correction 4
Attacked frame	 PSNR=25.34 dB	 PSNR = 19.80 dB	 PSNR = 13.54 dB
Extracted watermark			

Figure 5. Block Diagram of Video Watermarking

Table 1. Experimental results for car race video.

Size of inserted watermark	**PSNR (dB)**	**MSE**
8 x 8	40.1001	6.3544
16 x 16	39.6682	7.0188
32 x 32	39.4523	7.3765
64 x 64	39.4006	7.4648
128 x 128	39.3987	7.4893
256 x 256	39.3812	7.4982

TABLE 2. EXPERIMENTAL RESULTS FOR WAKNA ROAD VIDEO.

Size of inserted watermark	**PSNR**	**MSE**
8 x 8	42.3011	5.9673
16 x 16	41.8764	6.2643
32 x 32	41.1354	6.9065
64 x 64	39.8731	7.3108
128 x 128	39.4521	7.6874
256 x 256	39.0142	7.8432

IV. CONCLUSION

The proposed approach is more efficient because the watermark has been divided into small pieces and every piece of watermark has a very low space. So it increases the imperceptibility of watermarked frame just because of its small size. The proposed approach also increases privacy and security of the original watermark because it is very hard to detect the pattern of inserted watermark by using splittance technique.

Video watermarking is an essential need of copyright protection and a lot of research is still going on to find out the new methods for security and privacy of the multimedia contents. Current methods for video watermarking are extended form of image watermarking and there is a great scope of innovation. Research can be carried out to establish new strategies for digital video copyright protection which makes the system more robust and efficient.

V. REFERENCES

[1]. I. J. Cox, M. L. Miller, J. A. Bloom, J. Fridrich, and T. Kalker, "Digital Watermarkingand Steganography," in 2nd ed. San Mateo, CA, USA: Morgan Kauffman, 2008.

[2]. X. Niu, S. Sun, "A new wavelet-based digital watermarking for video, "In Proceedings of the IEEE Digital Signal Processing Workshop", pp. 1–6. Texas, 2000.

[3]. Y. Wang, J. F. Doherty, and R. E. Van Dyck, "A wavelet-based watermarking algorithmfor ownership verification of digital images," IEEE Trans. Image Process, vol. 11, no. 2,pp. 77–88, Feb. 2002.

[4]. P. Campisi and A. Neri, "Perceptual video watermarking in the 3D-DWT domain using amultiplicative approach," in Lecture Notes in Computer Science, vol. 3710. New York,NY, USA: Springer-Verlag, pp. 432–443, Sep. 2005.

[5]. E. Ganic, Eskicioglu, "Robust DWT-SVD domain image watermarking: embedding datain all frequencies," In Proceedings of the ACM Multimedia and Security Workshop, pp.166–174, 2004.

[6]. I.B Shahraray "Scene Change Detection and Content-Based Sampling of VideoSequences," in SPIE 2419 (Digital Video Compression: Algorithms and Technologies):2–13, 1995.

[7]. H. Khalilian, S. Ghaemmaghami, and M. Omidyeganeh, "Digital video watermarking in3D ridge let domain," in Proc. 11th ICACT, vol. 3, pp. 1643–1646, Feb. 2009.

[8]. S. Gandhe, U. Potdar, and K. Talele, "Dual watermarking in video using discrete wavelet transform," in Second International Conference on Machine Vision, ICMV ’09, pp. 216 –219, Dec. 2009.

[9]. George, M., J. Chouinard and N. Georganas, "Digital watermarking of images and video using direct sequence spread spectrum techniques," Proceeding of the IEEE Canadian Conference on Electrical and Computer Engineering, May 9-12, IEEE XplorePress, Edmonton, Canada, pp:116-121H. Huang, C. Yang, and W. Hsu, "A video watermarking technique based on pseudo-3D DCT and quantization index modulation," IEEE Trans. Inf. Forensics Security, vol. 5, no. 4, pp. 625–637, Dec. 2010.

[10]. S. Bhattacharya, T. Chattopadhyay, and A. Pal, "A survey on different video watermarking techniques and comparative analysis with reference to h. 264/avc," in IEEE Tenth International Symposium on Consumer Electronics, ISCE’06. IEEE, pp. 1–6, 2006.

[11]. H. Khalilian and I. V. Bajic, "Multiplicative video watermarking with semi- blind maximum likelihood decoding for copyright protection," in Proc. IEEE Pacific Rim Conf. Commun., Comput. Signal Process, pp. 125–130, Aug. 2011.

[12]. B. Yeo and B. Liu, "Rapid scene change detection on compressed video," in IEEE Trans. Circuits Syst. Video Technol., vol. 5, no. 6, pp. 533–544, Dec. 1995.

[13]. F. Hartung and B. Girod, "Watermarking of uncompressed and compressed video," in Signal Process., vol. 66, no. 3, pp. 283–301, May 1998.

[14]. A. Hampapur, R Jain and T. Weymouth, "Digital Video Segmentation," In: Limb J, Blattner M (eds) Proc Second Annual ACM Multimedia Conference and Exposition, ACM, New York, NY, USA, pp 357–364, 1994.

[15]. M. A. Akhaee, S. M. E. Saheaeian, and F. Marvasti, "Contourlet based image watermarking using optimum detector in a noisy environment," IEEE Trans. Image Process., vol. 19, no. 4, pp. 967–980, Apr. 2010.

[16]. B Shahraray "Scene Change Detection and Content-Based Sampling of Video Sequences," in SPIE 2419 (Digital Video Compression: Algorithms and Technologies): 2–13, 1995

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

A New Image Steganography Technique for Hiding the Data in Multi Layers of the PNG Images

Jyothula Dharma Teja*[1], A Chandra Sekhara Rao[1], Suresh Dara[2]

[1]Department of Computer Science and Engineering Indian Institute of Technology (ISM),Dhanbad, Jharkhand, India

[2]Department of Computer Science and Engineering B.V. Raju Institute of Technology, Narsapur, Telangana, India

ABSTRACT

Steganography is a process of hiding the secret data in disguised form to conceal the Presence of secret data and the original form of its existence. In this paper a new Steganography method is proposed which uses multiple layers to hide the data. The method proposed is a PNG image based technique .Unlike the other data hiding Methods in which bmp or gif file format is used this method uses PNG images in which the originality of RGB layer is highly preserved even after Stegonating the Image. In this multi layer approach we Proposed a method for both text to image, and image to image Steganography, In this method The plain text is encoded into the cover image, and resultant image is obtained with plain text embedded in it, and the resultant image is further encoded inside other cover image as second resultant image, with image embedded in it, by this approach the how secret data stored and the form of its existence is highly unpredictable, it is highly Robust from Attacks even a positive attack cannot produce the exact results of secret data as the data is stored in different forms in different layers. so in this proposed method security is highly enhanced and the hiding capacity is highly Improved.

Keywords : Steganography, data hiding, Least Significant Bit (LSB), Multi layer,RGB

Reference: to this paper should be made as follows: J.Dharma Teja, ACS Rao, Suresh Dara (2017) 'A NEW STEGANOGRAPHY TECHNIQUE FOR HIDING THE DATA IN MULTI LAYERS OF THE PNG IMAGES ', Int. J. Ad Hoc and Ubiquitous Computing, Vol. X, No. Y4, pp.000–000.

Biographical notes: Jyothula Dharma Teja is an M.Tech student in Department of Computer science and Engineering from Indian Institute of Technology(ISM),Dhanbad. India. His main research interests include signal processing, image processing, machine learning

A Chandra Sekhara Rao is an Assistant Professor at the Department of Computer Science and Engineering, Indian Institute of Technology(ISM), Dhanbad, India. His main research interests include Evolutionary Algorithms, Machine Learning, and Bioinformatics Suresh Dara is an Assistant Professor in Department of Computer Science and Engineering at B.V.Raju Institute of Technology, Narsapur, Telangana, India. His research interests include Machine learning, Bioinformatics, Evolutionary computation and big data analytics.

I. INTRODUCTION

Security is the main concern in today's modern world, to hide a sensitive piece of data from intruders and hackers became a difficult task. In cryptography we use certain Techniques to encode the data using a key and the encoded secret data is Decoded using the same key or different key shared by the sender to decode. In cryptography one can predict message is encoded but cannot decode without key. But in Steganography one cannot predict the data is encoded or its form of existance.In Steganography we encode the message inside a media and the original form of secret data is changed, so the secret data is hidden and its form is unknown to predict. Steganography can be involved in two categories: 1.Linguistic Steganography, 2.Technical Steganography. Here our interest is the technical Steganography Technical Steganography is further classified as follows 1.Text, 2. Image, 3. Audio, 4.Video, 5. Protocol. The data hiding is possible in the above formats because of the existence of high redundancy bits in the above digital media. Higher the redundancy bits higher the possibility of manipulation .In text Steganography the plain text is hidden in the image or in the video file In image Steganography the image is stored inside other image or in a video file .In audio Steganography the audio file is hidden inside another audio file or other media file. In video Steganography the video is usually hidden inside another video file. or a still image can some image files that are combined to form a video file.In protocol Steganography the network protocols are hidden for secret communication. Though each of them has their own importance The Image Steganography is widely used because of its wide possibilities in manipulating the pixels. Many data hiding methods are proposed recently but this proposed method has some unique features to hide the data. Here in our Article we discuss a method for both Text-Image and Image-Image Steganography. In this case a text message is hidden it in the image concealing the existence of its form as text data so that the intruder cannot guess there is an existence of secret data in the form of text encoded inside the image which further encoded inside other cover image. The above method we are going to propose is an image Steganography method, In this new method we use a multi layer of security.This lets the method robust from Steganalysis which is detection of the existence of Steganography in the given digital media

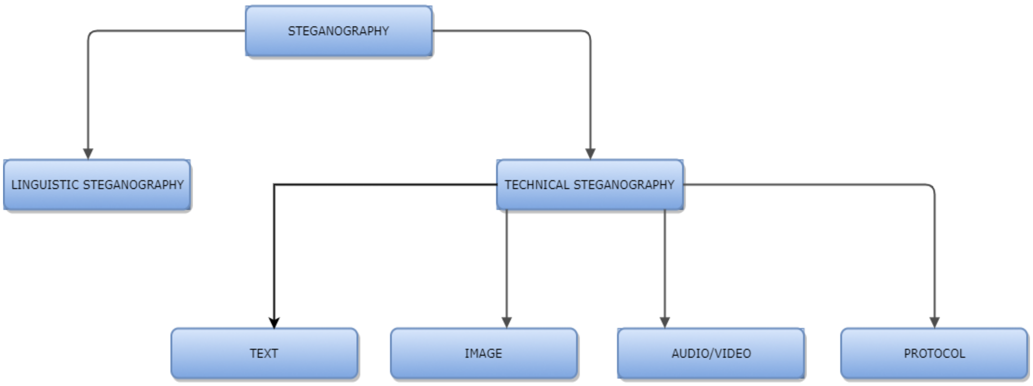


Figure 1. Classification of Steganography

II. BASIC TERMINOLOGY

Cover image: Cover image is the image the data is usually hidden inside

Secret data or payload: Secret data is the data to be hidden, it can be in any form as text or an image or a video file or an audio file.

Steganography algorithm: There are many techniques in Steganography,the algorithm is selected based on the capacity and form of secret data and security issues , LSB(LEAST SIGNIFICANT BIT) is one of the most commonly used technique.

Steganoted image or Stegogramme: Steganoted image is the resultant image obtained after encoding the secret data in the cover image using the Steganography algorithm.

BASIC MODEL OF STEGANOGRAPHY

The basic model describes how the data is embedded and extracted

Model For Embedding The Data

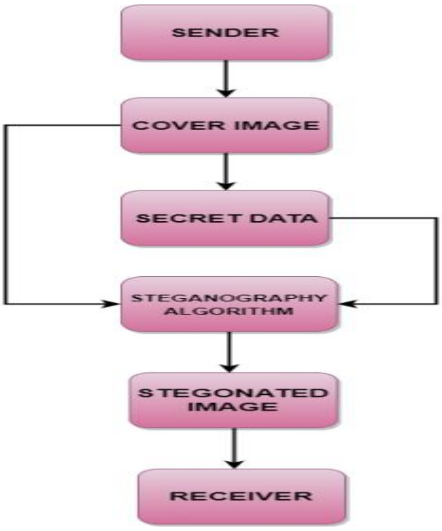


Figure 2. Flowchart describing the Embedding the process

Model For Extraction Of The Embedded Data

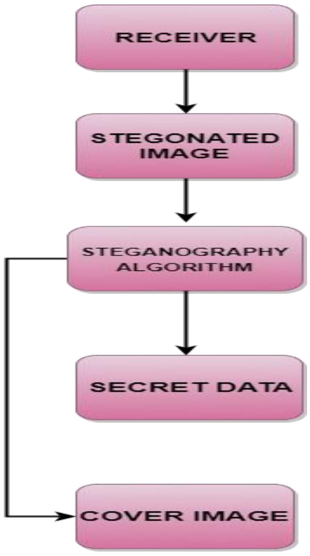


Figure 3. Flowchart describing the Embedding process

III. RELATED WORK

Steganography is applied to both transform domain and spatial domain In transform domain the widely used methods are JSTEG (JPEG Steganography), DWT etc. In spatial domain the LEAST SIGNIFICANT BIT (LSB) method is widely used.

3.1 LSB(Least Significant Bit) METHOD

LSB in bmp, LSB in gif, LSB in PNG is widely used. In Simple LSB technique for RGB color images each pixel color component is divided into 8bit binary strings in the ASCII format and the least significant bit is modified on the selected color to hide the data string in it.

Example: Let's hide a data String using LSB method, **01111010** is the String to hide into an 8-bit color image. The binary equivalent of those pixels may be like this:

01101100 10101111 11011010
01101010 10101101 10111010
10011011 10111001

The binary string **01111010** is replaced in every lsb bit from left to right to in the pixel vales of the image, the replaced bit pattern would be

01101100 10101111 11011011
01101011 10101101 10111010
10011011 10111000

The binary string **01111010** (decimal value 122) is secretly hidden inside the LSB'S of the pixels .

But This Technique is easy to detect and porn to attack, so this new method is proposed to solve this problem.

IV. THE MAIN GOAL OF THE STEGANOGRAPHY TECHNIQUES ARE [1,2]

1. Large data hiding capacity
2. High security

3. Higher PSNR values

V. PROPOSED METHOD

In this method, the secret message to be encoded is converted to a stream of 6 bit binary data per character. This binary data is encoded into the blue color component of the first image. The image (first image) thus formed has the secret message encoded into it. This image is further encoded into another image (second image). Each pixel of the first image requires three pixels of the second image for embedding their values into them. The red value of the pixel of the first image is encoded into the RGB values of the pixel of the second image. Similarly, the green and blue values of the pixel of the first image are embedded into the next two consecutive pixels of the second image. During extraction the embedded image is first extracted. This extracted image has secret text encoded into it. This image is further processed by the decoding algorithm to extract the secret text.

5.1 Encoding algorithm:

1. The secret message is broken down to individual characters.
2. Each character is assigned a distinct 6 bit binary notation example: a=000001
3. Let the total no. of binary digits in the message be “bin_total”
4. The image(first image) in which the secret message is to be encoded is converted in to an array of pixel colors.
5. The blue color of each pixel is manipulated to contain 0 or 1
6. Now start the loop from 1 to bin_total
7. Read the pixel (x , y)’s blue color
8. If message bit is 1 and blue color is even, then subtract 1 from blue; blue=blue-1; else do nothing to blue component
9. If message bit is 0 and blue color is odd, then subtract 1 from blue; blue=blue-1; else do nothing to blue component
10. Increment x value till the end of the image width. once width is reached, reset x to 0 and increment y.
11. When the loop is complete, take a second image in which the first image is to be encoded.
12. Now the pixel colors of the first image is encoded into the pixel colors of the second image.
13. Read a pixel color of the first image and store them in 3 variables R=red, B=blue, G=green(example: R=125,G=167,B=234)
14. Read the pixel color of the first image and pad the color values to the nearest 10thvalue (RGB=234,147,255 to 230,150,240), if a color value is greater than 240 then it is made equal to 240.
15. Now add each decimal place value of the variable R(ex:125=1,2,5) to the padded color values of the pixel of the second image.(ex: RGB=230,150,240 to 230+1, 150+2, 240+5= 231, 152, 245)
16. Now embed the G, B values in to the next 2 pixels of the second image respectively.
17. Repeat this processes until all the pixels of the first image is embedded in to the second image.
18. It takes 3 pixels of the second image to embed 1 pixel of the first image.

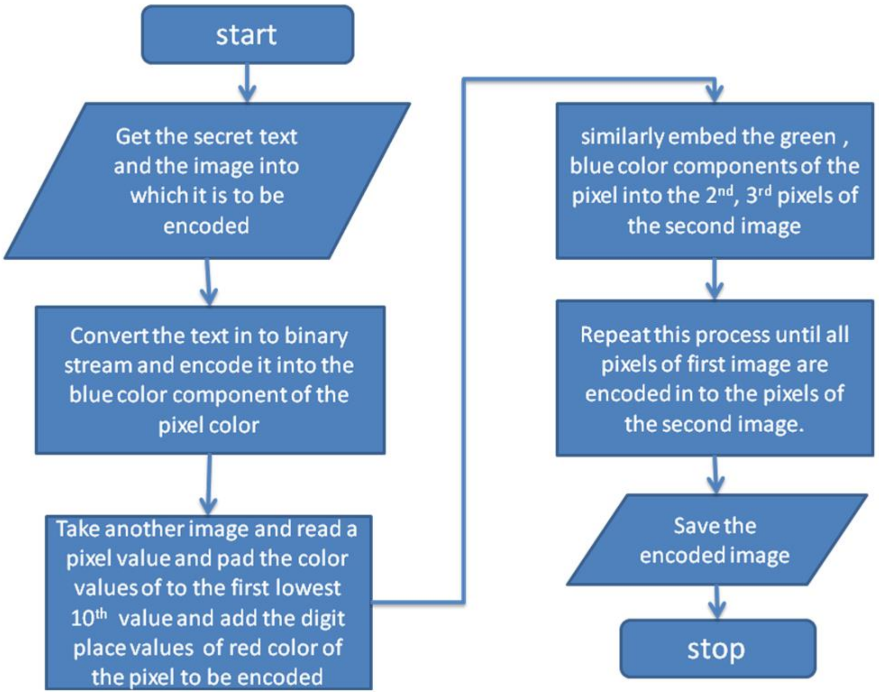

start
Get the secret text and the image into which it is to be encoded
Convert the text in to binary stream and encode it into the blue color component of the pixel color
Take another image and read a pixel value and pad the color values of to the first lowest 10th value and add the digit place values of red color of the pixel to be encoded
similarly embed the green , blue color components of the pixel into the 2nd, 3rd pixels of the second image
Repeat this process until all pixels of first image are encoded in to the pixels of the second image.
Save the encoded image
stop

Figure 4. Flow chart describing the Encoding Algorithm

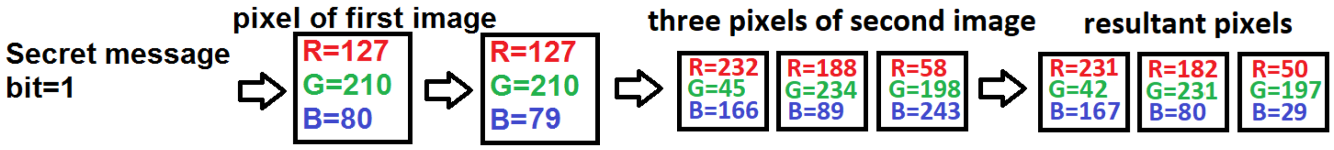


Figure 5. Modification in the Pixels while Encoding

5.2 Decoding algorithm:

1. The pixel colors of the image to be decoded are read. Each pixel of the embedded image is extracted from 3 pixels of the encoded image.
2. The pixel color of the first pixel are read, and the red, green, blue values are subtracted from the next lowest tenth value(ex: RGB=231,152,245 to 230+1, 150+2, 240+5)
3. The resultant values are put into decimal places from red to blue(ex :RGB=230+1, 150+2, 240+5 to 125)
4. This extracted value is red color component of the pixel of the encoded image, similarly processing the next two pixels give the green and blue color values of the encoded pixel.
5. Repeat this loop until all the pixels of the encoded image are extracted.
6. Now save the image that is decoded from the above process for further information extraction.
7. Now the decoded image has a secret message encoded in it.
8. The blue component of the (x,y)pixel of the decoded image are read. If the value is even then the encoded bit is 0.
9. If the value is odd then encoded bit is 1
10. After every six cycles the bits are assembled to form a character(ex: 000001=a)
11. Increment x value. if x value reaches the width of the image, reset x to 0 and increment y value.
12. Repeat the loop until all the characters of the secret message are decoded.

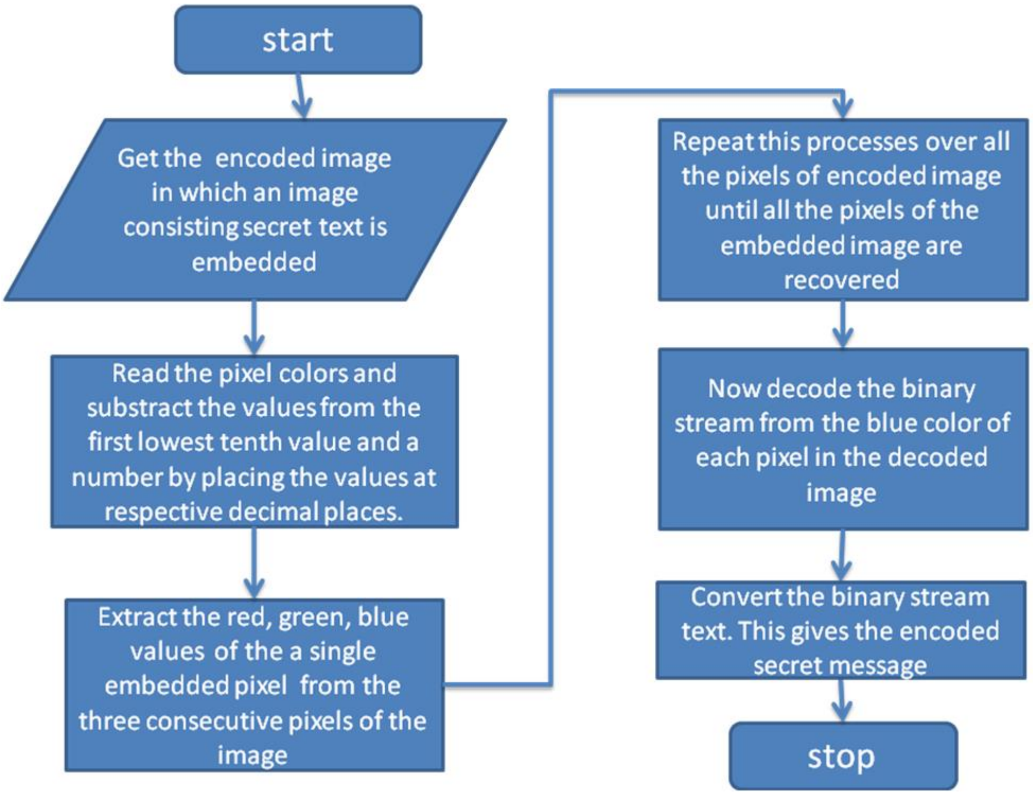

start
Get the encoded image in which an image consisting secret text is embedded
Read the pixel colors and substract the values from the first lowest tenth value and a number by placing the values at respective decimal places.
Extract the red, green, blue values of the a single embedded pixel from the three consecutive pixels of the image
Repeat this processes over all the pixels of encoded image until all the pixels of the embedded image are recovered
Now decode the binary stream from the blue color of each pixel in the decoded image
Convert the binary stream text. This gives the encoded secret message
stop

Figure 6. Flow chart describing the Decoding Algorithm

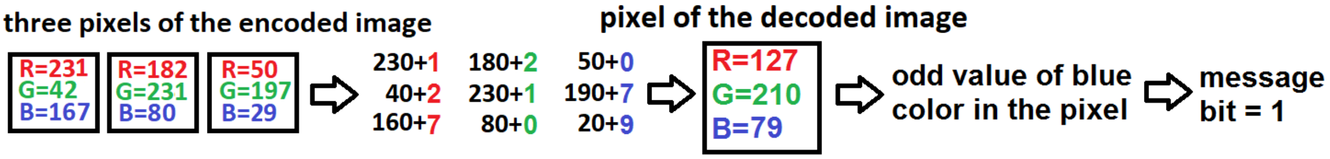


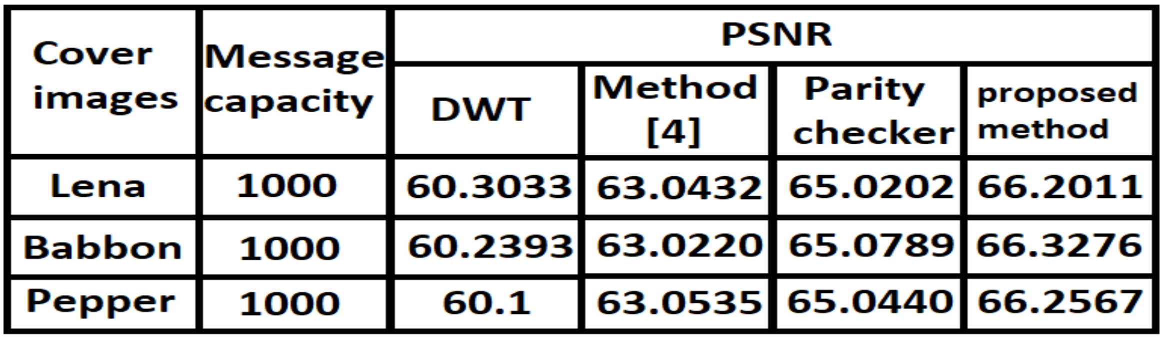
Figure 7. Modification in the Pixels while Decoding

VI. EXPERIMENTAL RESULTS

The results in this format are calculated using the PEAK-TO-SIGNAL-RATIO(PSNR). The PSNR measures the matching of the original image with the Stegonated image by measuring the maximum possible power signal of the original image with the noised image. Here in our case the original image is the cover image and the noised image is the stegonated image. Higher the PSNR better the results archived in this context we tabulated the comparison of PSNR of various methods with the proposed method. The proposed Algorithm was implemented in MATLAB (R2015 a) running on Windows 10 Operating System. The images used are 265x265 standard PNG Format images namely Lena, Baboon, Pepper, Boat and the tested message capacity and the Method names are tabulated to compare

6.1 COMPARISON OF RESULTS

Table 1. PSNR value comparison of 3-Methods and suggested method



Cover images	Message capacity	PSNR			
		DWT	Method [4]	Parity checker	proposed method
Lena	1000	60.3033	63.0432	65.0202	66.2011
Babbon	1000	60.2393	63.0220	65.0789	66.3276
Pepper	1000	60.1	63.0535	65.0440	66.2567

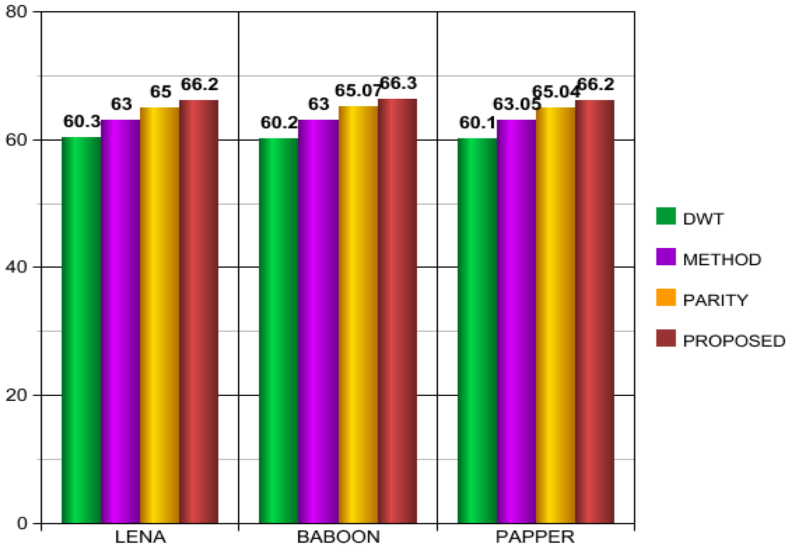
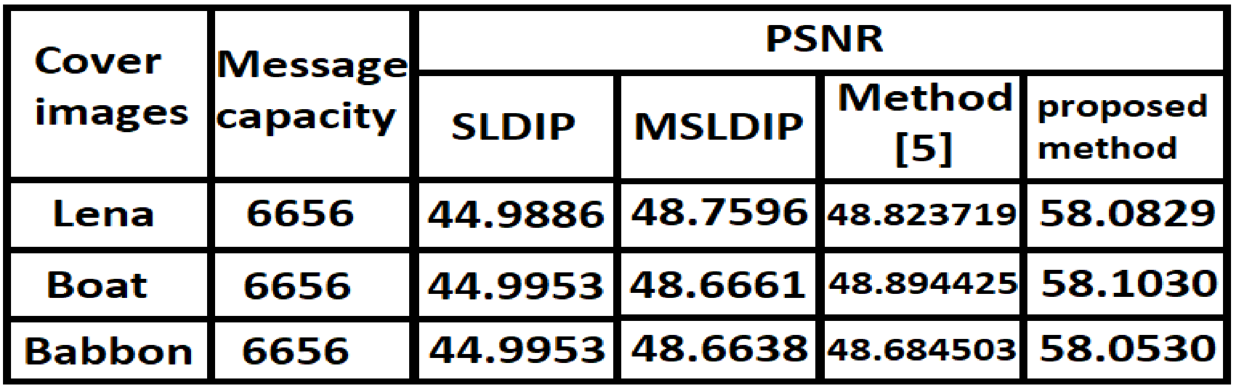


Figure 8. Bar graph of the PSNR Values for the 4 Methods Mentioned

Table 2. PSNR value comparison of 4 different Methods and proposed method



Cover images	Message capacity	PSNR			
		SLDIP	MSLDIP	Method [5]	proposed method
Lena	6656	44.9886	48.7596	48.823719	58.0829
Boat	6656	44.9953	48.6661	48.894425	58.1030
Babbon	6656	44.9953	48.6638	48.684503	58.0530

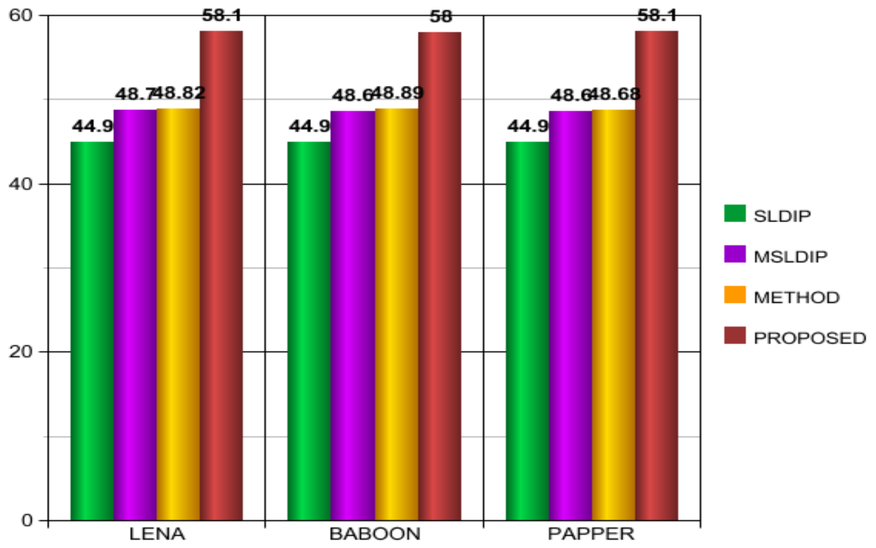


Figure 9. Bar graph of PSNR values for the 4 different methods

Table 3. PSNR value comparison of 3 different Methods and Suggested method

Cover images	Message capacity	PSNR		
		Jpeg-Jsteg	Method [5]	proposed method
Lena	4382	37.77	50.717675	59.8805
Babbon	6026	36.49	49.117879	58.4644
Pepper	4403	37.77	50.763116	59.8795

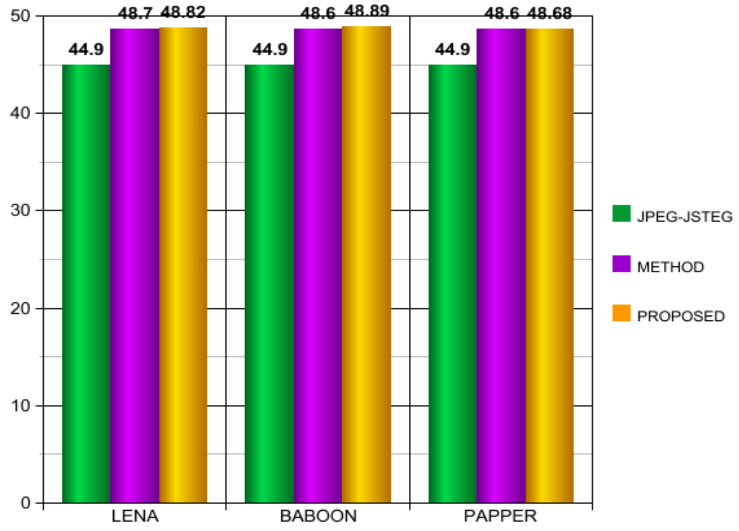


Figure 10. bar graph of PSNR values for the 3 different methods compared

VII. CONCLUSION

In this article a new Steganography method is Proposed for multi layer data hiding, The proposed method has high PSNR Values compared with Few Existing Methods like SLDIP[1](substitute last digit in pixel), MSLDIP[1] (Modified substitute last digit in pixel), JPEG-JSTEG[7], Parity Method[6],DWT[8], the Results are tabulated. Our main aim by this article is to secure the data and to preserve the image quality and To increase the capacity of data hiding. The above proposed method do not destroy the originality of the image ,by using this new multi layer approach data security is highly enhanced, The high PSNR values obtained because RGB layers of the Image are preserved well ,though some methods have similar PSNR values, still this method is highly suggestible as the security is main concern in this field. Further studies suggest the Application of this method for video data hiding and few Security Enhancements.

VIII. REFERENCES

[1]. Radwan, A. A., & Swilem, A. seddik AH,"A high capacity SLDIP (substitute last digit in pixel) method.In fifth international conference on intelligent computing and information systems (ICICIS 2011) (Vol. 30).

[2]. Deepa S., Umarani R., "A Study on Digital Image Steganography ", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 3,Issue 1, January 2013.

[3]. Abdelmgeid A. A., Al - Hussien S. S., " New Text Steganography Technique by using Mixed-Case Font ", International Journal of Computer Applications, Vol 62, No.3, January 2013

[4]. Marwa M. E., Abdelmgeid A. A., Fatma A. O. "A Modified Image Steganography Method based on LSB Technique." International Journal of Computer Applications,Vol. 125, No. 5, September 2015.

[5]. Abdelmgeid A. A., Al – Hussien S. S., " New Image Steganography Method By Matching Secret Message With Pixels Of Cover Image (SMM) ", International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), Vol. 3, Issue 2, Jun 2013.

[6]. Tahir A. and Amit D.” A Novel Approach of LSB Based Steganography Using Parity Checker” International Journal of Advanced Research in Computer Science and Software Engineering, Vol 5, Issue 1, January 2015.

[7]. S. K. Muttoo , Sushil K. "Data Hiding In JPEG Images", BVICAM'S International Journal of Information Technology Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi, Vol. 1, No. 1 January – June, 2009

[8]. Arun R. , Nitin S. , Eep K. "Image steganography method based on kohonen neural network." International Journal of Engineering Research and Applications (IJERA) Vol. 2, Issue 3, May-Jun 2012.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Intrusion Detection in IoT based on Neuro-Fuzzy Approach

Shafalika Vijayal, Mohit Mittal
Department of Computer Science Engineering MIET, Jammu, India

ABSTRACT

The internet of things (IoTs) is part of latest developments having combination of RFID, sensor nodes, communication technologies and protocols. IoT is one of the latest technology that has amass significant research recognition due to their ability to monitor the physical world phenomenon and their applicability to an extensive range of applications. IoTs have a wide range of applications including smart cities, smart homes, industrial sectors etc. The current scenario is highly demanding for deployment of smart sensors into existing applications to deliver a fully automated system. The major issue faced by IoT's existing system is security issue. This paper focuses on intrusion detection in IoT using neuro-fuzzy approach. The proposed model discusses about how anomalies detection scheme is improved using neuro-fuzzy approach.

Keywords : Internet of Things (IoT); SOM neural network; Neuro-Fuzzy technique; intrusion detection; anomaly detection.

I. INTRODUCTION

With the advent of IoT, the technology is expanding its purview not just in terms physical hardware but also software and middleware. The internet has played a vital role in providing the connections. IoT enables the physical devices like vehicles, buildings, electronic devices, sensors, actuators to communicate (hear, see, think, perform) and coordinate decisions through technology and data flow. IoT [9] transforms the simple objects to smart objects. With communication the prime focus between devices the flow of data securely in the main concern. The cognitive functions of humans have changed the way machines should perform.

Almost anything can be connected to internet: cars, watches, spectacles, meters at home, and manufacturing machines. But the pitfalls prevail and covering them through the best known foundations of world class security using hardware and software level protection is the concern [10].As connectivity of devices is increasing so is the threat to malware, hacking and other types of attacks with smart gadgets like TV, media PC's, fridge's. A fridge was reportedly involved in sending spam emails as web attack compromised smart gadgets in the year 2014.About 25% of the messages did not pass through the laptops, desktops or smart-phones. Instead, the malware managed to get itself installed on other smart devices such as kitchen appliances, home media systems on which people stored copied DVDs and web connected televisions. Many of these devices have computer processors onboard and act as self contained web server to handle communication and other sophisticated functions.

With artificial intelligence giving this feature, Internet of things is the internetworking of these features by integrating physical devices (wireless SOC, Prototyping boards and platforms) and making them communicate (RFID, NFC, ANT, BLUETOOTH , ZIGBEE, Z-WAVE, IEEE 802.15.4, WIFI). The terms coined as smart home, smart car, smart healthcare, smart city fall under the domain of IoT.

Security and privacy play an important role in markets globally due to the sensitivity of consumer privacy because of lack of common standards and protocols. The core functionality of IOT depends on the exchange of information between trillions of internet connected devices. Thus security is a very crucial aspect to be covered. With security the data/Information exchange can be classified as either -The best. In the next section, intrusion detection systems will be discussed in detail. In section III, see the latest review on intrusion detection IoT. Based on this latest problem, section IV focuses on the Intrusion detection based on neural networks. Than after proposed methodology is explained. In last section, conclusion part provides the analysis of the technique proposed.

II. INTRUSION DETECTION SYSTEMS

Intrusion detection systems [11], [12] are employed in systems or networks to detect any kind of malevolent activity intended to fetch information or harm the systems distributed over network. The IDS are placed at strategic points on the network to monitor the traffic from various devices (computers, laptops, workstations) if the communication flow gets unusual it is reported to the network administrator. The intruders intend to harm the network by penetrating into the security system as legitimate users.

A. Types of intruders

The intruders can be of following types-

1) Masquerader
2) Misfeasor
3) Clandestine users.

IDS can be any software system or a hardware tool capable to find an intruder doing malicious activity by relating to user activity or already existing signatures of attacks on machines or network packets. IDS is not just detecting the intruders but also provides measures to prevent them. An alarm system is used to inform users about the origin of the failure. The alarms can be implemented as filtered or non-filtered. The later listing the origin of failure along with the devices that were affected by the attack.IDS hence acts as a network observer which informs about the attack by generating an alert before the system or network gets affected.

B. Types of attacks that IDS can detect

1. Internal attacks - are the ones that are generated by nodes within the network.
2. External attacks - are the ones that are initiated by third party nodes which do not exist within the network.

IDS can detect the attacks by monitoring, analyzing, detecting and then raising an alarm.

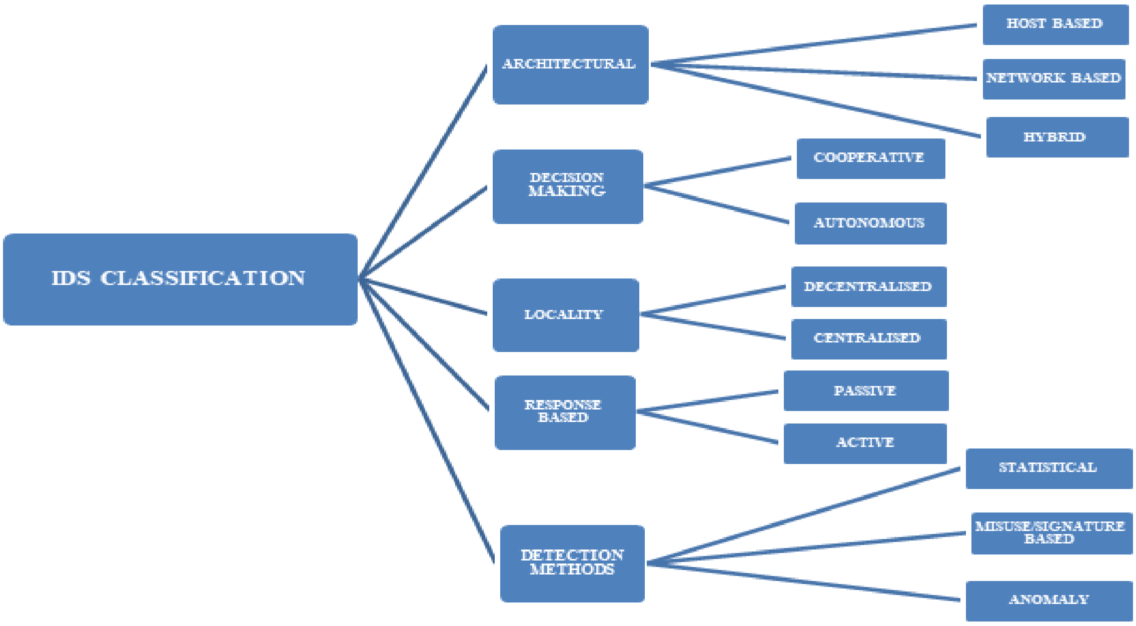


Figure 1. IDS Classification

C. Variants of IDS by detection approach

1) Host based IDS

In these Intrusion Detection Systems hosts are evaluated. The hosts can be a single device or multiple devices on the network. The system evaluates the in and out flow from the device and generates an alert if there is any malicious activity suspected [13]. The current updated system files are evaluated with the snapshots of the previous ones to check for any abnormal behaviour.

2) Network based IDS

In these Intrusion Detection Systems network is analysed to detect any intrusion. Sensors are implemented to keep a check of packets travelling in or from the network. The sensors are placed at various points over the network.

3) Vulnerability assessment IDS

Through these IDS vulnerability of the hosts on internal network or firewalls is checked.

D. Detection Techniques used in IDS

1) Signature based IDS

Also known as rule based detection technique. In this a database of signatures is already created. In this approach signatures refer to the attacks that have already been occurred are assigned a pattern. The database is checked for any signature after analyzing the packets flowing in the network and if any pattern gets matched to the one in database is blocked by raising an alert. This technique is very simple to use but requires a lot of space as the patterns or signatures keep on adding with increase in malicious activity [15]. The disadvantages of this technique are that it cannot identify previously unknown or new attacks. And also requires knowledge to form patterns or signatures.

2) Anomaly based IDS

Also known as event based IDS [12], [13]. In these IDS malicious activities are identified by evaluating the events. This technique is useful in detecting unknown attacks. The behavior of the network is analyzed if there occurs any unusual activity; it is reported as an intrusion. The behavior of the network is analyzed by studying the protocols through which communication is established.

3) Specification Based IDS

This technique is similar to anomaly based detection. In this technique, the normal behavior of the network is defined manually, so incorrect positive rate is less. This technique attempts to combine best of signature-based and anomaly based detection approaches by trying to clarify deviations from normal behavioral patterns that are created neither by the training data nor by the machine learning method. The technique is time consuming as development of attack or protocol specification is done manually, providing a disadvantage of this approach[14].

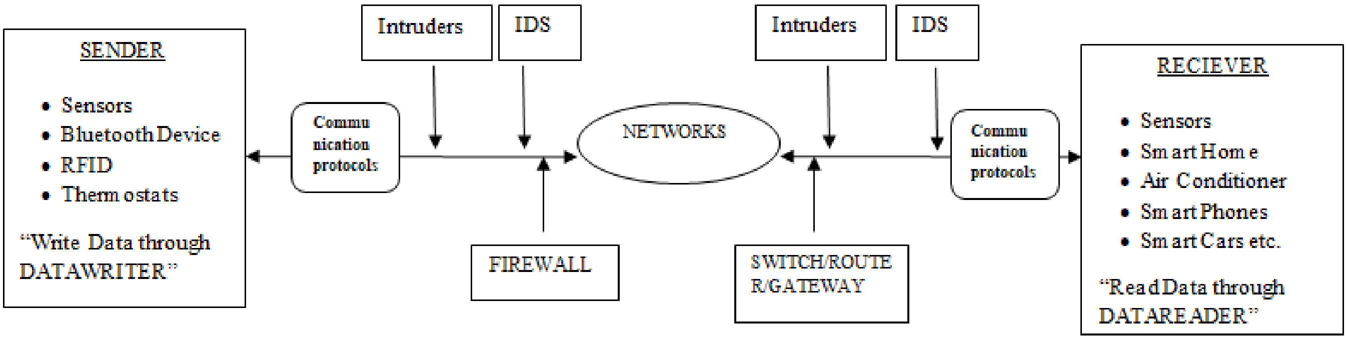


Figure 2. Intrusion detection in IoT.

III. RELATED WORK

Intrusions in network can be detected on statistical level as well as rule based level. The former analyses the intrusion by monitoring the behavior of the user over a period of time dependant on the threshold and user profile. The later interprets the anomalies and external penetrations into the network which lead to abnormal behavior of network.

Elike Hodo et. al. [8] have proposed an offline IDS Model as an ANN to gather and detect the various Information from various parts of the IoT network. The model detects the normal and threat patterns by setting the input nodes in three layers feed forward network. The network is trained by taking repeated steps of gradient decent.[8] use a 5 node sensors IoT network of which 4 act as client, 1 x as server relay node for data analysis. The intruder in the network is considered to b external targeting the relay node to disrupt traffic as DOS attack with one node and DDoS attack with three network nodes. The network is trained with 2313 samples, validates with 496 samples and 496 test samples to construct a ANN confusion matrix which yields 99 % result accuracy to detect DOS and DDOS attacks on legitimate IoT network. Anomaly IDS has prime focus on defining what is normal, so this model gives a good performance in terms of true and false positives rate of traces of network packets.

Kumar et. al. [7] have proposed an algorithm that learns characteristics of both normal and intrusive packets. They consider that the number of intrusive packets to be less in accordance to normal packets in the network .IDS here is explained through DARPA dataset. The K means clustering samples the datasets into normal and abnormal by initially setting value of K equal to two. The distribution implies to a fuzzy rule implemented as sql queries which places the two separate clusters in a vector by identifying the abnormal packets through certain fields like- type, count, land and svr_rate and listing them into a table. The percentage of the combinations of these characteristics defines the extent of intrusion. The impact of these rules assigns a weight to the packet and converting it to a training pattern for Neural Network Technique. Authors later use Back

Propagation to exploit neural network as it uses weights generated to learn the intrusions and differentiate it from normal packets. The paper focuses on dual behavior of network and reduces the number of false alarm rates significantly

IV. INTRUSION DETECTION BASED ON ARTIFICIAL NEURAL NETWORK

Artificial neural network (ANN) [17] is mostly implemented to solve the complex problem (mostly related real world scenarios). It comprehensively embedded into system and helps to resolve the intrusion detection problems encountered by the existing system. Under intrusion detection, statistical analysis incorporates statistical comparison among existing events to set off baseline criteria in advance. It is commonly involves in the detection of deflections from typical behavior and diagnosis of similar events to those which are indicative of an attack [3].

Authors in paper [1] and [2] have been discussed an alternative system to the statistical analysis component of anomaly detection system which is based on ANN. Nowadays, the field of IoT implementation is drastically expanded on result of it the implementation of ANN for intrusion detection is lacks behind in each IoT scenario.

ANN techniques are generally categorized into two learning algorithms: supervised learning and unsupervised learning. In supervised learning the input as well as target values are provided. It means that the target values are present according to which input values are optimized. In simple words, the teacher is present on which weight values can be optimized. On the other hand in unsupervised learning, only input values are provided for optimization [17], [19]. The weight values are updated according to input values. In another words, no teacher exists for optimization.

A. Supervised learning

Supervised leaning [17] is used for adaptation. Multi-Level Perceptron (MLP) is most common ANN which is generally used for pattern recognition problems. Multilayered feed-forward neural networks are supervised approach for non-parametric regression methods. It has main functionality in dataset by minimizing the loss function. Loss function is used for training process for ANN as quadratic error function.

In supervised neural network the input is induced in the network. The training process is starts after that. The product of input values and default weight values are calculated. This resultant values are input into transfer function. After this threshold values are either inhibit or exhibit. On completion of learning process, the final values are represented in the form of neural network weights.

J. Cannady et. al in paper [4] have discussed about how to apply MLP model for misuse detection. In the proposed method, MLP prototype had various characteristics such as 4 fully connected layers, 9 input nodes a nd 2 output nodes (normal and attack). The simulation of this model under normal traffic evaluates several attacks as ISS scans, SATAN scans and SYNFlood.

B. Unsupervised learning

Kohonen's Self-Organizing Maps (SOMs) [17], [18] are come under the category of neural network family. Professor Tuevo Kohonen has invented SOM neural network in 1982. 'Self-Organizing' name suggests that no supervision is present. 'Maps' word designates that attempt to map their weights to the given input values. The neurons in different layers are arranged according to topological function like gridtop, hextop or randtop. Distances among the neurons can be calculated with help of different distance functions such as dist, boxdist, linkdist and mandist.

SOM network identifies a winning neuron i*. Except the winner neuron set aside, all other neurons will update within a certain neighbourhood. Ni* (d) of the winning neuron are updated, using the Kohonen rule:

$$iw\ (q) = (1- \alpha)\ iw(q-1) + \alpha\ p(q)$$

Here the neighborhood Ni* (d) contains the indices for all of the neurons that lie within a radius d of the winning neuron i*.

$$Ni(d) = \{\ j\ ,di,j \leq d\}$$

When a vector p is presented, the weights of the winning neuron and its close neighbors move toward p. As a result of it, neighboring neurons have learned vectors similar to each other.

The authors in paper [5] and [6] have implemented SOM technique for intrusion detection. SOM approach made clusters of network traffic and determine attacks. It also provides 2D-space visualization of clustered network traffic. Intrusions are then taken out from this view, by highlighting divergence from the norm with visual metaphors of network traffic. The whole approach is tested for various attacks: IP spoofing, FTP password guessing, network scanning and network hopping; log file systems are analyzed from firewalls. Moreover, this approach requires a visual examination of network traffic by an administrator to detect attacks.

V. PROPOSED METHODOLOGY

This section will modulate the solution of the problem related to anomaly detection using neuro-fuzzy approach. The input can be DARPA datasets which is easily available online. The dataset is induced to SOM neural network. As SOM is an unsupervised; the training algorithm processes the input and categorized the output depending on the input given. The weight values are adjusted according to the input values. Here, two categories are maintained: normal and abnormal values. As compared to other clustering techniques like K-means, SOM is far better than K-means clustering. 'k' number of centroids are selected to optimize the solution. SOM is totally dependent on the input and has strong learning algorithm. In general practice, the numbers of intrusion packets are less in number as compare to normal values.

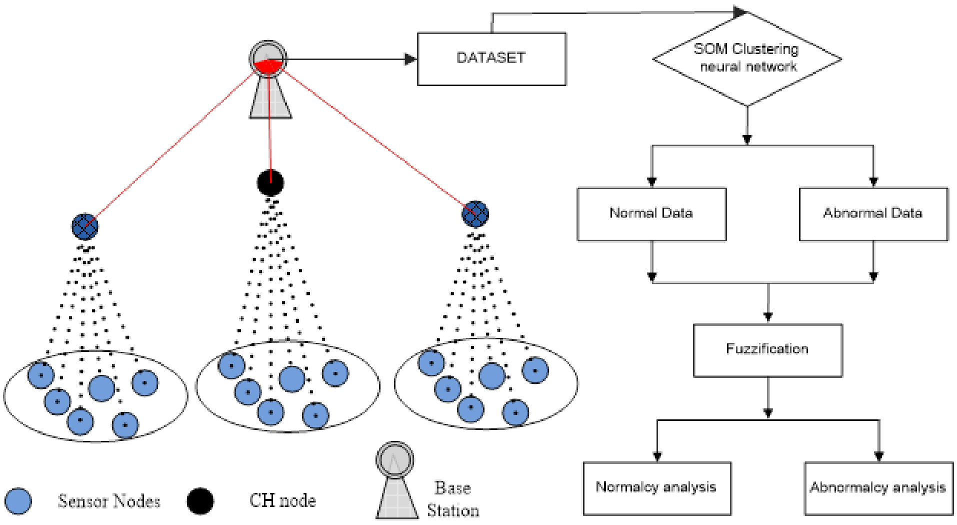


Figure 3. Intrusion detection based on neuro-fuzzy approach

This SOM approach helps in division of dataset into normal and abnormal cluster. Analysis is done over the intrusive data values to get knowledge about major characteristics of abnormalcy. Then only whole picture will come out to be clear cut the actual factors of intrusion. This same analysis is also done over the normal data packets, so that we can get the exact distinguished factors of abnormalcy. This process helps to get better solution in anomaly detection. So, these factors are also induced with these partitioned the data values into fuzzy logic model. Here, mamdani model is used to get exact nature of abnormalcy. Fuzzy rules formulate the dataset into separate vector. Therefore, once the fuzzy logic collects the data packets than only it can able to classify normal packets from the abnormal one or deviated one.

The process of detection initiates thorough analysis of normalcy and abnormalcy in the data packets. The various parametric values will be calculated out of the abnornmal packets with their corresponding values. With the help SQL query processing, scrutinized the distinct values that have been interpreted from the each parameter. As a resultant of this, a list has been generated containing the details of abnormal data values and its characteristics and normal data values with its characteristics as shown in fig. 3. By processing this repeatedly over the dataset, we can able to get much efficient anomaly detection system as due to usage of neuro-fuzzy approach.

VI. CONCLUSION

IoT is named to the collection huge volume of devices into one system connected via radio signals. The major issue has been seen from past years from the existing IoT system is security. To overcome this problem various artificial intelligence techniques or machine learning algorithms are used nowadays. As complexity of the system is so high; to evaluate presence of intrusion requires complex computational algorithms to solve the problem in efficient way. To cope with various scenarios we have proposed hybrid approach for anomaly detection. A neuro-fuzzy approach is one of the best to modulate, evaluate the problem. As per the requirements, we have proposed SOM neural network for categorization of dataset into normal data and abnormal data. For further better optimization of the results, we have proposed the mamdani model that specifies the fuzzy rule set on the normal data and abnormal data for scrutinized further based on membership values. The proposed methodology has use hybrid approach; it will provide better targeted results.

VII. REFERENCES

[1]. Denning, Dorothy, "An Intrusion-Detection Model", IEEE Transactions on Software Engineering, vol.13, no.2, 1987.

[2]. Fox, Kevin L., Henning, Rhonda R., and Reed, Jonathan H. "A Neural Network. Approach Towards Intrusion Detection", 13th National Computer Security Conference, 1990.

[3]. Helman, P. and Liepins, G., "Statistical foundations of audit trail analysis for the detection of computer misuse", IEEE Trans. on Software Engineering, 1993.

[4]. Cannady J. and Mahaffey J, "The application of Artificial Neural Networks to Misuse detection: initial results", Georgia Tech Research Institute, 1998.

[5]. Girardin L. and Brodbeck D. "A Visual Approach for Monitoring Logs", 12th System Administration Conference (LISA '98)", pages 299-308, 1998.

[6]. Girardin L., "An eye on network intruder-administrator shootouts - UBS UBILAB", 1st Workshop on Intrusion Detection and Network Monitoring (ID '99)", 1999.

[7]. K. S. Anil Kumar and V. Nandamohan, "Novel Anomaly Intrusion Detection using Neuro-Fuzzy Inference System", IJCSNS, pp.6-11, 2008.

[8]. Elike Hodo, Xavier Bellekens, Andrew Hamilton, Pierre-Louis Dubouilh, "Threat analysis of IoT networks Using Artificial Neural Network Intrusion Detection System",

International Symposium on Networks, Computers and Communications (ISNCC), 2016.

[9]. Ala Al-Fuqaha, Mohsen Guizani, Mehdi Mohammadi, Mohammed Aledhari and Moussa Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols and Applications",IEEE Communications Surveys & Tutorials, 2015.

[10]. Luigi Atzori, Antonio Iera and Giacomo Morabito, "The Internet of Things: A survey", Computer Networks, vol. 54, pp. 2787-2805, 2010.

[11]. V. Jyothsna and V. V. Rama Prasad, "A Review of Anomly Based Intrusion Detection System", IJCA, Vol. 28, no. 7, pp. 26-35, 2011.

[12]. Nilesh B. Nanda and Ajay Parikh, "Classification and Technical Analysis of Network Intrusion Detection System", International Journal of Advanced Research in Computer Science, vol. 8, pp. 657-661, 2017.

[13]. E. Kesavulu Reddy, "Neural Networks for Intrusion Detection and Its Applications", World Congress in Engineering, vol. 2, 2013.

[14]. Tariqahmad Sherasiya and Hardik Upadhyay, "Intrusion Detection System for Internet of Things", IJARIIE, vol. 2, issue.3, pp. 2244-2248, 2016.

[15]. Pavan Pongle and Gurunath Chavan, "Real Time Intrusion and Wormhole Attack Detection in Internet of Things", IJCA, vol.121, no.9, pp. 1-9, 2015.

[16]. Rupinder Singhm Jatinder Singh and Ravinder Singh, "Fuzzy Based Advanced Hybrid Intrusion Detection System to Detect Malicious Nodes in Wireless Sensor Networks", Wireless Communication and Mobile Computing, pp. 1-14, 2017.

[17]. Laurene Fausett, Fundamentals of Neural networks: Architecture, Algorithm and Applications, Pearson Education 1994.

[18]. Mohit Mittal and Krishan Kumar, "Data Clustering In Wireless Sensor Network Implemented On Self Organization Feature Map (SOFM) Neural Network" ICCCA, 2016.

[19]. Mittal M., Kumar K., Network Lifetime Enhancement of Homogeneous Sensor Network Using ART1 Neural Network, Sixth International Conference on Computational Intelligence and Communication Networks, pp. 472-475 2014.

[20]. Mittal M., Kumar K., Quality of Services Provisioning in Wireless Sensor Networks using Artificial Neural Network: A Survey, International Journal of Computer Application (IJCA), pp. 28-40 2015.

[21]. Mittal M., Bhadoria R. S., Aspect of ESB with Wireless Sensor Network, Exploring Enterprise Service Bus in the Service-Oriented Architecture Paradigm", igi-global publications, pp. 319, 2017.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Comparison of Intrusion Detection Techniques in Cloud Computing

Aditya Bakshi[1], Sunanda[2]

[1]Department of Computer Science and Engineering, School of Computer Science and Engineering Shri Mata Vaishno Devi University , Lovely Professional University, Katra, India1, Phagwara, India

[2]Department of Computer Science and Engineering Shri Mata Vaishno Devi University Katra, India

ABSTRACT

This paper has focused on specifying different Intrusion detection techniques in cloud computing. There are different types of attacks that are affecting the cloud are also discussed in this paper. The role of firewall and different intrusion detection techniques in cloud computing for preventing various attacks has also been discussed.

Keywords : Cloud Computing, Firewall, Intrusion Detection System.

I. INTRODUCTION

Cloud computing is the latest computing technology which provides various services on demand and pay per use basis. Fundamental idea behind the evolution of this technology is diversity of computing relative to users. Every user has own needs and expectations from computers and to fulfil them there is a need of various features from computing components i.e. software, hardware and network. It is almost impossible to have every possible computing environment by every user. Especially in case of software development where technology changes every day and clients have varied requirements, software development organizations cannot purchase every development environment for clients. These conditions lead to the evolution of cloud computing where every computing is provided in virtual environment. There are cloud servers created and maintained by computing giants firms which provide numerous services asked by users. Basically cloud services are categorized in three broad categories. First one is Software as a Service (SaaS) which provides various applications especially bounded to software to users. Second one is Infrastructure as a Service (IaaS) which provides various infrastructure environments to users and last one is Platform as a Service (PaaS) which deals with various platforms like OSs, etc [4] [5]. All these services are available to users on pay per use and on demand basis which reduces the cost from earlier stage which was at unaffordable stage to minimal level. Users have to just pay the rent for the time to which they are using services. Apart from this unique feature, cloud computing provides various other features like availability, maintainability, scalability, interoperability, etc. These all facilities cannot be achieved anyhow by standalone users in their local infrastructure due to various unavoidable conditions whereas cloud providers support them due to devoted services.

II. COMMON ATTACKS IN CLOUD

It is crystal clear that cloud computing is the next generation technology which is suitable for all users ranging from any background and having different local computing resources [1]. Although it has attracted researchers and organizations towards its advancements but still it is in its infancy. Moreover there are various security issues due its openness because cloud architecture involves network as Internet and intranets (in some cases). Some of the major intrusions are described as follows

Insider Attack: This is the attack which is performed by insider cloud users. Those users may try to breach the security of cloud by gaining unprivileged access by using their credentials. This is one of the most disastrous threat to cloud because once the internal security architecture will be breached then overall system can be compromised easily.

Flooding attack: This attack is performed by using Zombies which are innocent host and are compromised by attacker to flood cloud environment by various type of request. Those requests may combine ICMP, TCP, UDP, etc. which are sent to just flood the system and in meanwhile various other targets may be compromised to gain access to resources[2].

User to root access: In this attack, those users are compromised which are having root access to the cloud system. Those users can perform administrator level works due to having root level permissions and compromising their credentials may lead to gain of overall system to the attacker [3]. However it is not a single attack based on any paradigm which will be applied and user will be compromised but it involves various other techniques like social engineering as well as eavesdropping, etc. The main motto behind this attack is to gain credentials to reach to the root level of the cloud server which can be further compromised by using the same.

Port Scanning: Port scanning is the technique to scan for all ports of any system. Although it is a manual process to check for each and every port for their status as open or close but there are various automated tools which provide detailed description about any system based on the provided IP address. These tools are sometimes used as a tool to attack cloud environment. Once all open ports which are not being used by any specific service can be used as a back door and automated programs may be deployed to transmit all inform via the same.

Attacks on Virtual Machine (VM) or Hypervisor: Cloud environment is completely based on virtual architecture. It virtualizes both the environments either internal or outer structure. Virtual machine is a dedicated machine based virtually on real environment and may be used to hold other services which may need sophisticated system. The most popular technique for clubbing and splitting VMs is based on hypervisor. There are various known attacks which try to compromise either VMs or target hypervisor to completely choke the system. These attacks always target the layer which works between two layers and compromising any one of the layers would result in the overall compromise of the system. Backdoor or channel attack: Attacker can perform DDoS attack by compromising Zombie system. It may lead to get access to the cloud environment as a backdoor entry which can be used to perform various malicious activities. However in case of malicious activities performed by compromising authorized system is very difficult to detect due to openness and accessibility[6].

Apart from the above discussed attacks there are various other attacks which lead to severe security problems. The common solution to the problem is firewall implementation. However it does not solve the problems at all which forces the intrusion detection system (IDS) or sometimes intrusion detection and prevention system (IDPS) implementation. First of all we see the features of firewall and various other firewalls which can be implemented and then after various other IDPSs and their comparison in cloud environment [7].

III. FIREWALL

Firewall comprises various set of rules which act as the first line defence mechanism involved in the system. It protects and filters all the incoming and outgoing requests from the system. However, it is completely static in nature working on the pre-defined rules of network. It is unable to protect the system in cases where requests are evasion in nature and here IDPSs play crucial role for the system [8] [9] [10]. Some of the major firewall techniques that are used in cloud environment are Static Packet Filtering Firewall, Stateful Packet Filtering Firewall, Stateful Inspection Firewall and Proxy Firewalls.

Firewalls restrict to some extent in security attacks but not as an overall solution. For sustaining more security in different types of attacks, IDS or IPS can be served as solution that could be incorporate in cloud. However, the different parameters and techniques are required for improving the efficacy of an IDS/IPS in cloud computing. The parameters comprises of different techniques used in IDS and its configuration within the network. Some traditional IDS/IPS techniques such signature based detection, anomaly detection, state protocol analysis etc. can also be incorporated in cloud. The next section covers the common IDS/IPS techniques.

IV. CLOUD IDS TECHNIQUES

A) Signature based detection

This technique incorporates signatures of various known attacks. These signatures are stored in database server of IDS and any incoming or outgoing requests are matched with them. Any matching signature request is discarded immediately from the network or other consequences may be applied like changing the contents, modifying the target, etc. However it is the best technique for known attacks but proves to be very ineffective in case of unknown attacks. Any attack or security breach which is attempted by modifying the content is unable to be detected by this technique. One of the key reason for using signature based detection is because its rules can be easily reconfigured. Reconfiguration of rules is required for updating the signatures of unknown attacks. These signatures are helpful for detecting the network traffic [11].

In cloud, the known attack can be easily detected by using signature based intrusion detection technique. The signature based technique is applied on the front end of cloud for detecting the external intrusion or at back end of cloud for detecting internal intrusions. If signatures are not updated, it cannot be used to detect unknown attacks in cloud.

B) Anomaly detection

Anomaly detection technique tries to detect intrusions that are anomalous to the actual definition. This technique involves various profiles that are used to filter the traffic as genuine or malicious activity. All such profiles are stored in advance as well as dynamically updated based on the uses and traffic pattern. Some of the known products based on this technique are working very well in real life scenarios [12]. Apart from the normal computing, it is also very useful in case of cloud computing. It involves data collection related to the behaviour of legitimate users over training period, and then applies various test which are statistical in nature, are used to observe behaviour and determines genuine user. It is very useful in cases of unknown attacks where definitions or any specific signatures are unknown in advance. The main idea behind use of this detection technique is to decrease the false alarm rate and work either perfectly either with known or unknown attacks [13].

Anomaly detection techniques detects unknown and known attacks which are segregated at different levels. In cloud, by using anomaly based detection, large number of events (network level or system level) occurs, which makes difficult to monitor or control intrusions.[1].

Capability of soft computing to deal with uncertain and data that is partially true, makes them very useful technique in intrusion detection. There are various techniques from this computing like Fuzzy Logic, Association rule mining, Artificial Neural Network (ANN), Genetic Algorithm (GA), Support Vector Machine (SVM), etc. that can be incorporated to improve the accuracy of detection and efficiency of anomaly detection based IDS and signature based IDS.

C) Artificial Neural Network (ANN) [1] based IDS

ANNs generalises data from incomplete data for intrusion detection and classifies also as normal or intrusive behaviour. Types of ANN used in IDS are as: Back Propagation (BP), Multi-Layer-Feed-Forward (MLFF) nets and Multi-Layer-Perceptron (MLP). Distributed Time Delay Neural Network (DTDNN) has been claimed as the best detection technique in this category till now. It contains capability of classifying and fast conversion rates of data and proves to be a very simple and efficient solution. Its accuracy can be improved by combining various other techniques related to soft computing.

ANN based solutions of IDS proves a better solution over other techniques for network data which are unstructured in nature. Accuracy of intrusion detection involved with these techniques is completely dependent on training profile and layers that are hidden.

D) Fuzzy logic based IDS

FIDS are used for detecting and inspecting various network traffic related to SYN and UDP floods, Ping of Death, E-mail Bomb, FTP/Telnet password guessing and port scanning. Some evolving techniques under Fuzzy Neural Network (FuNN) collaborates both type of learning as supervised and unsupervised learning [1]. EuFNN has better accuracy in intrusion detection than normal ANN techniques and experimental results shown in [1] prove accuracy. Real time intrusions can be also detected in real time environment by involving association rules of Fuzzy System. The experimental results generate two result sets that are mined online from training data. It is very suitable for DoS or DDoS attacks that are implemented on large scale.

E) Association rule based IDS

There are various intrusions that are formed based on known or variants of known attacks. Apriori algorithm for determining the signatures of such attacks are used and they are also capable to determine the variants of such attacks can be determined and detected by frequent itemsets. Data mining technique used in Network based intrusion detection with signature based algorithm generates signatures for misuse detection. However, drawback of the proposed algorithm is involved time consumption which is more than considerable for generating signatures. Scanning reduction algorithm solved this problem which reduces the number of database scans for effectively generating signatures from previously known attacks. However, there are very high false positive rates occur which generate due to unwanted and unknown patterns [1].

F) Support Vector Machine based IDS

SVM is better than other artificial intelligence techniques used with IDS. There are various available experiments which show its efficiency over other techniques. It uses limited sample data to detect intrusions where accuracy does not get affected due to dimensions of data. False positives rate is also very less than other techniques as experimented in [6]. This is because that various other techniques require large sample dataset whereas it works on a limited sample dataset. Basically SVM works on binary data so for better accuracy, it can be combined with other techniques which can improve its accuracy in detection. SVM is combined with SNORT and some basic rule sets of firewall which allows it to generate a new and effective technique for intrusion detection. The SVM classifier is also used with SNORT to reduce false alarm rate and improve accuracy of IPS. SVM IDS techniques can prove the best techniques for intrusion detection in cloud which can enhance its current feature and extends its security level upto a considerable level.

G) Genetic Algorithm based IDS

GAs use confidence based fitness functions for intrusion detection which classifies network in a very efficient manner. These values can be used and determined for the profile generation as well. These services are very much useful in cases where intrusion behaviors are very dynamic in nature. These techniques can be collaborated with other techniques which are resource intensive and prioritize the overall performance of the system. These techniques use training period and determine the fitness value based on the trained profiles. However, GA can be integrated with other such techniques for better results in cloud technology. This feature is more important than any other techniques involved for intrusion detection. It is also suitable in scenarios wherever there is a need of mutual authentication in cloud among users.

H) Hybrid techniques

Hybrid techniques combine various such technologies together for a better result in sense of intrusion detection. This is such a kind of technology which contains various flavours related to other techniques. NeGPAIM is based on hybrid technique combining two low level components including fuzzy logic for misuse detection and neural networks for anomaly detection, and one high level component which is a central engine analyzing outcome of two low level components. This is an effective model and does not require dynamic update of rules.It is more suitable to be integrated with soft computing techniques which are traditional ones or focused towards intrusion detection. With pros and cons of every technique, this is also not an exception. Some of the limitations under this technique are mainly oriented towards training profiles, period and rules. However, there are various other techniques which can be clubbed with this one to improve the efficiency of the overall system. The lead role in this technique is of algorithm which makes it stand clear form other techniques.

V. CONCLUSION

In this paper, different types of attacks on cloud computing are discussed. This paper has also done a comparison on the different intrusion detection techniques for the securing the cloud from various attacks.

VI. REFERENCES

[1]. Chirag Modi, Dhiren Patel, Bhavesh Borisaniya, Hiren Patel, A Survey of intrusion detection technques in Cloud, Journal of Network and Computer Applications, Elsevier, 2013, pp. 42-57

[2]. Mohamed, A., Grundy, J., Ibrahim, A. S.: Adaptable, model-driven security engineering for SaaS cloud-based applications. Automated Software Engineering, vol. 21, pp. 187--224. Springer (2013)

[3]. Ye Du, Li, R. Z. M.: Research on a Security Mechanism for Cloud Computing based on Virtualization. Telecommunication Systems, vol. 53, pp. 19—24, Springer (2013)

[4]. Edurado, F. B., Monge. R., Hashizume K.: Building a Security Reference Architecture for Cloud Systems., Requirements Engineering, pp. 1—25. Springer (2015)

[5]. Jin, H., Dong, M., Ota, K., Fan, M., Wang, G.: NetSecCC : A Scalable and Fault Tolerant Architecture for Cloud Computing Security. Peer-to-peer Networking and Applications, pp. 1—15, Springer (2014)

[6]. P, Hu., Sung C. W., Ho, S., Chan, T. H.: Optimal Coding and Allocation for Perfect Secrecy in Multiple Clouds, Information Forensics and Security, vol. 11, pp. 388-399, IEEE (2014)

[7]. Junwon, L., Cho, J., Seo, J., Shon, T., Won, D.: A Novel Approach to Analyzing for Detecting Malicious Network Activity Using a Cloud Computing Testbed, Mobile Networks and Applications, vol. 18, pp. 122-128, Springer (2012)

[8]. Jin, L., Li, Y. K., Chen, X., Lee, P. P. C., Lou, W.: A Hybrid Cloud Approach for Secure Authorized Deduplication, Parallel and Distributed Systems, vol. 26, pp.1206--1216, IEEE Transactions (2014)

[9]. Rahat, M.,Shibli, M. A., Niazi, M. A.: Cloud Identity Management Security Issues and Solutions : A Taxonomy, Complex Adaptive Systems Modeling, vol. 2, pp. 1 – 37, Springer (2014)

[10]. Seungmin, R., Chang, H., Kim, S., Lee, Y. S.: An Efficient Peer-to-peer Distributed Scheduling for Cloud and Grid Computing, Peer-to-peer Networking and Applications, vol. 8, pp. 863 – 871, Springer (2014)

[11]. Li, Q., Han, Q., Sun, L.: Collaborative Recognition of Queuing Behavior on Mobile Phones, Mobile Computing, vol. 15, pp. 60 – 73, IEEE (2014)

[12]. Tak, G. K., Badge N., Manwatkar, P., Rangnathan, A., Tapaswi, S.: Asynchronous Anti Phishing Image Captcha Approach towards Phishing, International Conference on Future Computer and Communication, vol. 3, pp. 694 – 698, IEEE (2010)

[13]. Malhotra, K., Gardner, S., Patz, R.: Implementation of elliptic-curve cryptography on mobile healthcare devices, IEEE (2007)

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

A Proposed Algorithm to Enhance Security in CRN

Kriti, Dr. Ajay Kaul
Department of Computer Science and Engineering SMVD University, Jammu & Kashmir, India

ABSTRACT

With the growing demand of the wireless spectrum leading to its scarcity, motivated the use of Cognitive Radio as the successful way to deal with this problem. The efficient exploitation of the spectrum is done by the Cognitive Radio that allows the licensed spectrum to be utilized by the unlicensed users effectively. This paper illustrates an algorithm which enhances the security in CRN irrespective of the security threat or the attack done by the malicious users. It uses the concept of cluster head generated through random numbers and the formation of slots for the free spectrum based on the round robin algorithm. The efficient management of the spectrum is done so that each user utilizes the spectrum in an effective way without causing harm to the primary users.

Keywords : primary user (PU); secondary user (SU), cognitive radio network (CRN), primary user emulation (PUE)).

I. INTRODUCTION

With the passage of time, tremendous advances have been shown by the wireless communication that includes the network which is being complemented by the systems that are not only self- organizing but also heterogeneous with the infrastructure being hybrid adding the communication nodes of peer-to-peer. The researchers are attracted towards the cognitive radio in these days, where the frequency band, the sharing of frequency is realized for the assignment to the primary system [1]. The secondary cognitive terminal senses the frequency band being assigned to the primary systems, by transmitting the signals without causing any interference between the two. But the situation of interference is fluctuated, in the cognitive networks, in terms of time, frequency and location. Therefore for the ad-hoc cognitive network, the basic techniques for routing are not effective [2].

NEED FOR CRN- the motivation

To establish and manage, a wireless networks for cognitive radio, a new model is proposed using the trainable and the adaptive radio. A number of intelligent tasks are implied in cognitive radio in the motion of cognition and hence robust knowledge is required to represent the facility of sharing and knowledge reuse. The capability of reconfiguring the infrastructure defines the cognitive network, which adapts itself to the continuously changing environment in the network, for machine learning techniques. To support the decision making, the learning engines have been proposed for the services and applications which are context-aware. In turning these models of learning, are the challenges, into viable commercial products [8].

II. BACKGROUND AND HISTORY OF COGNITIVE RADIO

Spectrum is a very limited product and due to the spectrum insufficiency facing by the wireless based service providers lead to high overcrowding stages. The main reason that leads to useless utilization of the radio spectrum is the licensing system itself. If the allocated radio spectrum is not used by primary users then it also cannot be utilized by SUs. As a result, wireless systems are intended to work only on a devoted band of spectrum for fixed and rigid allocations. It cannot change the band as changing the surroundings. As illustration if one channel of spectrum band is greatly used, the wireless system cannot alter to work on any other more lightly used band.

The authorized access of spectrum is usually defined by owner of spectrum; transmit power, frequency, space and the license duration. In general, a license is allocated to one licensee and the use of band by this owner must have the requirement e.g. highest power of transmit, base station location. In present spectrum licensing system, the license cannot vary the application or giving the access to another licensee. This restriction causes in low consumption of the frequency spectrum. Spectrum hole is defined as a group of frequencies given to a licensee, except that user is not using the band at exact time and exact geographic location [10].

The allocation of radio band is in power of the central government. Federal Communications Commission (FCC) printed a statement in beginning of 20th century which was ready by the spectrum plan mission force that was intended to improving this expensive source. The allotment of the unlicensed frequency bands leads to the congestion of these bands.

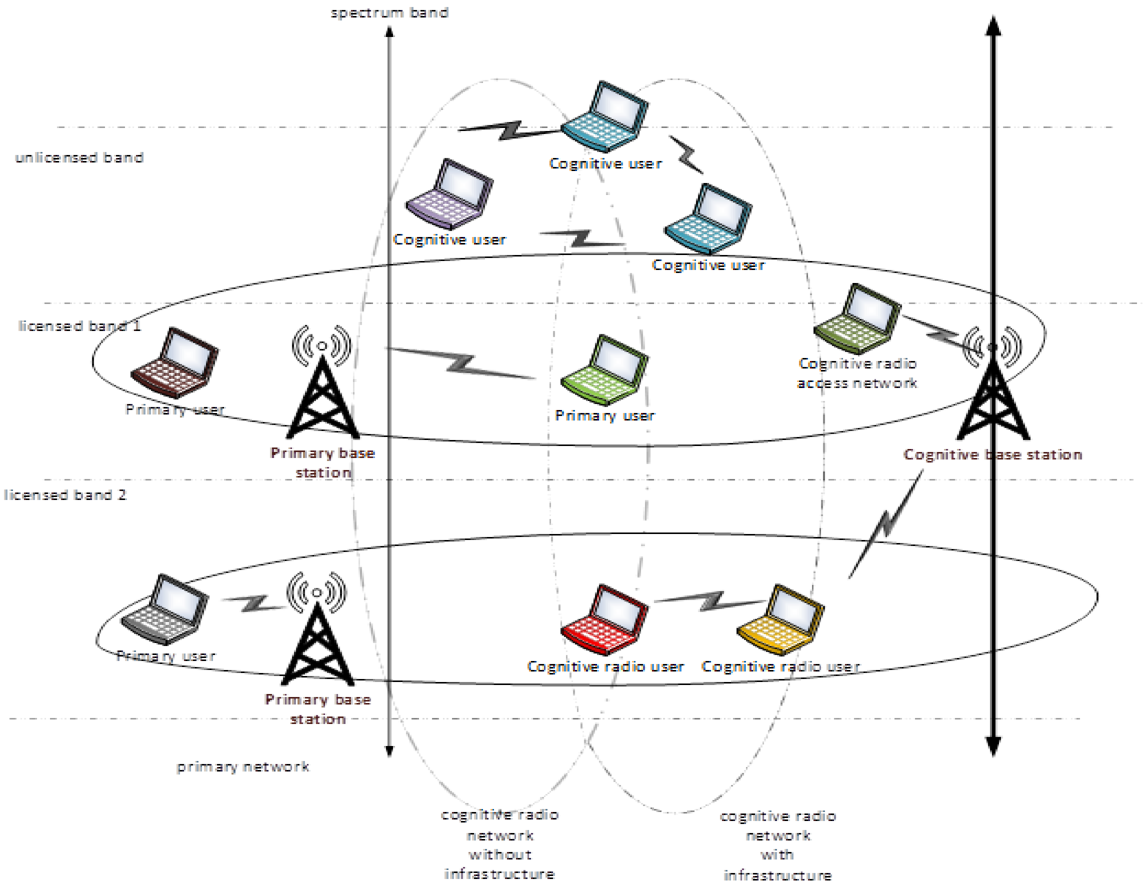


Figure 1. Cognitive Radio Network

A. Cognitive Radio Network: an introduction

The ability to sense the spectrum by software radio intellectuality and to seek the spectrum hole by automatic sensation of electromagnetic environment, cognitive radio is used which adjust the optimum condition by bi-lateral signal parameter of the communication protocol and the algorithms.

Cognitive Radio (CR) is the category of wireless system in which either an entire network or a single node varies its communication or response parameter to correspond effectively. It avoids obstruction with primary user (PU) and secondary user (SU). It is considered to be an intelligent communication system which is sensitive of surrounding atmosphere and uses the techniques to gain knowledge from the surroundings and adjust its internal conditions to arithmetic changes in the arriving RF by creating consequent variation in definite working factors.

A CR is intended to be alert and responsive to that alters in its neighboring that makes spectrum sensing an imperative necessity for the understanding of secondary networks. Spectrum sensing method allows SUs to use the vacant spectrum segment adaptively to the radio atmosphere [12].

B. Fundamentals of CRN

1. **CR characteristics:** The two fundamental resources for communication which are energy and bandwidth are scarce which in turn limits the service quality and channel capacity. The new communication and network paradigm fetched the attention of the researchers to utilize the scarce resources efficiently and intelligently.

2. **CR function:** The various functions performed by CR include spectrum sensing, analysis, management and handoff etc. Spectrum sensing and analysis includes detection of white space in the spectrum and then utilize it. And in order to avoid harmful interference to PU when they again start using the spectrum, CR does sensing. The spectrum management and handoff function enable the choice of best frequency band available.
3. **Network Architecture and application:** The secondary network and the primary network are the major components in CR network architecture, where the SU are the unauthorized user that utilizes the unused unlicensed band temporarily owned by PU. Also CR functionality is present in both SU and secondary base station. The spectrum broker is a central network that coordinates spectrum usage if one common spectrum band is being used by several secondary networks. Cognitive communication increases spectrum efficiency and also supports services that require high bandwidth by sensing, detecting and monitoring the RI environment surrounding it.

C. The Classification Based on Network Architecture

One is centralized in which the central unity is held responsible for controlling and coordinating the spectrum allocation and access of SU. The other classification is based on the access behaviour of SU. One is cooperative in which all SU focuses towards a common goal. The other is non cooperative where they no longer cooperate to achieve common objective [13].

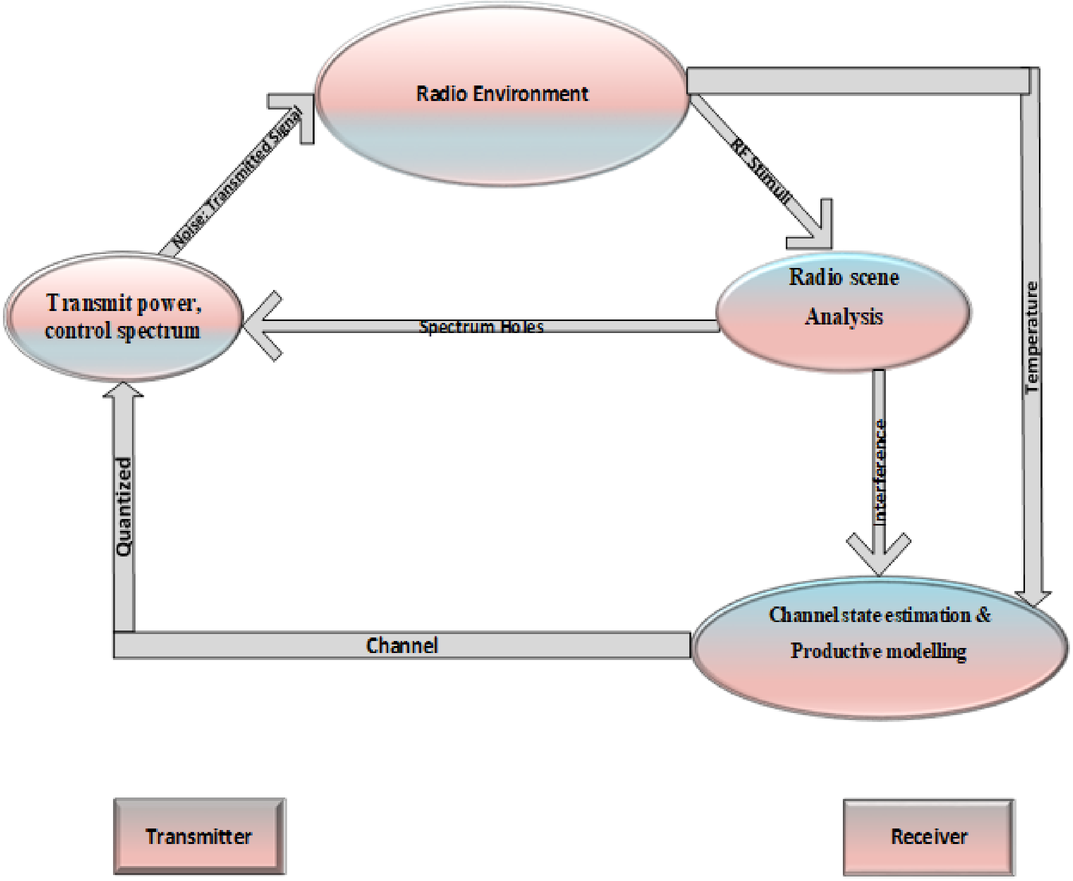


Figure 2. Cognitive cycle model

A. Security Threats in CRN and the PUE Attack

The PUE attack is further classified into two types:

1) Selfish PUE Attacks: To maximize the spectrum usage for itself is the objective of the attacker. When a fallow spectrum band is detected by the attacker, this prevents the SU from competing against particular band.
2) Selfish PUE Attacks: When the DSA process obstructed, the prime objective of the attack is fulfilled to harm the legitimate secondary user. This attack leads to denial of service. The fallow spectrum band is not necessarily used by the malicious attacker to serve its communication purpose unlike the selfish attacker [18].

III. METHODS TO DETECT MALICIOUS USERS

The performance of the system for cooperative sensing is significantly affected by the presence of the nodes which are malicious. Due to the malfunctioning of the device or some selfish reason the node acts maliciously. For example if the absence of signal is detected by the node but it might generate false positive, and the wrong decision is taken by the access point considering the presence of the primary signal and hence the malicious node can transmit its own signal selfishly over the free channel available.

Different type of malicious node has been considered 'Always Yes' node or 'Always No' nodes are simple malicious nodes. In case of 'Always Yes' the value given all the time is above the threshold and in case of 'Always No' the value given is below the threshold. With 'Always Yes', the probability of false alarm Pf is increased whereas with 'Always no' the probability of detection decreases.

B. PRE-FILTERING OF THE SENSING DATA

To identify and then remove the node which is malicious, pre-filtering of the data sensed is essential which in turn affect the final decision significantly at the access point hence giving values that are extremely false.

C. TRUST FACTORS:

To give the reliable measure, the trust factor is used for a particular user. While the calculation of the mean for the values of energy which are obtained for the various users, the trust factor are hence used as the weighing factor.

D. QUANTIZATION:

The need to quantize the energy value before sending to the access point is essential since limited bandwidth is offered by the control channels. Hence leads to the extensive studies of the schemes for optimal quantization for the distributed detection. However, it is highly complex and moreover the problem of optimization is non-linear to find the optimal threshold value [20].

Hence in many of the frequency bands, there is low usage of spectrum due to the conventional fixed spectrum allocation policy. And to exploit this under-utilized spectrum the promising technology proposed is CRN [21].

IV. PROPOSED ALGORITHM

As mentioned in [22], a trust- based system in be defined to prevent the PU and SU from various attacks. The author in the paper has built a trust model for CRN. Basically after checking trustworthiness which depends on a trust value it assigns free spectrum to the SU. The communication activity of the SU depends on the availability of the free spectrum. Hence using stochastic approach the author proposed Markov model showing the availability of spectrum for SUs and addressing the corresponding SU as authenticated user. The cognitive nodes are all included to calculate the trust level for all its surrounding nodes and which in turn are stored for later use. Also based on the new interactions, these values will be updated.

Table 1

Existing Method Drawbacks	**Proposed Method Advantage**
Defining the correct threshold for trust level.	It does not involves defining any threshold.
To set up a trust value it involves each cognitive node.	The cognitive node involved can be malicious node, hence it can be easily detected using methodology given below.
PU has to check the trustworthiness of SU on the demand to the available free spectrum.	PU has no role to be played in spectrum assignment hence reducing its overhead.
Assignment of free spectrum by PU.	The SU searches for free spectrum.

Each cognitive node will calculate trust for all its surrounding nodes and store these values for later use; these values should be updated in a specific time period based on new interactions. Hence in the proposed algorithm the secondary users are the in charge for allocating the free spectrum to them. One of the SU is selected as the cluster head for the region of availability of spectrum for particular primary user. The selection of the SU as cluster head is primarily done on the basis of random number generation. Then following the round robin pattern the available spectrum is divided into equal slots depending upon the SU demanding for that free spectrum. This pattern does not involve either defining trust value or the overhead caused to the PU to allot the free spectrum. Moreover it will help detecting the malicious user in the vicinity as uneven slots of spectrum will otherwise be created due to its intervention in the spectrum assignment.

Proposed Methodology:

Steps to be followed are:

- Step 1:

Creation of CRN.

A network that consist of:

1) **Various primary users and the secondary users.**

2) **The primary base station and the secondary base station.**

- Step 2:

SU searches for free spectrum

For free spectrum of PU

Select one of the SU as the cluster head using random number generator.

Selected SU form slots for each SU demanding the free spectrum based on Round Robin algorithm.

Allocate the spectrum

If any unevenness in allotment detected for particular SU report it as malicious node.

ELSE

REPEAT UNTIL

Spectrum is available and demand is still not fulfilled

End

V. OPEN ISSUES AND FUTURE RESERCH DIRECTION

The issue of security fragility is one issue that cannot be resolved easily. The requirement of fundamental security is violated by the SU because of sensing where the legitimate analysis of the traffic is performed for the utilization of the spectrum, it is compromise in security.

In the wireless communication, the security and reliability trade-off is an important consideration in the presence of eavesdropping attack. Also various security algorithms thus that are efficient from the energy point of view as well as the low in complexity.one desirable in making CR technology viable solution in the wireless communication for the future generation.

VI. CONCLUSION

In cognitive radio networks, some malicious secondary users may create interference by accessing the primary user's available spectrum band. Such malicious SUs can seriously break down the whole network performance. To tackle this problem, we want to redefine a trust based model to check the trustworthiness of the secondary user who wants to use primary user's free spectrum band. After allowing the SU to form a cluster head to other SU in a PU vicinity, that SU allocate the spectrum forming slots to each SU in demand of the spectrum. Also the malicious activity can be easily detected. Hence the proposed algorithm not only reduces the overhead of the PU to spectrum allocation but also detect malicious node in an efficient way.

VII. REFERENCES

[1]. FCC, "Spectrum Policy Task Force Report," Vol. ET Docket Issue 02-155, November 2002.

[2]. Danijela Cabric, Shridhar Mubaraq Mishra, Robert W. Brodersen, " Implementation Issues in Spectrum Sensing for Cognitive Radios," Signals, systems and computers, 2004. Conference record of the thirty-eighth Asilomar conference on, Vol. 1, pp. 772-776. IEEE, 2004.

[3]. MacKenzie, Allen B. and Stephen B. Wicker. "Game theory in communications: Motivation, explanation, and application to power control." In Global Telecommunications Conference, 2001. GLOBECOM'01. IEEE, vol. 2, pp. 821-826. IEEE, Year 2001.

[4]. Zhu, Xiaorong, Lianfeng Shen and T-SP Yum. "Analysis of cognitive radio spectrum access with optimal channel reservation."

Communications Letters, no. 4 pp. 304-306.IEEE, Year 2007.

[5]. Ji, Zhu and KJ Ray Liu. "Cognitive radios for dynamic spectrum access-dynamic spectrum sharing: A game theoretical overview." Communications Magazine, IEEE 45.5 pp. 88-94, Year 2007.

[6]. Cheng, Shilun and Zhen Yang. "Energy-efficient power control game for cognitive radio systems." In Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2007. SNPD 2007. Eighth ACIS International Conference on, vol. 1, pp. 526-530. IEEE, Year 2007.

[7]. Liu, Jishun, Lianfeng Shen, Tiecheng Song and Xiaoxia Wang. "Demand-matching spectrum sharing game for non-cooperative cognitive radio network." In Wireless Communications & Signal Processing, 2009. WCSP 2009. International Conference on, pp. 1-5. IEEE, 2009.

[8]. P. Rostaing,T. Pitarque, E. Thierry, " PERFORMANCE ANALYSIS OF A STATISTICAL TEST FOR PRESENCE OF CYCLOSTATIONARITY IN A NOISY OBSERVATION," In Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on, Vol. 5, pp. 2932-2935, 1996.

[9]. V. Srivastava and M. Motani, "Cross-layer design: a survey and the road ahead," Communications Magazine, IEEE, Vol. 43, Issue 12, pp. 112-119, December 2005.

[10]. Mishra, Shridhar Mubaraq, Danijela Cabric, Chen Chang, Daniel Willkomm, Barbara Van Schewick, Adam Wolisz, and Robert W. Brodersen, "A real time cognitive radio testbed for physical and link layer experiments," New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on, pp. 562-567. IEEE, 2005.

[11]. Danijela Čabrić and Robert W. Brodersen, "Physical Layer Design Issues Unique to Cognitive Radio Systems,"Personal, Indoor and Mobile Radio Communications, 2005. PIMRC 2005. IEEE 16th International Symposium on, Vol. 2, pp. 759-763. IEEE, 2005.

[12]. Ian F. Akyildiz, Won-Yeol Lee, Mehmet C. Vuran, and Shantidev Mohanty, "Next generation/ dynamic spectrum access/cognitive Radio Wireless Networks: A Survey," COMPUTER NETWORKS JOURNAL ELSEVIER," Vol. 50, pp. 2127-2159, 2006.

[13]. William A. Gardner, Antonio Napolitano, and Luigi Paura, "Cyclostationarity: Half a century of research," Signal Processing, Vol. 86, Issue 4, pp. 639-697, 2006.

[14]. Kyouwoong Kim, Ihsan A. Akbar, Kyung K. Bae, Jung-sun Um, Chad M. Spooner, and Jeffrey H. Reed, "Cyclostationary Approaches to Signal Detection and Classification in Cognitive Radio," New frontiers in dynamic spectrum access networks, 2007, pp. 212-215,2007.

[15]. Alex Chia-Chun Hsu, David S. L. Wei† and C.-C. Jay Kuo, "A Cognitive MAC Protocol Using Statistical Channel Allocation for Wireless Ad-hoc Networks," Wireless Communications and Networking Conference, pp. 105-110. IEEE, 2007.

[16]. Lundén, Jarmo, Visa Koivunen, Anu Huttunen, and H. Vincent Poor. "Spectrum sensing in cognitive radios based on multiple cyclic frequencies."Cognitive Radio Oriented Wireless Networks and Communications, 2007. CrownCom 2007. 2nd International Conference on, pp. 37-43. IEEE, 2007.

[17]. Qing Zhao and B.M. Sadler, "A Survey of Dynamic Spectrum Access," Signal Processing Magazine, IEEE, Vol. 24, Issue 3, pp. 79-89, may 2007.

[18]. ShiyuXu, Zhijin Zhao, Junna Shang, “Spectrum Sensing Based on Cyclostationarity,” Power Electronics and Intelligent Transportation System, 2008. PEITS'08. Workshop on, pp. 171-174. IEEE, 2008

[19]. Jun Ma, Guodong Zhao and Ye (Geoffrey) Li, "Soft Combination and Detection for Cooperative Spectrum Sensing in Cognitive Radio Networks," IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, pp. 4502-4507, Vol. 7, no 11, November 2008.

[20]. Kim, Hyoil, and Kang G. Shin, "In-band spectrum sensing in cognitive radio networks: energy detection or feature detection" in Proceedings of the 14th ACM international conference on Mobile computing and networking, pp. 14-25. ACM, 2008.

[21]. Zayen, Bassem, A. M. Hayar, and Dominique Nussbaum, "Blind spectrum sensing for cognitive radio based on model selection," in Cognitive Radio Oriented Wireless Networks and Communications, 2008. CrownCom 2008. 3rd International Conference on, pp. 1-4. IEEE, 2008

[22]. Parvin, Sazia, Farookh Khadeer Hussain, Omar Khadeer Hussain, and Abdullah Al Faruque. "Trust-based Throughput in Cognitive Radio Networks." Procedia Computer Science 10 (2012): 713-720.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Static Malware Analysis : A Case Study

Satya Narayan Tripathy, Sisira Kumar Kapat, M. Soujanya, Susanta Kumar Das
Department of Computer Science Berhampur University, Odisha, India

ABSTRACT

In the arena of digitized era, everyone needs internet connectivity for seeking and sharing of information. Starting from sharing information to social networking, each task requires internet. Some of the malware take advantage of this, and use user activities to activate. Hence the vector will be SDN (Software Defined Network) and SNS (Social Networking Sites). In both the cases, the user cannot be pretended to be a malware specialist or a computer professional who can detect the malicious activity easily. Although a lot of anti-malware tools are available, but it is good if the user can predict the malware. This paper focuses to analyze a malware easily and effectively, which a normal user can capture.

Keywords : Malware, Static Analysis, Case Study, Portable Executable, Common Strings.

I. INTRODUCTION

Malware infection is trending in SNS and SDN with the growth of new technologies. The recent topics in malware infection include Ransomware, ad fraud malware, android malware, botnets, banking Trojans and adware. This paper focuses on the static analysis of malware; including analysis of three malwares as examples. This research activity is based on simple concepts which a normal user can understand. The complete research is to detect (or can predict) malware existence in user computer. This paper considers the user is not a computer professional so we do not consider the malware family rather we focus on the files or process collected from user computer. Here three examples are given to make a general understanding about the malwares.

Malware is a computer program which directly or indirectly affects another computer program [9]. To be a malware a computer program must satisfy either of the malware criteria as, follows.

1. It must replicate itself and/or
2. Copy, remove or modify other files or programs.

In most of the cases, the computer malware is a PE (Portable Executable) file. The portable executable file has several sections. Out of the several sections, PE header is the most common one. The user can open any suspected file in an editor (like notepad). If the first two letters are found to be 'MZ', then the user may confirm that, the file is a PE file. One example is presented in **Error! Reference source not found.** and **Error! Reference source not found.**. The file which is placed in the 'startup' will be initialized first when the computer system starts. So the most common classic place of malware is 'startup'. The user can search the location before searching any other location in the system.

The existing malware analysis is based on some tools and techniques. The most common tools used for malware analysis are x32dbg/x64dbg, API monitor, PE explorer etc. and virtual machine is the most common environment for malware analysis. Some malware uses anti-VME (anti virtual machine environment) technique [8]. This implies the malware cannot be executed successfully in a virtual machine for analysis. There are two methods available for malware analysis, such as static malware analysis and dynamic malware analysis.

In static malware analysis [10], the malware is analyzed before it runs in the device. It includes, disassembling the source program which is not possible for each case. In dynamic analysis, the malware is executed and monitored the behavior. Dynamic analysis can monitor the instructions by executing the malware in a virtual environment such that, it won’t affect the host operating system. The malware is then passed through behavioral study. In both the cases some merits and demerits are there but still static analysis beats dynamic analysis in speed [11,12].

II. RELATED WORK

According to Andrew & Srinivas[1], signature based detection was good. So they focused on modifying the existing signature detection technique. They analyzed the malware and collected the API call sequence. Each windows API is mapped into 32 bit integer id number. The obtained signature is used for similarity measurement with the existing signature database. For similarity measurement Euclidian distance, sequence alignment and similarity function methods is used. The analysis is based on static scanning means no sandboxing, proxy testing or code de-obfuscation. This technique holds good for polymorphic and metamorphic malware.

Madhu et al.[2], presented a methodology for composing signature of malware codes from Portable Executable is presented. They presented two methods for malware detection such as Static Analyzer for Vicious Executables (SAVE) and Malware Examiner using Disassembled Code (MEDiC). MEDiC uses assembly calls for analysis and SAVE uses API calls (Static API call sequence and Static API call set) for analysis. According to them, Assembly can be superior to API calls in that it allows a more detailed comparison of executable. API calls, on the other hand, can be superior to Assembly for its speed and its smaller signature.

Karishma et al.[3] collect some data set from benign files and spyware files and then they used java programming for Hexadecimal dumps for byte sequence Generation .Then they generated n-gram, for that they used hexadecimal dumps to converted into 'n' of fixed size and stored in HashMap, for later updation in the database. These are used for feature extraction by using Frequency-based Feature Extraction (FBFE) approach. The features with high frequency are being considered for training the classifier. The features with low frequency are ignored after this step the classifier is trained for model training they used Naïve Bayes Algorithm for classifying. The limitation of their approach is, regular explicitly searches for a process.

Ankur[4] discusses about basic outline of malicious codes and especially spywares and their detection using different techniques and also they told that the installation of spyware normally involves Internet Explorer. The main reason is the popularity of internet explorer that has made it target of spywares. Its deep integration with the Windows environment makes it prone to attack into the Windows operating-system. Internet Explorer also serves easy environment for spyware in the form of Browser Helper Objects, which modify the browser's behavior to add toolbars or to redirect traffic.

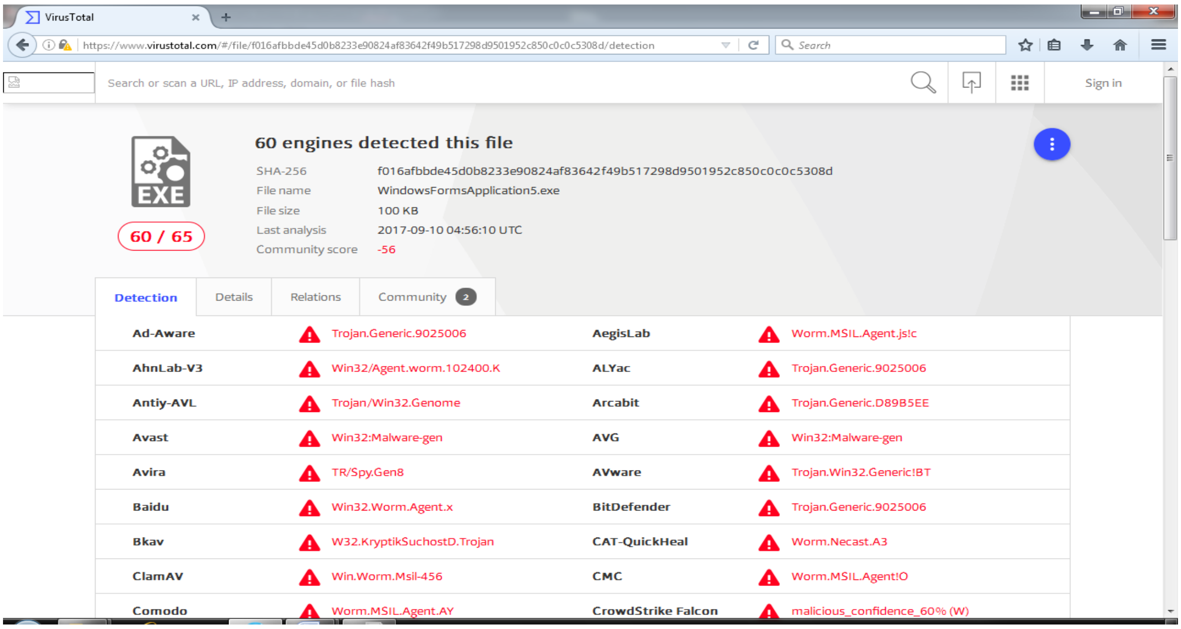
Gerardo et.al.[5] introduces a detection technique that assume that a side effect of the most common metamorphic engines it is the dissemination of a high number of repeated instruction in the body part of virus. Also they evaluate their technique. That method is more effective for static analysis rather than dynamic analysis. They used frequency analysis of instruction presetting disassemble code to detect the malware.

Case study

This paper considers three cases, where the malwares considered are, “New folder.exe”, “tongji.js” and “suchost..exe”. These malwares were in one of the infected device which was left infected purposefully. The files which are collected from the device are analyzed and discussed below as the case studies.

Case-1: suchost..exe (svchost..exe)

The ‘suchost..exe’ file is a PE (Portable Executable) file. This file runs another process called ‘svchost..exe’ which sounds similar to svchost.exe.The svchost.exe is a system file which is well visible in the task manager of any computer system.To confirm this file as a malware, this file if processed in virustotal assuming that, all the anti-malware engines used in virustotal is updated. The below Figure 1. (showing suchost..exe is a malware, virustotal output) shows that out of 65, 60 malware engines confirms this file to be a malware.


VirusTotal
https://www.virustotal.com/#/file/f016afbbde45d0b8233e90824af83642f49b517298d9501952c850c0c0c5308d/detection
Search
Search or scan a URL, IP address, domain, or file hash
Sign in
EXE
Detection
Details
Relations
Community 2

60 engines detected this file

60 / 65

SHA-256	f016afbbde45d0b8233e90824af83642f49b517298d9501952c850c0c0c5308d
File name	WindowsFormsApplication5.exe
File size	100 KB
Last analysis	2017-09-10 04:56:10 UTC
Community score	-56

Ad-Aware	⚠ Trojan.Generic.9025006	AegisLab	⚠ Worm.MSIL.Agent.js!c
AhnLab-V3	⚠ Win32/Agent.worm.102400.K	ALYac	⚠ Trojan.Generic.9025006
Antiy-AVL	⚠ Trojan/Win32.Genome	Arcabit	⚠ Trojan.Generic.D89B5EE
Avast	⚠ Win32:Malware-gen	AVG	⚠ Win32:Malware-gen
Avira	⚠ TR/Spy.Gen8	AVware	⚠ Trojan.Win32.Generic!BT
Baidu	⚠ Win32.Worm.Agent.x	BitDefender	⚠ Trojan.Generic.9025006
Bkav	⚠ W32.KryptikSuchostD.Trojan	CAT-QuickHeal	⚠ Worm.Necast.A3
ClamAV	⚠ Win.Worm.Msil-456	CMC	⚠ Worm.MSIL.Agent!O
Comodo	⚠ Worm.MSIL.Agent.AY	CrowdStrike Falcon	⚠ malicious_confidence_60% (W)

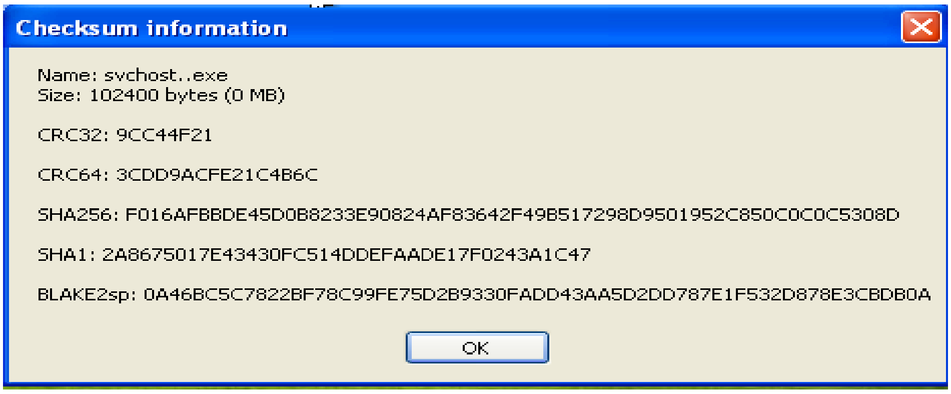
Figure 1. (showing suchost..exe is a malware, virustotal output)

Whenever suchost..exe runs,

1. It copies itself to two positions as (collected from source file)
   a. AppData\Roaming\Microsoft\Windows\Start Menu\Programs\Startup \svchost..exe
   b. SystemDrive /\Documents and Settings \Users /\Documents\suchost..exe
2. It creates two processes as suchost..exe and svchost..exe, which can be clearly visible in task manager. It may not run in win-xp environment and it requires .NET framework to run in post win-7 environment.
3. Creates two registry keys
   a. (System drive :\...)\AppData\Roaming\Microsoft\Windows\Start Menu\Programs\Startup\svchost..exe
   b. (System drive :\...)\Documents\suchost..exe
4. The two registry keys given above are placed in four registry locations such as,
   a. HKEY_CLASSES_ROOT/Local Settings/Software/Microsoft/windows/ shell/MuiCache

b. HKEY_CURRENT_USER/Software/Classes/Local Settings/Software/ Microsoft/windows/shell/MuiCache
c. HKEY_USERS/s-1-5-21-2528868183-3685655094-3686021654-1000/ Software/Classes/Local Settings/Software/Microsoft/windows/shell/ MuiCache
d. HKEY_USERS/s-1-5-21-2528868183-3685655094-3686021654-1000_Classes/ Local Settings/Software/Microsoft/windows/ shell/MuiCache

5. The Checksum information collected from win-xp sp2, 32 bit environment is as shown in Figure 2 . (Showing Checksum information of suchost..exe):

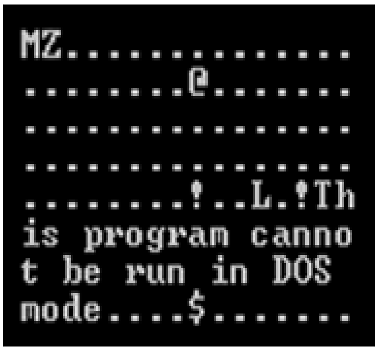


Checksum information

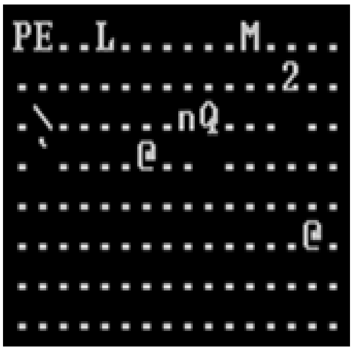
```
Name: svchost..exe
Size: 102400 bytes (0 MB)

CRC32: 9CC44F21

CRC64: 3CDD9ACFE21C4B6C

SHA256: F016AFBBDE45D0B8233E90824AF83642F49B517298D9501952C850C0C0C5308D

SHA1: 2A8675017E43430FC514DDEFAADE17F0243A1C47

BLAKE2sp: 0A46BC5C7822BF78C99FE75D2B9330FADD43AA5D2DD787E1F532D878E3CBDB0A
```

Figure 2 . (Showing Checksum information of suchost..exe)

To be clear about this file, the user may open this file in a notepad or the user may use some external software for analysis. Some portions of 'suchost..exe' is presented in the figures. From the **Error! Reference source not found.** and **Error! Reference source not found.**it is clear that the file starts with letter 'MZ' and figure shows the PE header of 'suchost..exe' file, which refers to the file is a PE (Portable Executable) file and figure.



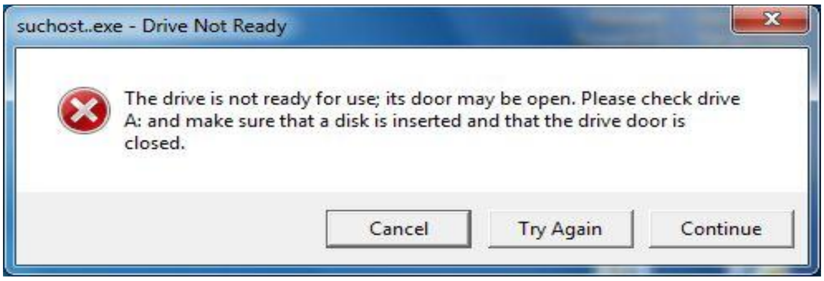
```
MZ..............
........@.......
................
................
........!..L.!Th
is program canno
t be run in DOS
mode....$.......
```



```
PE..L......M....
.............2..
.\......nQ... ..
.`....@.. ......
................
..............@.
................
................
```

Figure 3. (Shows 'MZ' is the first two letters) **Figure 4.** (Shows PE header of suchost..exe)

In 32 bit environment the file shows a message dialog box, whose message body is "drive not ready" and this is an infinite loop hence the user will be unable to close the dialogue box. The message box is shown in Figure 3. (Drive not ready message by suchost..exe)below.



The drive is not ready for use; its door may be open. Please check drive A: and make sure that a disk is inserted and that the drive door is closed.

Figure 3. (Drive not ready message by suchost..exe)

Case-2: tongji.js

```
<script type="text/javascript" src="http://web.nba1001.net:8888/tj/tongji.js">
</script>
```

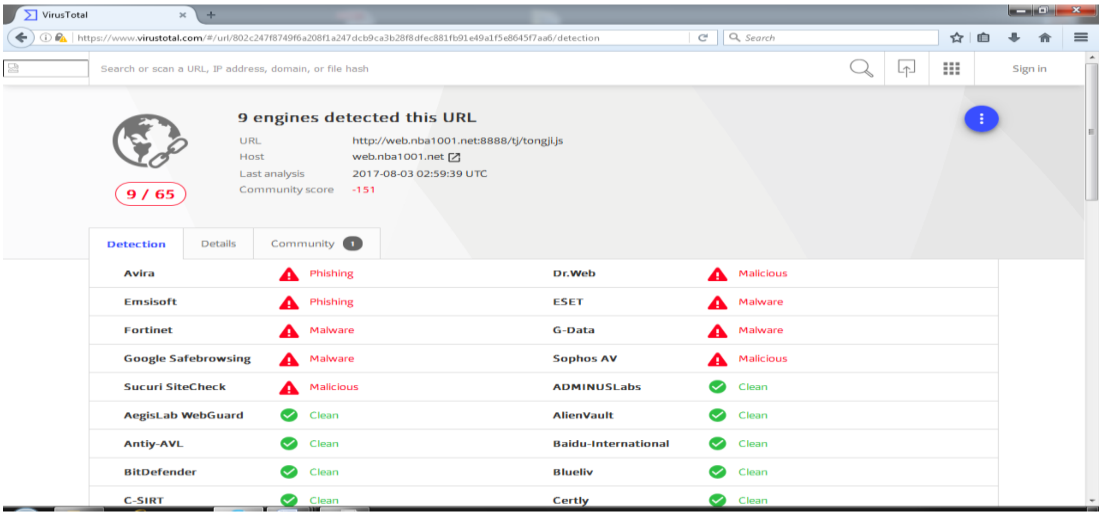
Figure 4. (Embeded malicious source in an HTML file)

The code statement is collected from one of the infected device. This is a malware which appends its execution code statement to an HTML file. This statement calls the original javascript program using the URL and executes it whenever the html program is connected to the network. The architecture of this malware is shown in Figure 5. (Architecture of 'tongji' malware).

HTML Document
Malware Code (Append)

Figure 5. (Architecture of 'tongji' malware)

This URL is being analyzed with virustotal and the result is shown in Figure 6. (showing tongji link given above is malicious, virustotal output).



VirusTotal
https://www.virustotal.com/#/url/802c247f8749f6a208f1a247dcb9ca3b28f8dfec881fb91e49a1f5e8645f7aa6/detection
Search
Search or scan a URL, IP address, domain, or file hash
Sign in
9 / 65
Detection
Details
Community 1

9 engines detected this URL

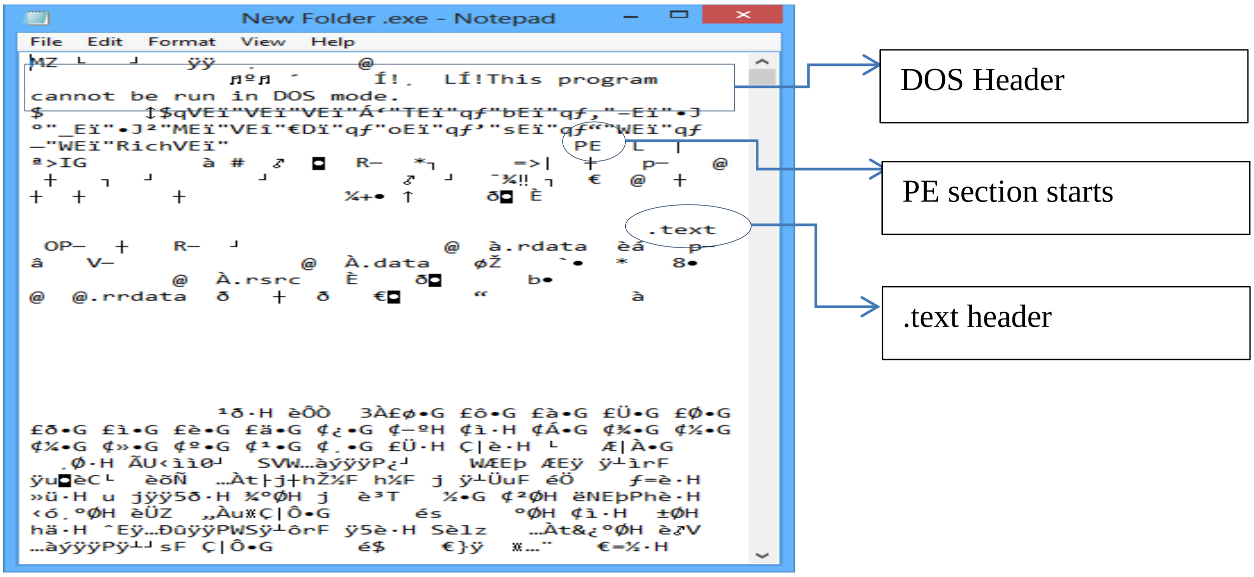
URL	http://web.nba1001.net:8888/tj/tongji.js
Host	web.nba1001.net
Last analysis	2017-08-03 02:59:39 UTC
Community score	-151

Avira	Phishing	Dr.Web	Malicious
Emsisoft	Phishing	ESET	Malware
Fortinet	Malware	G-Data	Malware
Google Safebrowsing	Malware	Sophos AV	Malicious
Sucuri SiteCheck	Malicious	ADMINUSLabs	Clean
AegisLab WebGuard	Clean	AlienVault	Clean
Antiy-AVL	Clean	Baidu-International	Clean
BitDefender	Clean	Blueliv	Clean
C-SIRT	Clean	Certly	Clean

Figure 6. (showing tongji link given above is malicious, virustotal output)

Case-3: NewFolder.exe

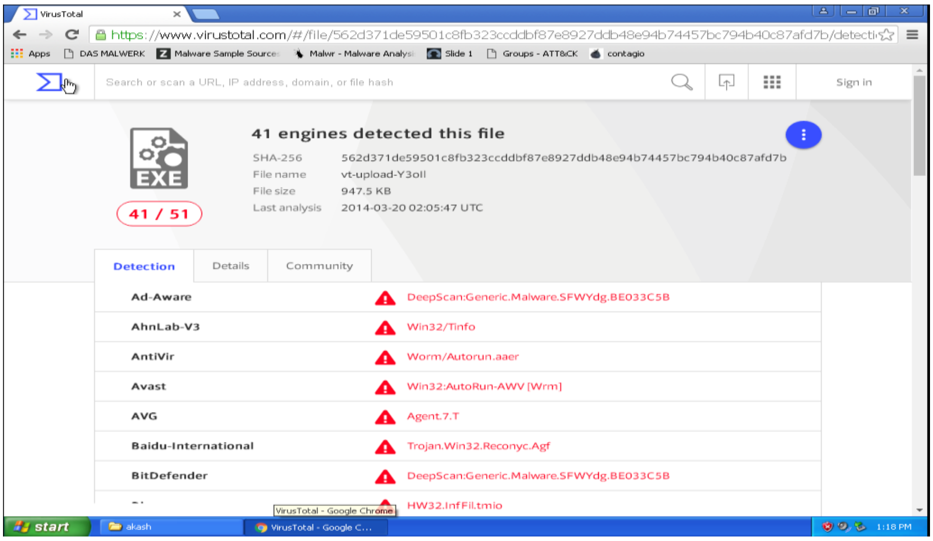
This is another malware whose labeled sample Figure 7. (Sample of 'NewFolder.exe' malware) is as follows.



```
MZ └      ┘      ÿÿ              @
                   ╔º╔  ´          Í! ¸    LÍ!This program
cannot be run in DOS mode.
$          ‡$qVEï"VEï"VEï"Á‹"TEï"qƒ"bEï"qƒ,"–Eï"•J
º"_Eï"•J²"MEï"VEî"€Dï"qƒ"oEï"qƒ'"sEï"qƒ“"WEï"qƒ
—"WEï"RichVEï"                                PE   L   |
ª>IG          à  #   ♂   ◘   R—   *┐        =>|   +    p—   @
 +    ┐   ┘          ┘                ♂   ┘    ¯¾!! ┐  €   @   +
+   +       +               ¼+•  ↑        ð◘  È
                                                        .text
 OP—   +    R—   ┘                   @   à.rdata   èá    p—
â    V—                   @   À.data    øŽ       `•    *    8•
              @   À.rsrc     È       ð◘         b•
@   @.rrdata   ð    +   ð    €◘        “               à



                      ¹ð·H èÔÒ   3À£ø•G £ô•G £à•G £Ü•G £Ø•G
£ð•G £ì•G £è•G £ä•G ¢¿•G ¢–ºH ¢ì·H ¢Á•G ¢¾•G ¢½•G
¢¼•G ¢»•G ¢º•G ¢¹•G ¢¸•G £Ü·H Ç|è·H └   Æ|À•G
  ¸Ø·H ÃU‹ìì0┘   SVW…àýÿÿP¿┘     WÆEþ ÆEÿ ÿ┴ìrF
ÿu◘èC└   èõÑ   …Àt├j┼hŽ¾F h¾F j ÿ┴ÜuF éÖ     ƒ=è·H
»ü·H u jÿÿ5ð·H ¾ºØH j   è³T    ½•G ¢²ØH ëNEþPhè·H
‹ó¸ºØH èÜZ   „Àu⋇Ç|Ô•G          és        ºØH ¢ì·H  ±ØH
hä·H ˆEÿ…ÐûÿÿPWSÿ┴ôrF ÿ5è·H Sèlz    …Àt&¿ºØH è♂V
…àýÿÿPÿ┴┘sF Ç|Ô•G          é$     €}ÿ   ⋇…¨    €=½·H
```

Figure 7. (Sample of 'NewFolder.exe' malware)

Another virus like NewFolder.exe places its original file somewhere else and the host file which calls the original file to be executed is placed in the memory like pen drive or any other location in the system. This virus gets executed in the background in hidden mode. Hence it is difficult to be traced. The system will slow down, if this virus is executed. Virustotal depicts out of 51malware engines, 41 malware engines shows the 'NewFolder.exe' is a malware which is shown in Figure 8. (showing NewFolder.exe is malicious, virustotal output).


VirusTotal
https://www.virustotal.com/#/file/562d371de59501c8fb323ccddbf87e8927ddb48e94b74457bc794b40c87afd7b/detecti
Apps
DAS MALWERK
Malware Sample Source:
Malwr - Malware Analysi
Slide 1
Groups - ATT&CK
contagio
Search or scan a URL, IP address, domain, or file hash
Sign in
EXE
41 / 51
Detection
Details
Community
VirusTotal - Google Chrome
start
akash
VirusTotal - Google C...
1:18 PM

41 engines detected this file

SHA-256 562d371de59501c8fb323ccddbf87e8927ddb48e94b74457bc794b40c87afd7b
File name vt-upload-Y3oll
File size 947.5 KB
Last analysis 2014-03-20 02:05:47 UTC

Ad-Aware	DeepScan:Generic.Malware.SFWYdg.BE033C5B
AhnLab-V3	Win32/Tinfo
AntiVir	Worm/Autorun.aaer
Avast	Win32:AutoRun-AWV [Wrm]
AVG	Agent.7.T
Baidu-International	Trojan.Win32.Reconyc.Agf
BitDefender	DeepScan:Generic.Malware.SFWYdg.BE033C5B
[illegible]	HW32.InfFil.tmio

Figure 8. (showing NewFolder.exe is malicious, virustotal output)

III. CONCLUSION

This paper is based on non-debugging and non-disassembling technique, which is quite easy for a normal user. With a little knowledge about the text strings observed in the editor (when a file is opened), a malware can be detected. If there is a little knowledge of API is there, then it is very much easy for a normal user. Different authors used static methodology but, they mean to discuss the malware before execution. Here the files are analyzed before execution. If at all the file is executed, then the user can analyze the malware. Here www.virustotal.com is used for malware confirmation. For the analysis, the user can use the common strings which are quite helpful for most of the cases. The examples given are real-time examples and based on the view of a normal user.

IV. REFERENCES

[1]. A.H.Sung, J. Xu, P.Chavez, S.Mukkamala, "Static Analyzer of Vicious Executables (SAVE)", Conference Paper, DOI: 10.1109/CSAC.2004.37, Source: IEEE Xplore, https://www.researchgate.net/publication/4115464 , 2005

[2]. Madhu K. Shankarapani, SubbuRamamoorthy, Ram S. Movva, SrinivasMukkamala, "Malware Detection using assembly and API call sequences", J ComputVirol (2011) 7:107-119, DOI 10.1007/s11416-010-0141-5, 2010

[3]. KarishmaPandey, MadhuraNaik, JunaidQamar , MahendraPatil.," Spyware Detection Using Data Mining", International Journal for Research in Applied Science &Engineering,Technology(IJRASET) Volume 3 Issue III, March 2015

[4]. Ankur Singh Bist, "Spyware Detection Techniques", INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY Bist, 3(2): February, 2014

[5]. Gerardo Canfora, Antonio NiccolòIannaccone, Corrado Aaron Visaggio, "Static analysis for the detection of metamorphic computer virusesusing repeated-instructions counting heuristics", J ComputVirol Hack Tech, DOI 10.1007/s11416-013-0189-0, 2013

[6]. Andreas Moser, Christopher Kruegel, EnginKirda, "Limits of Static Analysis for Malware Detection", 23rd Annual Computer security applications conference, 2007, http://rosaec.snu.ac.kr/meet/file/20090204paperc.pdf

[7]. Maryann Gong, Uma Girkar, Benjamin Xie, "Classifying Windows Malware with Static Analysis", https://courses.csail.mit.edu/6.857/2016/files/5.pdf

[8]. Norkhushaini Awang, Arifin Salleh, Mohamad Yusof Darus, "Manual Malware Analysis Using Static Method", International Journal of Computer Networks and Communications Security, 1(7), ISSN 2308-9830, pp. 324-328, 2013

[9]. Simon Kramer, Julian C.Bradfield, "A general definition of malware", DOI 10.1007/s11416-009-0137-1, J Comput Virol (2010) 6:105–114

[10]. B. Jaya Prasad, Haritha Annangi, Krishna Sastry Pendyala, "Basic Static Malware Analysis using open source tools", https://securitycommunity.tcs.com/infosecsoapbox/sites/default/files/Static%20Malware%20Analysis%20Techniques%20.pdf

[11]. Mohd. Ishrat, Manish Saxena, Dr. Mohd. Alamgir, "Comparison of Static and Dynamic Analysis for Runtime Monitoring", International Journal of Computer Science & Communication Networks, ISSN:2249-5789, Vol 2(5), 615-617

[12]. Ronghua Tian, "An Integrated Malware Detection and Classification System", Ph.D. Thesis, Deakin University, 2011

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

An Intrusion Detection System for MANETS

Insha Majeed, Sakshi Arora
Department of Computer Science Engineering Shri Mata Vaishno Devi University Katra, Jammu and Kashmir, India

ABSTRACT

Mobile Ad hoc Networks (MANETs) are very important and unique applications as they does not require a fixed network infrastructure and moreover both works of transmitter and a receiver are done by single node. Nodes are able to communicate directly with each other or through their neighbours to relay messages. MANETs can be envisioned for military and emergency communication due to their capability of cooperative routing but are also more vulnerable to routing attacks than wired networks. Thus it is imperative to develop an efficient intrusion-detection mechanisms to protect MANETs. To detect new routing attacks in MANETs we have applied specification based intrusion detection approach that defines normal behaviour of the protected networks. In this paper, we propose a complete distributed intrusion detection system that consists of four models for MANETs with formal reasoning and simulation experiments for evaluation. Optimal Link State Routing (OLSR) is representative of a proactive, link-state routing protocol of MANETs, and the Ad Hoc On Demand Distance Vector (AODV) is the other popular, reactive, request-on- demand routing protocol. Both OLSR and AODV have IETF RFCs.

Keywords : AODV, Topology Control, RREP, RERR, hop, OLSR, DEMEM, DRETA.

I. INTRODUCTION

A. Mobile Adhoc Network

MANET is a set of wireless mobile nodes with no pre-established infrastructure. They can be applied to various popular wireless technologies including cellular phone services, disaster relief, emergency services, battlefield scenarios, and other applications because MANETs have mobile nodes with reliable routing services. Moreover, MANETs are decentralized networks with unpredictable network topology because of node mobility. Because of decentralisation nature all mobile nodes need to discover the dynamic topology and deliver messages by themselves and mobile nodes in MANETs act as both hosts and routers. The mobile nodes set up the routing tables by exchanging routing messages with each other and then deliver data packets for others. The most fundamental and critical issue related to MANETs is to develop a system to maintain routing tables reliably. The dynamic nature and cooperativeness of MANETs presents extensive challenges for network security.

B. Vulnerabilities in MANET

Because of the characteristics discussed above network-based access control mechanisms such as firewalls cannot be directly implemented as they do not have well defined boundary and make MANETs more vulnerable than normal wireless networks with base stations. A MANET is trust-all-peers design assuming every node to provide accurate routing information and acts as a router to cooperatively forward packets. Ad hoc networks are vulnerable to several routing attacks: address spoofing, black hole, man-in- the-middle, modification of packets, and distributed denial-of- service (DDoS).

C. Cryptographic Approaches in Securing MANET

Largely research about securing MANET routing protocols use cryptographic approaches based on public key infrastructure (PKI). For example, ARAN[37] and SAODV[42] apply PKI on AODV to generate digital signature to protect the integrity of its routing messages. A secure OLSR protocol also uses digital signature to guard OLSR routing messages. PKI is also applied to protect other routing protocols of MANETs. The cryptographic approaches do not cover up critical fields, for example hop count, the value of which will change over time. Another limitation is that they cannot prevent insider attacks. Therefore, some other mechanism has to be developed to set off the limitations of cryptographic approaches, and develop the Intrusion Detection to secure MANET routing.

D. New challenges of IDS in MANET

MANET with such features and limitations poses difficult challenges in developing IDS for MANET as compared in wired networks. First, the characteristic of nodes tobe honest and cooperative can be advantageous for malicious node to launch many routing attacks. Second, thedistributed network without centralized admin in MANET, the IDS will not detect the routing attacks if all distributed detectors does not have monitoring information from others. Third, the detectors will require up-to- date evidence in real-time to detect the attacks with low false positive and negative rates but is difficult due to highly dynamic and unpredictable mobility. The attacks can propagate and paralyze the network quickly due to the lack of trust management between nodes.

E. Contributions

Our proposed Intrusion Detection System (IDS) detects insider attacks on MANET routing protocols, and it is the first complete IDS with several innovative models including two intrusion detection models, a message exchange model, and an authentication model. This effectual ID system can overcome the challenges discussed above and adapt to the unique MANET environment with low message and computation overhead. We apply a specification based detection approach to correctly detect routing attacks that contravene detection constraints. The constraints define the normal behaviours of the target routing protocol and violations of these constraints are potential routing attacks. The distributed detectors use these constraints to detect corrupt routing messages, causing the violations, and then correct the corrupt message contents to stop the attacks. MANET nodes generally have less computational power and bandwidth. MANET is very susceptible to message overhead generated by IDS. First, we put forward two specification-based intrusion detection models for two representative routing protocols: OLSR and AODV. Then we propose distributed intrusion detection models to improve the first two models and make them practical and scalable.

II. RELATED WORKS

A. Introduction

Security mechanisms, like authentication services and access control cannot alone deter all possible attacks such as insider attacker. Therefore we need the security mechanisms which can deal with bad insider nodes possessing valid key and access rights. Intrusion detection provides second line of defense.The routing works for MANETs can be characterized as: first category is based on authentication-based approaches such as Authenticated Routing for Ad-Hoc Networks (ARAN)[37], Ariadne[17], and Secure AODV[42]. The work in second category are IDSes targeted at mobile ad hoc networks like MANET IDS framework, statistical anomaly based IDS for detecting insider attacks, and security analysis for selected protocols. The last category is malicious packet dropping detection. MANET needs to detect selfish nodes and enforce cooperative participation as routing services require cooperation among all nodes. In this category, either statistical or reputation-based approaches are used.

B. Authentication Approaches

The cryptographic approaches [17][2][42][37][31][41] proposes authentication protocols for controlling the routing data message exchange in several protocols. Authenticated Routing for Ad-Hoc Networks (ARAN)[37] assume that each node has its own public and private key distributed by a trusted server. Zapata and Asokan[42] proposed Secure AODV, which uses asymmetric cryptography to secure the AODV routing protocol. To prevent replay attacks Adjih[4] proposed a secure OLSR protocol which uses a signed time stamp to validate the freshness of a message. Papadimitratos and Haas has proposed a secure link state routing for mobile ad hoc networks[31]. These works use public key based signatures to keep the header readable and to protect routing message header from being modified. Adrian Perrig proposed TESLA[2] which is a symmetric key based broadcast authentication protocol. Hu proposed Ariadne[17], more secure version of Dynamic Source Routing (DSR), which applied TESLA to reduce computation overhead.

C. IDS Approaches in Mobile Ad-hoc Net-works

Distributed cooperative IDSs are proposed for MANETs to determine the lack of central authority. Zhang and Lee[43] proposed first integrated IDS architecture in 2000. For statistical anomaly detection Huang and Lee[6] in 2003, presented a cooperative cluster-based architecture. Sterne [11] presented a cooperative intrusion detection architecture for addressing the challenges in MANET. For misuse detection within MANETs Subhadhrabandhu[12] evaluated several selection strategies for placement of IDS modles. Ramanujan[34] presented a system which can detect, avoid, as well as recover from malicious attacks. Based on mobile agent technology Kachirski and Guha[3] describes a wireless IDS. An IDS approach based on a stateful analysis of AODV control packet streams was employed by Gwalami[14]. This approach applies State Transition Analysis Technique (STAT) [19], Peng Ning and Kun Sun[28] presented an examination of insider attacks in the AODV protocol.

Several IDS approaches are proposed for detecting malicious packet dropping for MANETs (i.e. both routing and data packets). [8][9][26][27] used the method of assigning a value to the “reputation” of a node and using this information to hoe out misbehaving nodes and use only trusted and verifiably fine nodes. [7] and [36] are credit and statistics-based approaches to solve packet dropping problems in MANET respectively. To find out whether nodes are not forwarding packets at the desired rate because of congestion or because of malicious behaviour a statistical approach is presented by Rao and Kesidis in [35] using estimated congestion at intermediate nodes.

D. Specification based Intrusion Detection Systems

In wired network Intrusion detection systems have imployed two models: anomaly based and signature based approaches. A signature-based IDS[19][25] can monitor activities on the network and then compares those with known attacks. An anomaly based IDS [6][7][36][43][5] can monitor the network traffic and compare it with normal behaviour patterns statistically. A new approach ideal for new environments, such as MANETs is the specification based approach. A specification based IDS detects attacks (including known and unknown) according to normal behavors of protected services, such as routing services in MANETs. To do this, the IDS first analyzes the protected protocol specification, and introduces vulnerabilities of the protocol, including useful descriptions of exploits in the protocol. Second, the IDS provides detection rules to enforce the protocol normal behavior. Since malicious nodes have limited known attack methods to take advantage of protocol vulnerabilities, their attacks will cause the violations of the rules and be detected. Thus, the specification based IDS can achieve much lower false positives and negative than those by an anomaly based IDS. And the specification based IDS is able to detect new or unknown attacks which a signature based cannot detect.

III. A SPECIFICATION-BASED INTRUSION DETECTION MODEL FOR AODV

A.Introduction

AODV, a reflex and stateless routing protocol, builds up routes when wanted by the source node is helpless against different sorts of attacks [28]. The determination based intrusion detection procedure is proposed to identify attacks on AODV. The approach utilizes the limited state machines to determine AODV routing conduct and distributed network screens for recognizing run-time infringement of the details. One extra field called sequence number in the protocol message is proposed to empower the checking. In our calculation, we utilize a tree information structure and a node shading which successfully recognizes the genuine attacks continuously, with least overhead.

B. Overview of AODV

AODV builds up routes on demand by a source node utilizing Route Request (RREQ) and Route Reply (RREP) messages. Route Request (RREQ) message has RREQ ID (RID) that is broadcast by a node to discover a route to its destination . Turn around route to the source node in the routing tables is set up and sequence number is refreshed by a node when it gets RREQ message. A route answer (RREP) is unicast back to the source node when the node is the destination or the node has a route to the destination that meet the freshness necessity. The sequence number (SN) in AODV tells about freshness of the routing data and furthermore ensures circle free routes. Sequence number can be expanded just under two conditions: 1.when RREQ is broadcast by the source node and 2. the destination node answer with a RREP. The hop count (HC) is incremented by 1when a message(RREQ or RREP) is forwarded each hop. Route blunder packets (RERR) are spread to the start node along the turn around route when a link is broken, and all other nodes will drop the section in their routing tables. Figure 3.1 shows how Aodv works.The estimations of the fields in the routing messages are signified in Table3.1.

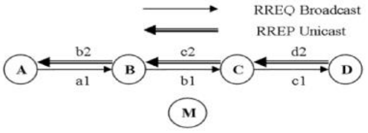
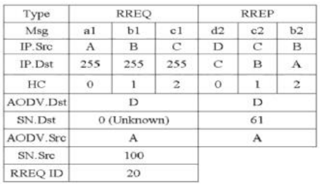


Figure 3.1. AODV state

Table 1. RREQ and RREP values

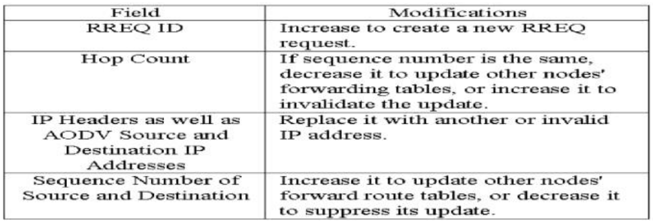


Type	RREQ			RREP		
Msg	a1	b1	c1	d2	c2	b2
IP.Src	A	B	C	D	C	B
IP.Dst	255	255	255	C	B	A
HC	0	1	2	0	1	2
AODV.Dst	D			D		
SN.Dst	0 (Unknown)			61		
AODV.Src	A			A		
SN.Src	100					
RREQ ID	20					

C. The vulnerable Fields for AODV Control Messages

AODV is proficient and adaptable as far as network performance, yet it enables attackers to effortlessly publicize misrepresented route data to divert routes and to dispatch different sorts of attacks. In each AODV routing packet, some basic fields, for example, hop count, sequence numbers of source and destination, RREQ ID, IP headers and additionally IP locations of AODV source and destination,are fundamental to redress protocol execution. Any abuse of these fields can make AODV malfunction. Table 2 shows defenseless fields in AODV routing messages and the impacts when they are altered.

Table 2.Aodv fields



Field	Modifications
RREQ ID	Increase to create a new RREQ request.
Hop Count	If sequence number is the same, decrease it to update other nodes' forwarding tables, or increase it to invalidate the update.
IP Headers as well as AODV Source and Destination IP Addresses	Replace it with another or invalid IP address.
Sequence Number of Source and Destination	Increase it to update other nodes' forward route tables, or decrease it to suppress its update.

Some of the attacks are underneath:

- ✓ Single Attacks are:Forging Sequence Number,Forging Hop Count
- ✓ Examples of Aggregated Attacks

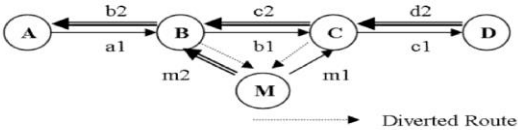


Figure 3.2. Man in the Middle Attack

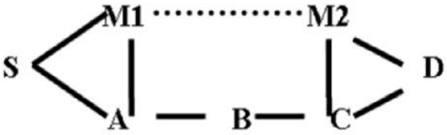


Figure 3.3. Tunnelling Attack

D. Specification-based Monitoring of AODV

Specification-basd observing contrasts the conduct of items and their related security specifications that catch the right conduct. Specification-based detection does not recognize intrusions specifically - it identifies the impact of the intrusions as run-time infringement of the specifications. The specification-based detection approach has been effectively connected to screen security-basic projects [23], applications, and protocols[22].

1. Assumptions

We utilized the accompanying suppositions:

1. MAC locations and IP locations of all versatile nodes are enrolled and authenticated with the network screens.

2. Network screens can cover all nodes and play out all required functionality.

3. Every network screen are dependable and can simply convey safely and dependably.

4. Every node neither drops AODV messages nor sticking remote channels.

2. The Finite state Machine Constraints

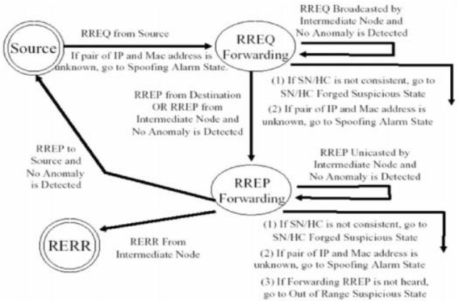

Source
RREQ from Source
RREQ Forwarding
If pair of IP and Mac address is unknown, go to Spoofing Alarm State.
RREQ Broadcasted by Intermediate Node and No Anomaly is Detected
(1) If SN/HC is not consistent, go to SN/HC Forged Suspicious State
(2) If pair of IP and Mac address is unknown, go to Spoofing Alarm State
RREP from Destination OR RREP from Intermediate Node and No Anomaly is Detected
RREP to Source and No Anomaly is Detected
RREP Forwarding
RREP Unicasted by Intermediate Node and No Anomaly is Detected
(1) If SN/HC is not consistent, go to SN/HC Forged Suspicious State
(2) If pair of IP and Mac address is unknown, go to Spoofing Alarm State
(3) If Forwarding RREP is not heard, go to Out of Range Suspicious State
RERR
RERR From Intermediate Node

Figure 3.4. Normal State Diagram

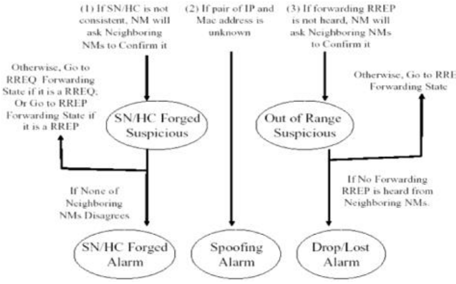

(1) If SN/HC is not consistent, NM will ask Neighboring NMs to Confirm it
(2) If pair of IP and Mac address is unknown
(3) If forwarding RREP is not heard, NM will ask Neighboring NMs to Confirm it
Otherwise, Go to RREQ Forwarding State if it is a RREQ. Or Go to RREP Forwarding State if it is a RREP
SN/HC Forged Suspicious
Out of Range Suspicious
Otherwise, Go to RREP Forwarding State
If None of Neighboring NMs Disagrees
If No Forwarding RREP is heard from Neighboring NMs.
SN/HC Forged Alarm
Spoofing Alarm
Drop/Lost Alarm

Figure 3.5. Alarm and Suspicious State Diagram

IV. A SPECIFICATION-BASED INTRUSION DETECTION MODEL FOR OLSR

Optimized Link State Routing (OLSR) is a proactive table-driven routing protocol created by INRIA [10]. The protocol is a refinement of conventional link state protocols utilized in wired networks; in the last mentioned, the neighborhood link state data is spread inside the network utilizing broadcast strategies. This flooding impact will devour impressive data transmission if specifically utilized in the MANET area, and in this way, OLSR is intended to ideally disperse the nearby link state data around the network utilizing a progressively settled sub-network of multipoint transfer (MPR) nodes; these are chosen from the current network of nodes in the MANET by the protocol. OLSR utilizes two principle control messages: Hello messages and Topology Control (TC) messages to disperse link state data. These messages are intermittently broadcast in the MANET keeping in mind the end goal to build up the routing tables at each node freely. In OLSR, just nodes that have bidirectional (symmetric) links between them can be neighbors. Hi messages contain neighbor records to permit nodes to exchange neighbor data, and set up their 1-hop and 2-hop neighbor records; these are utilized to compute multi-point hand-off (MPR) sets.

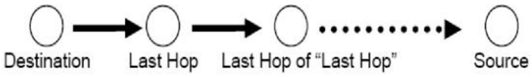
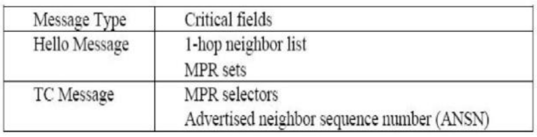


Figure 4.1. A route from Topology Table

Table 4.1. Hello and TC Messages' Critical fields



Message Type	Critical fields
Hello Message	1-hop neighbor list MPR sets
TC Message	MPR selectors Advertised neighbor sequence number (ANSN)

A. OLSR Vulnerabilities and Attacks

A few examinations have been done on the vulnerabilities of OLSR [15][4]. When all is said in done, an attacker can manufacture packets, block and adjust packets experiencing it, or decline to forward packets, causing bargains of confidentiality, integrity, and avaixlability. In this work, we just concentrate on those vulnerabilities that could trade off the integrity of the network, i.e., the routing tables in the nodes. In OLSR, each node infuses topological data into the network through HELLO messages and TC messages. In this way, a malevolent node can infuse invalid HELLO and TC messages to disturb the network integrity, making packets route inaccurately or to the benefit of the attacks.

Table 4.1 showcases the basic fields in the TC message and the Hello message on which the calculation of the routing table depend. The 1-hop neighbour list in Hello message is utilized by its neighbour to make the 2-hop neighbor rundown an MPR set. The sender's MPR set is MPR sets in hello message. The MPR selector in TC message is utilized as a part of figuring routing tables at nodes accepting the messages.

In this manner, an attack can:

1) provide an off base 1-hop neighbor list in a Hello message
2) provide an off base MPR set in a Hello message
3) provide off base MPR selectors in a TC message
4) modify the MPR selectors and ANSN before it forwards a TC message

B. Intrusion Detection Model

Here we portray our specification-based way to deal with recognizing OLSR attacks. Specification-based detection is especially reasonable for identifying attacks on network protocols on the grounds that the right conduct of a protocol is generally very much characterized and is archived in the protocols specification. Utilizing network observing, test is to separate a reasonable modal of conduct from the protocol specification; can be checked at runtime. We initially list suppositions utilized, & after that presents the right conduct modal of OLSR under these presumptions.

C. Assumptions

We expect a distributed ID design which permits helpful detectors to wantonly screen the Hello and TC messages, and also exchange their neighborhood information when important. IDS detectors in this design can screen all Hello and TC messages sent by each node of the network, dependably exchange IDS information effectively, and won't be traded off. We accept OLSR is the main routing protocol in the network and each node has just a single network interface.

D. Correct Behavior Model of OLSR

Figure 4.2 demonstrates the FSA model of the OLSR protocol that characterizes the right operation of an OLSR node in dealing with control activity. At this point when node gets a Hello control message, it will refresh its neighbors rundown and MPR set. A node refreshes the topology and routing table when it accepts the Topology Control message. What's more, the node will forward the TC on the off chance that it is an MPR node. Also, a node will occasionally broadcast Hello and TC messages.

We depict the requirements on the control activity between neighbor nodes for identifying irregularities inside the control messages.

1. Neighbor’s record in Hello message must be complementary. E.g., if node 2 is the neighbor of node 1, at that point node 1 must be node 2's neighbor.
2. The MPR nodes of a nodeshould achieve each of the 2-hop neighbors of the node & the MPR nodes are to transmit TC messages intermittently.
3. MPR sets of Hello messages must match relating MPR selectors of a TC message. E.g., if node 2 is node 1's MPR selector, node 1 must be node 2's MPR.
4. Forwarded TC message’s Integrity must be kept up.

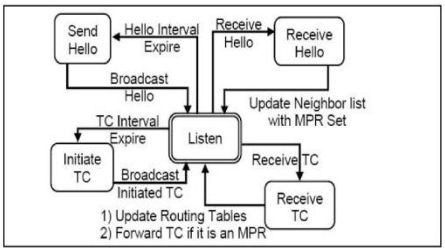


Figure 4.2. Routing Finite State Automata (FSA) for OLSR

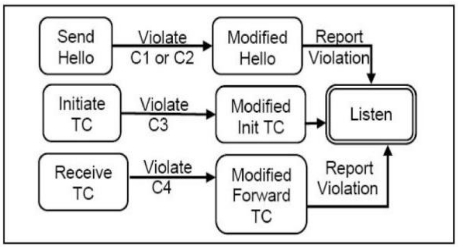


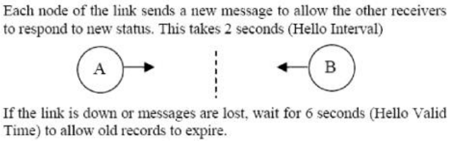
Figure 4.3. Finite State Automata for Security Specification

At the point when an OLSR control messages abuse any of the limitations, the FSA goesfrom an ordinary state to some caution states (Modified Init TC State, Modified Forward TC State, Modified Hello State).

Figure 4.3 (an expansion of FSA in Figure 4.2) delineates FSA utilized by the specification-based intrusion detection system.

E. Temporary Inconsistency

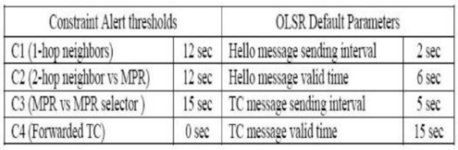
When the topology changes links are made or evacuated so temporary violation of imperatives C1, C2 and C3 may happen in a brief timeframe. To maintain a strategic distance from false cautions, the detector is tosit tight for the two nodes on the two sides of a link to take in the new link status before declaring an irregularity as attack. Also, when a link between node An and node B is made, node A refreshs the status of A-B link and sends Hello message, isn't predictable with past Hello message of node B. Again the detector is to sit tight for B to get new Hello message from An and send another Hello message mirroring the expansion of link A-B.


Each node of the link sends a new message to allow the other receivers to respond to new status. This takes 2 seconds (Hello Interval)

If the link is down or messages are lost, wait for 6 seconds (Hello Valid Time) to allow old records to expire.

Figure 4.4. Resolving temporary inconsistency

Table 4.2. Important Parameters for Temporary Inconsistency



Constraint Alert thresholds		OLSR Default Parameters	
C1 (1-hop neighbors)	12 sec	Hello message sending interval	2 sec
C2 (2-hop neighbor vs MPR)	12 sec	Hello message valid time	6 sec
C3 (MPR vs MPR selector)	15 sec	TC message sending interval	5 sec
C4 (Forwarded TC)	0 sec	TC message valid time	15 sec

F. Limitations

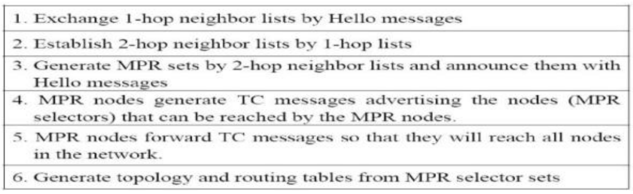
For a solitary attack or non-related attacks, the model can identify all attacks since we catch all conceivable approaches to change a solitary message at once. Be that as it may, if at least two attackers dispatch an associated attack in which inaccurate data is provided to various nodes reliably, the limitations will be unable to distinguish it. For instance,when attackers claim to be neighbours but are not there might not be distinguishable infringement.

G. OLSR Detection Model's Analysis

Here we dissect the OLSR protocol & the proposed detection model to demonstrate that the arrangement of limitations C1 — C4 can distinguish attacks in MANET.

Table 4.3 depicts the procedure for building up the routing table from the point of view of a node.

Table 4.3. OLSR Routing Table Establishment



1. Exchange 1-hop neighbor lists by Hello messages
2. Establish 2-hop neighbor lists by 1-hop lists
3. Generate MPR sets by 2-hop neighbor lists and announce them with Hello messages
4. MPR nodes generate TC messages advertising the nodes (MPR selectors) that can be reached by the MPR nodes.
5. MPR nodes forward TC messages so that they will reach all nodes in the network.
6. Generate topology and routing tables from MPR selector sets

As indicated by the RFC [10] (OLSR protocol), every node keeps up a topology a link set, utilized for figuring of route table. The link set has the link data of its 1-hop neighbor, developed from the Hello messages it gets. The topology has topology tuples as T[HoldingTime], T[DestAddr], T[LastHopAddr], T[Seq], which demonstrate that we can achieve T[DestAddr] through T[LastHopAddr]. A topology set is built from TC messages a node gets. The node processes the route table from its topology and link set.

Lemma 1: Under suppositions in D, all great nodes will have a right link set if imperative C1 holds.

Lemma 2: The MPR selector field of a TC message created by a MPR node must be right if limitation C3 holds.

Lemma 3: The MPR selector fields of all TC messages must be right if requirements C3 and C! Hold.

Lemma 4: For a node x, which is a n-hop neighbor of an alternate node y, x will get TC messages of y with n-1 forwarding if C2 holds.

V. DEMEM: DISTRIBUTED EVIDENCE-DRIVEN MESSAGE EXCHANGE INTRUSION DETECTION MODEL FOR MANET

A. Introduction

In this segment, we make two noteworthy commitments for intrusion detection systems (IDS) in MANET. To start with, we propose a handy and viable message exchange model: Distributed Evidence-driven Message Exchanging intrusion detection Model (DEMEM) for MANET.

DEMEM beats the difficulties to distributed IDS engineering of MANET(as depicted in segment 3 and 4), where detectors don't have adequate information to identify routing attacks. Rather than receiving expensive wanton observing, detectors in DEMEM just capture routing messages and approve these routing messages keeping in mind the end goal to identify routing attacks. Additionally, DEMEM isolates the obligations of security specialists and routing administrations to abstain from changing the routing protocols. Second, we coordinate DEMEM into a proactive routing protocol in MANET, OLSR (Optimal Link State Routing) [10]. DEMEM in OLSR utilizes detection requirements talked about in segment 4 [39]. The detection model demonstrates that by approving consistency among related routing messages as indicated by these detection imperatives, detectors can definitely distinguish both known and obscure routing attacks in OLSR. Three ID messages for DEMEM in OLSR are proposedto give the basic ID message exchange benefit, which is the fundamental suspicions of past detection models in area 3 and 4. ID-Evidence messages ensure each detector has adequate evidence for identifying infringement of limitations; ID-Forward messages trigger the chose forwarders sending ID-Evidence messages while the detector watches new evidence so as to limit message overhead, and ID-Request handles message misfortune. Along these lines, DEMEM not just performs down to earth, scalaxble, and exact intrusion detection in OLSR yet additionally endures message misfortune with low message overhead.

B. Distributed Evidence-driven Message Exchange intrusion detection Model

DEMEM is a strong, adaptable, and low message exchange overhead intrusion detection model for MANET. DEMEM beats the difficulties said in area A through the accompanying three principle includes: a distributed engineering, an intrusion detection layer, and an evidence-driven message exchange system.

C. Distributed IDS Architecture

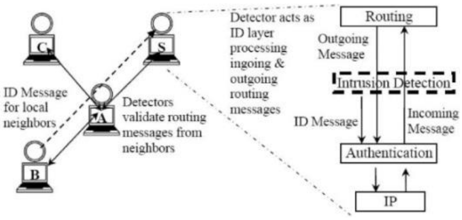

C
S
A
B
ID Message for local neighbors
Detectors validate routing messages from neighbors
Detector acts as ID layer processing ingoing & outgoing routing messages
Routing
Outgoing Message
Intrusion Detection
ID Message
Incoming Message
Authentication
IP

Figure 5.1. Architecture of DEMEM

DEMEM gets adjusted to the distributed and helpful behavior of MANETs. Each DEMEM node goes about as a detector screens its 1-hop neighbors by approving route messages it gets for intrusion detection purposes. So that, when a node sends a routing message, the majority of its neighbours approve the accuracy of the message. As found in Figure 5.2, node A is detector and screens nodes B, C, and S and similarly nodes B, C, and S are detectors that screen other nodes. Notwithstanding checking exercises inbetween 1-hop neighbors, 2-hop neighbors will need to exchange their watched data by customized ID (Intrusion Detection) messages to accumulate enough evidence for detection purposes. This approach takes out confounded topology upkeep and costly questionable want on observing required by progressive agreeable intruzion detection [43].

D.DEMEM in OLSR

1.Routing Attack Methods in OLSR

A proactive routing protocol (OLSR)uses periodic Hello and Topology Control (TC) messages to set up a total network topology. OLSR gives a hearty and finish routing topology and endures message misfortune caused by portability and clamor.

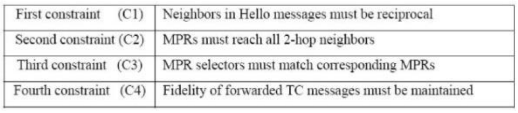
OLSR, the calculation of routing tables relies upon three basic fields in Hello and TC messages: 1-hop neighbors and MPRs in Hello message and also MPR selectors in TC messages. A node can send three sorts of essential OLSR messages: Hello, started TC, and forward TC messages. In this way, an attacker has four attack techniques against OLSR routing:

1. Forging 1-hop neighbors in a started Hello;
2. Forging MPRs in a started Hello;
3. Forging MPR selectors in a started TC; and
4. Forging MPR selectors in a forwarded TC.

The initial three attack strategies have a place with the main sort of attack model, and the fourth one has a place with the second kind of attack model.

2. Specification-based Intrusion Detection

In MANET, nodes sharing incomplete topology data and covered topology data from their routing packets must be steady. Despite the fact that it is hard to identify attacks propelled by forging started routing packets, substance of these fashioned packets won't be predictable with veritable routing packets that have covering routing data. In this way, the detector can recognize these fashioned packets by approving consistency among related routing messages. The specification-based intrusion detection model [39] in area 3 portrays four requirements (see Figure 5.2) to approve the rightness of Hello and TC messages in OLSR.



First constraint (C1)	Neighbors in Hello messages must be reciprocal
Second constraint (C2)	MPRs must reach all 2-hop neighbors
Third constraint (C3)	MPR selectors must match corresponding MPRs
Fourth constraint (C4)	Fidelity of forwarded TC messages must be maintained

Figure 5.2. 4 detection constraints

DEMEM helps the model [39] resolve this assumption with a practical message exchange technique.

3. Implementing DEMEM in OLSR

To make the model in segment 4 down to earth and successful, three Intrusion Detection (ID) messages are made for OLSR i.e, ID-Evidence, ID-Forward, and ID-Request messages. We additionally exhibit the components for taking care of three ID messages, particularly inside the Evidence Manager and the Forwarding Manager.

E.DEMEM FSM for OLSR

In OLSR, the Evidence Manager handles ID-Evidence, Hello and TC messages and records evidence in these messages. A Forwarding Manager can send three ID messages under three conditions appeared in Figure 5.4. The Validation Manager approves Hello and TC messages in view of the three limitations & related evidence from the Evidence Manager. In the event that the Validation Manager recognizes message irregularities that damage these requirements and the enduring time of irregularities surpasses the caution limits of the limitations, the Response Manager will perform legitimate attack recuperation.

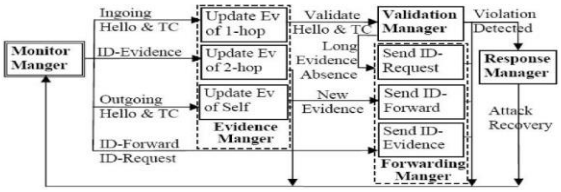


Figure 5.3. FSM within a DEMEM detector

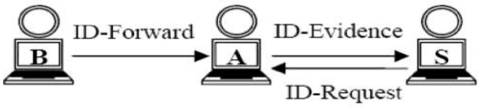


Figure 5.4. ID messages

Forwarding Manager: When the Validation Manager doesn't have adequate evidence fron a normal ID-Evidence message, it accept that the message is lost. The Validation Manager triggers the Forwarding manager to broadcast an ID-Request message to ask for the lost ID-Evidence message. Also the Forwarding Manager broadcasts an ID-Forward message when new evidence is sensed by Evidence manager in Hello message. And the Forwarding Manager broadcasts an ID-Evidence message for the neighbor when it gets message from the neighbors.

4 commonsense presumptions in light of existing works:

1. OLSR is the routing protocol and each node has one network interface. Various Message Interface Declaration (MID) and Host and Network Asociation (HNA) messages are not utilized here.
2. The substance of forwarded routing messages and the node personality in all routing and ID messages are authenticated by DRETA in part 6. In this manner, Constraint 4 of every 2 used to distinguish attack technique 4 out of 5.4.1 is secured here.
3. No deliberate packet dropping. A few trustworthy strategies [7][36] have been created for distinguishing ordinary unicast information packet drop attacks and also to broadcasting routing messages. We expect that detectors have been used to recognize purposefully packet dropping. DEMEM can likewise endure ordinary packet misfortune or drop.
4. No plotting attackers. Plotting attacks can make virtual links to perform worm-opening attacks. A few works [17] address this kind of attack. Likewise, included virtual links don't influence the presence of other ordinary routing links.

VI. DISTRIBUTED ROUTING EVIDENCE TRACING AND AUTHENTICATION INTRUSION DETECTION MODEL FOR MANET (DRETA)

A. Introduction

We propose the utilization of the DRETA (Distributed Routing Message Tracing and Authentication intrusion detection) model to give an effective and low-overhead insurance. DRETA has a distributed engineering, a detector in every node to screen and approve route messages. Isolated from the network layer, DRETA has a free layer, to block routing messages. Symmetric keys requiring much lower calculation overhead than public keys are utilised by DRETA, to give authentication administrations to all routing message. DRETA receives one-way key chain[16] and delay key disclosure[2] procedures for symmetric keys to be distributed in public channels like Public Key Infrastructure (PKI) does. Validation Messages (VMs), which utilize HMAC[24] are proposed for the integrity of forwarded messages.

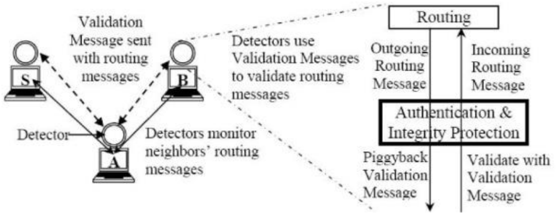


Figure 6.1. Validation Messages and Distributed detectors used to validate routing messages

We have implemented DRETA on two representative routing protocols i.e, OLSR and AODV. DRETA can protect the forwarded TC messages in OLSR and forwarded ID-Evidence messages in DEMEM.

B. Background

In section 2 and 3, we have introduced the Optimized Link State Routing protocol (OLSR)[10] and Ad-hoc On-demand Distance Vector routing protocol (AODV)[32] two representatives of proactive and on-demand routing protocols of MANET. We introduce one-way key chain[16] and delay key disclosure[2] techniques, which are adopted by DRETA.

2. Finite State Machine for DRETA

R'Msg: Forwarded Outgoing Routing Message(Sender is Not Originator)Ro Msg: Originated Outgoing Routing Message(Sender is Originator)

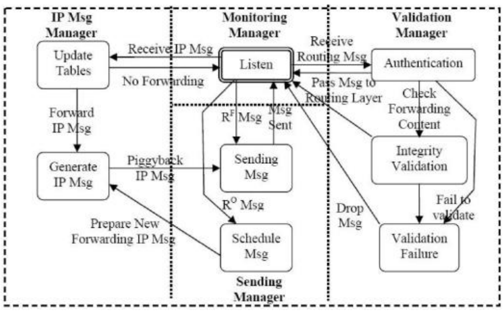


Figure 6.2. Finite State Machine within a node with DRETA

DRETA Implementations

Here we talk about executions of DRETA with AODV, OLSR, and DEMEM.

C. DRETA in AODV

DRETA have two security limitations to keep AODV messages from being malevolently changed when they are made (i.e, the msg originator is the noxious node). To begin with, DRETA never permit a middle of the road node to answer to a RREP on the grounds that the halfway node isn't the orignator of RREP. Likewise, the route information of destination in the middle of the road node might be obsolete, and it is troublesome & costly to approve the routng information. Subsequently, it is substantially more secure to just enable the destination to answer a RREP. Second, the nodes overlook Sequence Numbers of the destination in RREQ and RERR on the grounds that the no. may likewise be obsolete and along these lines not reliable.

In case the originator gives wrong data in its AODV message, mistaken data purposes routing harm to the originator itself. If the originator builds its SN an extensive amount, it won't influence AODV operation. Accordingly, attackers can't profit by malicious started AODV messages. The DRETA can thus secure forwarded AODV messages and also authenticated all AODV messages, DRETA effectively ensures the integrity of Aodv message.

D. DRETA in OLSR

OLSR has two primary routing messages, non-forwarded. DRETA in OLSR gives authentication to all messages:Hello messages and TC messages, and gives forwarded message insurance to TC message. In OLSR, just MPR nodes forward TC messages, so ETM (Evidence Tracing Messages) and KFM (Key Forwarding Messages) are just forwarded by MPR nodes. Subsequently, DRETA secures the forwarded TC messages in OLSR and averts attacks utilizing attack technique 4 in Figure 6.1.

E. DRETA in DEMEM

DEMEM forestalls attacks utilizing one of the initial three attack strategies in Figure 6.1. DRETA authenticates the three ID messages(ID-Evidence, ID-Forward, and ID-Request). Since the ID-Evidence message is a forwarded message, DRETA secures the integrity of the ID-Evidence message. In this way, DRETA and DEMEM agreeably guarantee the integrity of the routing messages in OLSR.

F. Experiment

We executed DRETA in GloMoSim, a reenactment intended for MANETs.

1. Experiment condition

GloMoSim bolsters 802.11, different routing protocols in MANETs, (for example, AODV and OLSR), and Ground Reflection (Two-Ray) radio model.DRETA utilizes the SHA-1 hash function to produce MACs and HMACs. The hash value estimate is 10 bytes and the key measure is 8 bytes. The key terminate time is 1 second.

2. Performance Metrics

We characterize three performance measurements to gauge DRETA's overhead:

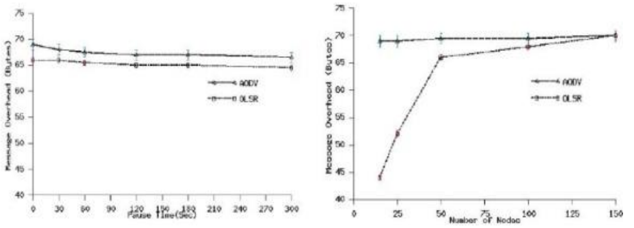


Figure 6.3. Message Overhead

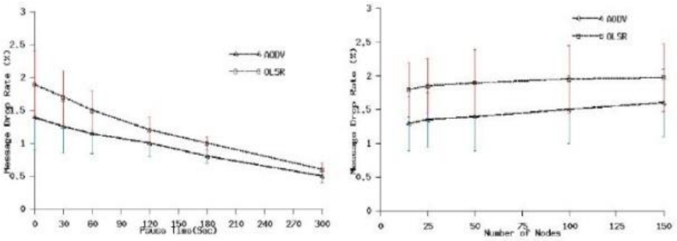


Figure 6.4. Routing message delay

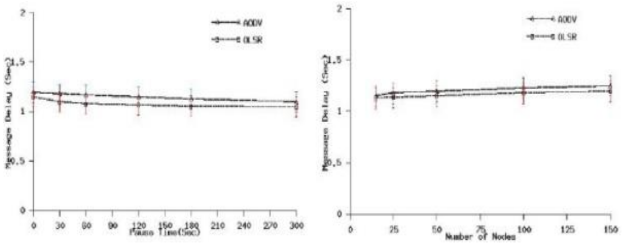
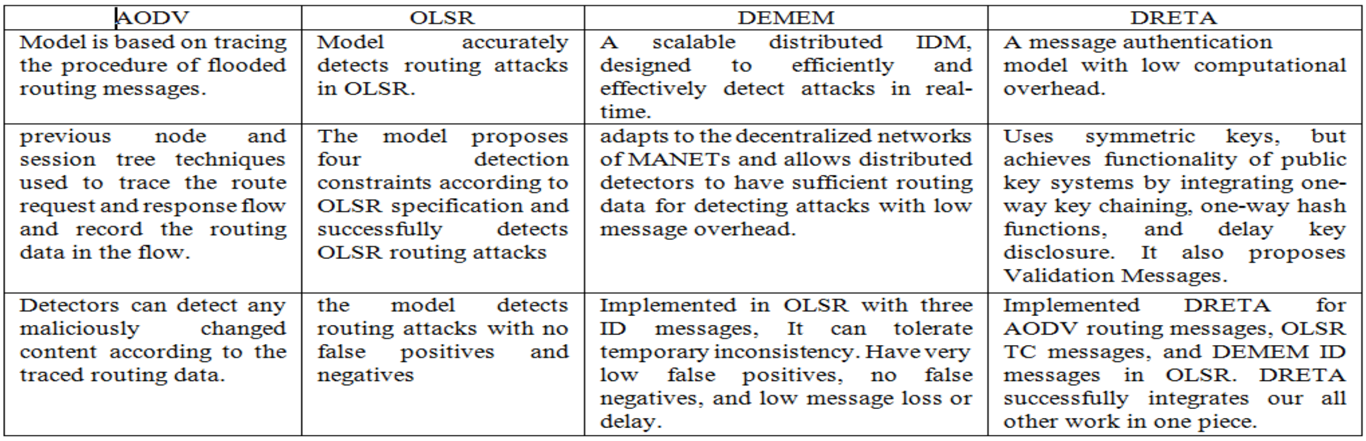


Figure 6.5. Detection accuracy

VII. CONCLUSIONS

In this paper, we developed four intrusion detection models which can be integrated with each other to become a complete intrusion detection system for MANETs.

Table 7.1



AODV	OLSR	DEMEM	DRETA
Model is based on tracing the procedure of flooded routing messages.	Model accurately detects routing attacks in OLSR.	A scalable distributed IDM, designed to efficiently and effectively detect attacks in real-time.	A message authentication model with low computational overhead.
previous node and session tree techniques used to trace the route request and response flow and record the routing data in the flow.	The model proposes four detection constraints according to OLSR specification and successfully detects OLSR routing attacks	adapts to the decentralized networks of MANETs and allows distributed detectors to have sufficient routing data for detecting attacks with low message overhead.	Uses symmetric keys, but achieves functionality of public key systems by integrating one-way key chaining, one-way hash functions, and delay key disclosure. It also proposes Validation Messages.
Detectors can detect any maliciously changed content according to the traced routing data.	the model detects routing attacks with no false positives and negatives	Implemented in OLSR with three ID messages, It can tolerate temporary inconsistency. Have very low false positives, no false negatives, and low message loss or delay.	Implemented DRETA for AODV routing messages, OLSR TC messages, and DEMEM ID messages in OLSR. DRETA successfully integrates our all other work in one piece.

VIII. FUTURE WORK

Here we discuss several future works that our proposed IDS does not support.

First, this IDS only supports attack recovery by an individual node, and we can develop a reputation-based cooperative intrusion response model for DEMEM and DRETA. Second, we can apply DEMEM and DRETA to the other two routing protocols, DSR(Dynamic Source Routing)[21] and TBRPF(Topology Broadcast based on Reverse-Path Forwarding)[30]. Third, we can develop an extension of DRETA for tolerating message loss and minimizing message dropping. Finally, we will develop detection of tunneling routing attacks from correlated attackers, which cannot be detected by our proposed IDS.

IX. REFERENCES

[1]. R. Canetti A. Perrig, D. Tygar, and D. Song. The TESLA broadcast authentication protocol. Cryptobytes, 5(2):2–13, 2002.

[2]. DEMEM: Distributed Evidence-driven Message Exchange intrusion detection Model for MANET. In Proceeding of the 9th International Symposium Recent Advances in Intrusion Detection (RAID), Hamburg, Germany, 2006.

[3]. O. Kachirski abd R Guha. Effective Intrusion Detection Using Multiple Sensors in Wireless Ad Hoc Networks. In 36th Annual Hawaii International Conference on System Sciences (HICSS 2003).

[4]. C. Adjih, T. Clausen, P. Jacquet, A. Laouiti, P. Muhlethaler, and D. Raffo. Securing the OLSR protocol. In Med-Hoc-Net 2003.

[5]. Yi an Huang and Wenke Lee. Attack Analysis and Detection for Ad Hoc Routing Protocols. In Proceedings of International Symposium Recent Advances in Intrusion Detection (RAID) 2004.

[6]. Yi an Huang and Wenke Lee. A Cooperative Intrusion Detection System for Ad Hoc Networks. In Proceedings of the ACM Workshop on Security in Ad Hoc and Sensor Networks (SASN) 2003.

[7]. Farooq Anjum and Rajesh R. Talpade. LiPad: Lightweight Packet Drop Detection for Ad Hoc Networks. In Proceedings of IEEE 60th Vehicular Technology Conference 2004.

[8]. S. Buchegger and J. Boudec. Performance Analysis of the CONFIDANT Protocol: Cooperation Of Nodes - Fairness In Distributed Ad hoc Networks. In Proceedings of MobiHoc 2002.

[9]. L. Buttyan and J.-P. Hubaux. Stimulating Cooperation in Self-organizing Mobile Ad Hoc Networks. Technical Report DSC/2001/046, Swiss Federal Institute of Technology, Lausanne, 2001.

[10]. T. Clausen and P. Jacquet. Optimized Link State Routing Protocol. IETF RFC 3626.

[11]. Daniel Sterne et. al. A General Cooperative Intrusion Detection Architecture for MANETs. In Proceedings of the 3rd IEEE International Information Assurance Workshop 2005.

[12]. Dhanant Subhadhrabandhu et. al. Efficacy of Misuse Detection in Adhoc Networks. In Proceedings of SECON 2004.

[13]. K. Bhargavan et al. VERISIM: Formal Analysis of Network Simulations. IEEE Transactions of Software Engineering, 28(2):129, 2002.

[14]. Sumit Gwalani, Kavitha Srinivasan, Giovanni Vigna, Elizabeth Belding-Royer, and Richard Kemmerer. An Intrusion Detection Tool for AODV-based Ad hoc Wireless Networks. In Proceedings of Computer Security Applications Conference 2004.

[15]. Andreas Hafslund, Andreas Tonnesen, Roar Bjorgum Rotvik, Jon Andersson, and Oivind Kure. Secure Extension to the OLSR protocol. In In OLSR Interop and Workshop 2004.

[16]. N. Haller. The S/Key one-time password system. Internet Society 1994.

[17]. Уih-Chun Hu, Adrian Perrig, and David Johnson. Ariadne: A Secure On-Demand Routing Protocol for Ad Hoc Networks. In Proceedings of MobiCom 2002.

[18]. Уih-Chun Hu, Adrian Perrig, and David Johnson. Packet leashes: A Defense against Wormhole Attacks in Wireless Ad Hoc Networks. In Proceedings of INFOCOM 2003.

[19]. K. Ilgun, R. Kemmerer, and P. Porras. State Transition Analysis: A Rule-based Intrusion Detection Approach. IEEE Transactions of Software Engineering, 2(13):181–199, 1995.

[20]. H. S. Javitz and A. Valdes. The SRI IDES Statistical Anomaly Detector. In Proceedings of the IEEE Symposium on Research in Security and Privacy 1991.

[21]. David Johnson and David Maltz. Dynamic Source Routing in Ad Hoc Wireless Networks. Mobile Computing, 1996.

[22]. C. Ko, P. Brutch, and J. Rowe et al. System Health and Intrusion Monitoring Using a Hierarchy of Constraints. In Proceeding of International Symposium Recent Advances in Intrusion Detection (RAID) 2001.

[23]. C. Ko, M. Ruschitzka, and K. Levitt. Execution Monitoring of Security-Critical Programs in Distributed Systems: A Specification-based Approach. In Proceedings of the 1997 IEEE Symposium on Security and Privacy, May 1997.

[24]. H. Krawczyk, M. Bellare, and R. Canetti. HMAC: Keyed-Hashing for Message Authentication. IETF RFC 2104.

[25]. U. Lindqvist and P. Porras. Detecting Computer and Network Misuse Through the Production-Based Expert System Toolset (P-BEST). In Proceedings of the 1999 Symposium on Security and Privacy.

[26]. S. Marti, T. J. Giuli, K. Lai, and M. Baker. Mitigating Routing Misbehavior in Mobile Ad Hoc Networks. In Proceedings of MobiCom 2000.

[27]. P. Michiardi and R. Molva. Core: A Collaborative REputation mechanism to enforce node cooperation in Mobile Ad Hoc Networks. In Communication and Multimedia Security 2002 Conference.

[28]. P. Ning and K. Sun. How to Misuse AODV: A Case Study of Insider Attacks against Mobile Ad hoc Routing Protocols. In Proceedings of IEEE Information Assurance Workshop 2003.

[29]. Jorge Nuevo. A Comprehensible GloMoSim Tutorial. 2004.

[30]. R. Ogier, F. Templin, and M. Lewis. Topology Broadcast based on Reverse-Path Forwarding. IETF RFC 3684.

[31]. Panagiotis Papadimitratos and Zygmunt J. Haas. Secure Link State Routing for Mobile Ad Hoc Networks. In Proceedings of IEEE

Workshop on Security and Assurance in Ad Hoc Networks 2003.

[32]. Charles Perkins, Elizabeth Belding-Royer, and Samir Das. Ad Hoc On Demand Distance Vector (AODV) Routing. IETF RFC 3561.

[33]. Mohapatra Prasant and Krishnamurthy Srikanth. Ad Hoc Networks: Technologies and Protocols.

[34]. R. Ramanujan, S. Kudige, T. Nguyen, S. Takkella, and F. Adelstein. Intrusion-Resistant Ad Hoc Wireless Networks. In Proceedings of MILCOM 2002.

[35]. R. Rao and G. Kesidis. Detection of malicious packet dropping using statistically regular traffic patterns in multihop wireless networks that are not bandwidth limited. Brazilian Journal of Telecommunications, 2003.

[36]. Y. Rebahi, V. Mujica, C. Simons, and D. Sisalem. SAFE: Securing packet Forwarding in ad hoc networks. In 5th Workshop on Applications and Services in Wireless Networks 2005.

[37]. Kimaya Sanzgiri, Bridget Dahill, Brian Neil Levine, Elizabeth Belding-Royer, and Clay Shields. A Secure Routing Protocol for Adhoc Networks. In Proceedings of International Conference on Network Protocols (ICNP) 2002.

[38]. Chin-Yang Tseng, Poornima Balasubramanyam, Calvin Ko, Rattapon Limprasittiporn, Jeff Rowe, and Karl Levitt. A Specification-Based Intrusion Detection System For AODV. In Proceedings of the ACM Workshop on Security in Ad Hoc and Sensor Networks (SASN) 2003.

[39]. Chinyang Henry Tseng, Tao Song, Poornima Balasubramanyam, Calvin Ko, and Karl Levitt. A Specification-based Intrusion Detection Model for OLSR. In Proceeding of the 8th International Symposium Recent Advances in Intrusion Detection (RAID), Seattle, 2005.

[40]. Shiau-Huey Wang, Chinyang Tseng, Calvin Ko, and Karl Levitt. A General Automatic Response Model for MANET. In Proceeding of First IEEE International Workshop on Next Generation Wireless Networks 2005 (IEEE WoNGeN ’05).

[41]. S. Yi, P. Naldurg, and R. Kravets. Security-aware routing protocol for wireless ad hoc networks. In Proceedings of ACM MobiHoc 2001.

[42]. M. G. Zapata. Secure ad hoc on demand (SAODV) routing. IETF Internet Draft, 2001.

[43]. Y. Zhang and W. Lee. Intrusion Detection in Wireless Ad Hoc Networks. In Proceedings of MobiCom 2000.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 4 | Issue 4 | ISSN : 2456-3307

A Comparative Analysis of Various Auto-Scalers in the Cloud Environment

Dhrub Kumar[1], Naveen Gondhi[2]

[1]Scholar, Department of Computer Science and Engineering, Shri Mata Vaishno Devi University, Katra, Jammu, India

[2]Assistant Professor, Department of Computer Science and Engineering, Shri Mata Vaishno Devi University, Katra, Jammu, India

ABSTRACT

The IaaS service model offers resources to its customers in the form of virtual machines (VMs) on a pay per use basis. These days, large enterprises and even small and medium businesses (SMBs) have started deploying their applications on clouds due to the various advantages it offers. The elastic feature of the clouds lets the deployed applications to scale their resources in accordance with the workload demands. This ensures that the applications provide the guaranteed QoS to its users as specified in the SLAs. To handle the automatic acquiring and releasing of resources as per application workload demands in the cloud environment (auto-scaling), various techniques have been proposed by researchers in the past. This paper performs a comparative analysis of various auto scaling techniques in cloud with respect to a number of factors viz. scaling technique, scaling type, scaling timing, and workload nature.

Keywords : Auto-scaling, Application Provisioning, Cloud Computing

I. INTRODUCTION

Providing on-demand, scalable and virtualized resources to its customers in a pay per use fashion are some of the key features of cloud computing. Many companies are shifting towards clouds for deploying their applications to avoid over-provisioning or under-provisioning of resources and to balance the cost-performance trade off [1]. The elastic feature of clouds is attracting large enterprises and even small and medium businesses (SMBs) to host their web applications on cloud infrastructures so as to handle varying workload demands. This, in turn, leads to improved QoS guarantees and reduced rental costs. For example, Animoto – an image processing web application experienced a sudden increase in workload requests that it has to increase its number of instances from 50 to 4000 in just three days in April, 2008 [2]. This way Animoto scaled up its resources to guarantee performance to its end users and later on scaled down its resources to reduce costs. Application deployment on cloud infrastructures brings many challenges. Ensuring automatic provisioning of sufficient amount of resources to application instances according to the current workload demands taking into account performance and cost constraints is one of the challenges. Fig. 1 demonstrates the fluctuations in requests to the FIFA 1998 Soccer World Cup website. The fluctuations in requests depend on a number of factors like what time of day is it, what day of week is it, and other seasonal factors.

Allocating suitable resources for such a workload is quite challenging. If resources are allocated according to average workload (under-provisioning), then cost of renting resources from cloud is low but at the same time performance will be affected as the end users may experience long delays or service unavailability. On the other hand, if resources are allocated according to peak workload (over-provisioning), then application QoS requirements will be met but at a higher cost as resources will remain idle most of the times. To tackle these problems of over and under provisioning and under provisioning, auto scaling is employed in cloud environments. From the application provider's perspective, lack of both expert knowledge about application dynamics and modeling expertise complicate the scaling of the cloud hosted applications [4].

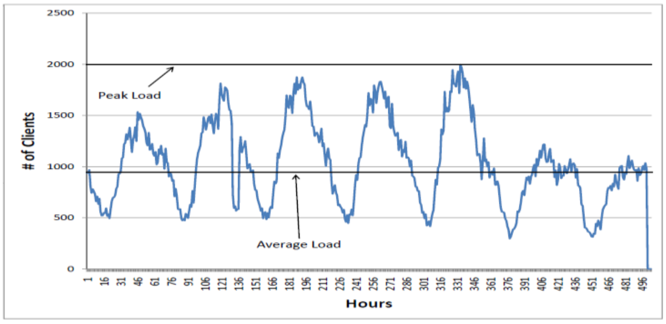


Figure 1. Workload of Soccer World Cup 1998

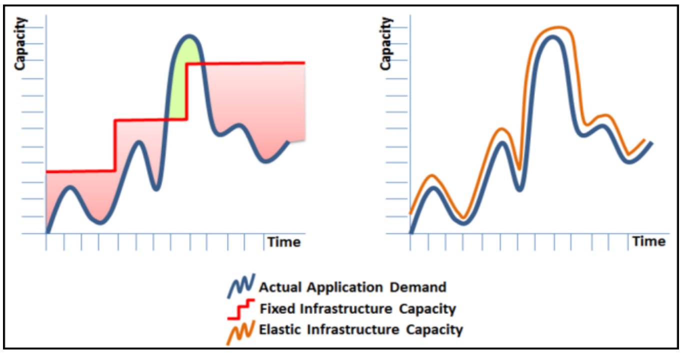
In section 2, the concept of auto-scaling is explored with respect to cloud environment. The work related to auto-scaling is summarized in section 3. Section 4 performs comparative analysis of various successful auto-scaling models proposed by various researchers in the past. Finally, section 5 concludes the paper.

II. AUTO-SCALING IN CLOUDS

In this section, we present the various concepts related to auto-scaling in cloud environment. We discuss what auto-scaling means in a cloud environment, the direction of scaling – horizontal or vertical, the timing of scaling – reactive or proactive, and the techniques used for auto-scaling.

Auto-Scaling

The process of acquiring and freeing resources as per application's workload demands in a dynamic, automatic fashion that takes into account resource costs and performance guarantees is called auto-scaling. According to [6], auto-scaling ensures that an application has the correct number of Amazon EC2 instances to handle the application's load. Fig. 2 shows two different ways of provisioning resources to application's workload [7]. The left portion shows the traditional way of provisioning resources where a business gradually increases its in-house infrastructure capacity to meet increasing application demands. On the right side of the graph, the cloud model allows business applications to scale resources up or down in line with the application's workload. This leads to better performance and reduces cost of renting infrastructure.



Capacity
Time
Capacity
Time
Actual Application Demand
Fixed Infrastructure Capacity
Elastic Infrastructure Capacity

Figure 2. Traditional model versus cloud capacity model

Static versus Dynamic Provisioning

Auto-scaling uses the dynamic provisioning approach. Unlike dynamic provisioning, where resource allocation may be changed during runtime, the static provisioning keeps the resources assigned to an application fixed i.e. adding or removing new VMs is not done even if a change in application workload is detected [3]. Under dynamic provisioning, there are two ways of scaling resources in response to changing workload demands viz. horizontal and vertical. Horizontal scaling deals with adding new VMs or removing the allocated ones. On the other hand, vertical scaling configures the resources (CPU, Storage etc) assigned to an already allocated VM. Vertical scaling uses a technique called as hot plug, which changes configuration of a VM on the fly without requiring it to shut down. According to [5], vertical scaling is better than horizontal scaling as VM instance acquisition time is shorter in vertical scaling.

Reactive versus Proactive Scaling

This relates to the timing of performing auto-scaling in cloud environments. The auto-scaler can be reactive or proactive. In case of reactive scaling, application is scaled only when certain pre-defined conditions are met, for example- when the CPU utilization stays over 80% for 2 minutes. This scheme is rule-based and often requires setting threshold values on part of the user and when these predefined thresholds reach certain values, some scaling action is triggered [8]. In comparison, proactive scaling relies on making predictions about future workload demands and then provisioning or de-provisioning resources accordingly.

Auto-scaling Techniques

A number of approaches have been tried in past by various researchers for implementing auto-scaling systems. Each implementation has its own scenario viz. the objectives to meet, the application architecture, the scaling parameters, the scaling method etc. In the past, researchers have used the following techniques to build an auto-scaling system:

1. Rule-based approach
2. Time Series analysis
3. Queuing theory
4. Control theory
5. Machine learning

1. Rule-based approach:

This technique is purely reactive, simple and easy to implement. It requires application providers to specify scaling indicators and set threshold values for these. On occurrence of the specified event, scaling action is triggered. Amazon’s Auto-scaling Service is also rule-based [9]. Rule-based approaches are less accurate as they take scaling action after the workload changes. Also, deciding selected thresholds for the application is also a challenging task and requires a deep understanding of the nature of workloads. In [11], Dutreilh et al. emphasized the careful tuning of thresholds to avoid oscillations in the system. To tackle this, a cool-down period is set during which no scaling decisions is allowed once a scaling action has been implemented.

2. Time Series analysis:

Most auto-scalers are exploiting the timely patterns associated with cloud workloads like day, week or month, for forecasting future workload requests. The methods commonly used for forecasting future workloads include:

a. Moving Average
b. Auto-regression
c. ARMA (auto-regressive moving average)
d. ARIMA (auto-regressive integrated moving average)
e. Exponential Smoothing

In [10], authors have evaluated various forecasting methods using Google cluster data and Intel NetBatch logs for predicting future workloads in cloud environment. Their findings suggest that no method is always accurate and the accuracy of the prediction made by a particular method depends on the frequency and type of the workload.

3. Queuing theory:

Queuing theory has been used in auto-scaling environments to predict future resource requirements by modeling the system. Queueing theory deals with the study of waiting lines or queues in mathematical form. Queueing theory uses probabilistic methods in order to predict queue length or average waiting time of workload requests in a cloud environment. In 1953, Kendall represented queuing model using the notation A/S/C, where A is the time between the arrivals, S is the time needed to service the job, and C represents the number of servers. Queuing model relies on online monitoring or other different methods in order to estimate parameters such as the input workload or service time [20].

3. Control theory:

Control theory controls an object by treating it as an input/output system, where the input corresponds to the control knobs and the outputs correspond to the metric being monitored [31]. Control theory has been widely used for designing auto-scalers in the cloud environment. Control systems can be classified as: open-loop, feed-back and feed-forward. Out of these, feed-back controllers are mostly used for auto-scalilng. Control theory works by first creating the application model in order to adjust the resources dynamically as per agreed SLAs. Control system should be adaptive to varying workload characteristics or the application itself. Control systems work in both reactive and proactive modes [32].

5. Machine learning:

Machine learning (ML) is closely associated with artificial intelligence, data mining and pattern recognition and has been broadly classified into supervised learning, semi-supervised learning and unsupervised. ML requires creating empirical models in order to understand application dynamics and make precise predictions. Various machine learning techniques like support vector machine, linear regression, neural networks, reinforcement learning etc were used by researchers as a predictive tool to make future workload predictions in cloud environment. In [22], authors observed that SVM provides more accurate results as compared to neural networks and linear regression models in terms of response time and throughput.

In [34], Gong et al. proposed a model called PRESS which uses statistical machine learning to perform resource auto-scaling by predicting future resource demands. In [21], Zhang et al. applied regression-based approximation to estimate the CPU demand, based on the number/type of requests. In [35], Islam et al. applied sliding window to linear regression and correction neural network for performing resource predictions in cloud environment. In [10], Xu et al. found optimal VM configurations in cloud computing environment by applying a unified reinforcement learning approach.

The following table summarizes the various auto-scaling techniques.

Table 1. Comparison of various auto-scaling techniques

Technique	**Working**	**Pros**	**Cons**
Rule based	Works on the principle of setting thresholds and corresponding actions.	Simple and easy to implement	Lacks accuracy and prediction
Time Series	Utilizes a series of historical data values in order to predict future values	Capable of predicting future workloads	Selection of history window is difficult
Queuing Theory	Works by modeling queues to describe the processes behind them and to predict their behavior	Allows the modeling of systems using probabilistic distributions like the	Most of the queuing model are still complex

		Poison and exponential distributions	
Control Theory	Controls the behavior of a dynamic system by comparing the output with a desired value	Use of feedback makes system quickly adapt to varying workload	Difficult to find static control setting so as to make the system stable (static output feedback stabilization problem)
Machine Learning	Deals with training a machine to learn from its past experience so as to improve performance	Automates analytical modeling and enable access to hidden insights	Overhead in learning from a large state space

III. RELATED WORK

This section compares the work done in the area of auto-scaling in the cloud computing environment. The comparison of various works is based on parameters viz. the underlying technique, type of scaling (horizontal or vertical), timing of scaling (Reactive, Proactive or Hybrid, nature of the workload, and the year of publication.

Table 2. Comparison of work done in auto-scaling in cloud domain

Ref	Underlying Technique	Type of scaling (H/V)	Timing of scaling (R/P/ Hybrid)	Metrics used	Nature of Workload	Year
12	Time Series	H	P	Execution time	Real world (Wikimedia Foundation)	2013
22	Queuing Theory	H	P	Request Rate	Real world (Wikipedia Traces)	2013
23	Time Series (Regression)	H	P	CPU (MIPS)	Synthetic	2015
24	Machine Learning	H	P	Response Time	Real world (NASA, Wikipedia, FIFA 98 world cup traces)	2015
25	Hybrid (Autonomic computing + Reinforcement	H	P	CPU utilization/ Response Time	Real world (ClarkNet and NASA traces)	2017

	Learning)					
26	Hybrid (Threshold based + Heuristic)	H	P	Response Time	Real world (EPA, SDSC and ClarkNet traces)	2014
27	Queueing Theory	V	P	Latency and Throughput	Real world (FIFA98 world cup traces)	2014
28	Machine Learning	H	P	CPU and Memory	Synthetic	2016
29	Time Series	H + V	P	Response Time	Real world (FIFA 98 world cup traces)	2016
30	Control Theory	H + V	P	Response Time	Nginx logs	2015

IV. CONCLUSION

The elastic nature of cloud computing enables the on demand provisioning and deprovisioning of resources in an automatic fashion. However, auto-scaling resources in cloud is a challenging task due to the unpredictable nature of web applications keeping in mind the SLA requirements of the end user. In this paper, we have presented the various aspects of auto-scaling in cloud and performed an exhaustive comparison of recent work done in the field of auto-sclaing in cloud environment.

V. REFERENCES

[1]. JoSEP, A. D., Katz, R., Konwinski, A., Gunho, L., PAttERSon, D., & RABKin, A. (2010). A view of cloud computing. Communications of the ACM, 53(4)

[2]. "Animoto case in rightscale blog," http://blog.rightscale.com/2008/04/23/animoto-facebook-scale-up/

[3]. Shoaib, Y., & Das, O. (2014). Performance-oriented Cloud Provisioning: Taxonomy and Survey.arXiv preprint arXiv:1411.5077

[4]. Gandhi, A., Dube, P., Karve, A., Kochut, A., & Zhang, L. (2014, June). Adaptive, Model-driven Autoscaling for Cloud Applications. In ICAC(Vol. 14, pp. 57-64)

[5]. Yazdanov, L., & Fetzer, C. (2012, November). Vertical scaling for prioritized vms provisioning. In Cloud and Green Computing (CGC), 2012 Second International Conference on(pp. 118-125). IEEE

[6]. http://docs.aws.amazon.com/autoscaling/latest/userguide/WhatIsAutoScaling.html

[7]. https://www.packtpub.com/books/content/elastic-load-balancing

[8]. Loff, J., & Garcia, J. (2014, December). Vadara: Predictive elasticity for cloud applications. In Cloud Computing Technology and Science (CloudCom), 2014 IEEE 6th International Conference on(pp. 541-546). IEEE

[9]. Amazon. 2016. Amazon Auto Scaling Service. (2016). http://aws.amazon.com/autoscaling/ Carlos Vazquez- time series

[10]. Xu, C. Z., Rao, J., & Bu, X. (2012). URL: A unified reinforcement learning approach for autonomic cloud management. Journal of Parallel and Distributed Computing, 72(2), 95-105

[11]. Dutreilh, X., Moreau, A., Malenfant, J., Rivierre, N., & Truck, I. (2010, July). From data center resource allocation to control theory and back. In Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on(pp. 410-417). IEEE

[12]. Calheiros, R. N., Masoumi, E., Ranjan, R., & Buyya, R. (2015). Workload prediction using ARIMA model and its impact on cloud applications' QoS. IEEE Transactions on Cloud Computing, 3(4), 449-458

[13]. Calheiros, R. N., Ranjan, R., & Buyya, R. (2011, September). Virtual machine provisioning based on analytical performance and QoS in cloud computing environments. In Parallel processing (ICPP), 2011 international conference on(pp. 295-304). IEEE

[14]. Ferretti, S., Ghini, V., Panzieri, F., Pellegrini, M., & Turrini, E. (2010, July). Qos–aware clouds. In Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on(pp. 321-328). IEEE

[15]. Gong, Z., Gu, X., & Wilkes, J. (2010, October). Press: Predictive elastic resource scaling for cloud systems. In Network and Service Management (CNSM), 2010 International Conference on(pp. 9-16). IEEE

[16]. Jiang, J., Lu, J., Zhang, G., & Long, G. (2013, May). Optimal cloud resource auto-scaling for web applications. In Cluster, Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium on(pp. 58-65). IEEE

[17]. Roy, N., Dubey, A., & Gokhale, A. (2011, July). Efficient autoscaling in the cloud using predictive models for workload forecasting. In Cloud Computing (CLOUD), 2011 IEEE International Conference on(pp. 500-507). IEEE

[18]. Fernandez, H., Pierre, G., & Kielmann, T. (2014, March). Autoscaling web applications in heterogeneous cloud infrastructures. In Cloud Engineering (IC2E), 2014 IEEE International Conference on(pp. 195-204). IEEE

[19]. Nguyen, H., Shen, Z., Gu, X., Subbiah, S., & Wilkes, J. (2013, June). AGILE: Elastic Distributed Resource Scaling for Infrastructure-as-a-Service. In ICAC(Vol. 13, pp. 69-82)

[20]. Urgaonkar, B., Shenoy, P., Chandra, A., Goyal, P., & Wood, T. (2008). Agile dynamic provisioning of multi-tier internet applications. ACM Transactions on Autonomous and Adaptive Systems (TAAS), 3(1), 1

[21]. Zhang, Q., Cherkasova, L., & Smirni, E. (2007, June). A regression-based analytic model for dynamic resource provisioning of multi-tier applications. In Autonomic Computing, 2007. ICAC'07. Fourth International Conference on(pp. 27-27). IEEE

[22]. Bankole, A. A., & Ajila, S. A. (2013, May). Predicting cloud resource provisioning using machine learning techniques. In Electrical and Computer Engineering (CCECE), 2013 26th Annual IEEE Canadian Conference on(pp. 1-4). IEEE

[23]. Al-Ayyoub, M., Jararweh, Y., Daraghmeh, M., & Althebyan, Q. (2015). Multi-agent based dynamic resource provisioning and monitoring for cloud computing systems infrastructure. Cluster Computing, 18(2), 919-932

[24]. Liu, J., Zhang, Y., Zhou, Y., Zhang, D., & Liu, H. (2015). Aggressive resource provisioning for ensuring QoS in virtualized environments. IEEE Transactions on Cloud Computing, 3(2), 119-131

[25]. Ghobaei-Arani, M., Jabbehdari, S., & Pourmina, M. A. (2017). An autonomic resource provisioning approach for service-based cloud applications: A hybrid approach. Future Generation Computer Systems

[26]. Tighe, M., & Bauer, M. (2014, May). Integrating cloud application autoscaling with dynamic vm allocation. In Network Operations and

Management Symposium (NOMS), 2014 IEEE(pp. 1-9). IEEE

[27]. Spinner, S., Kounev, S., Zhu, X., Lu, L., Uysal, M., Holler, A., & Griffith, R. (2014, September). Runtime vertical scaling of virtualized applications via online model estimation. In Self-Adaptive and Self-Organizing Systems (SASO), 2014 IEEE Eighth International Conference on(pp. 157-166). IEEE

[28]. Grozev, N., & Buyya, R. (2016). Dynamic Selection of Virtual Machines for Application Servers in Cloud Environments. arXiv preprint arXiv:1602.02339

[29]. Hirashima, Y., Yamasaki, K., & Nagura, M. (2016, July). Proactive-Reactive Auto-Scaling Mechanism for Unpredictable Load Change. In Advanced Applied Informatics (IIAI-AAI), 2016 5th IIAI International Congress on(pp. 861-866). IEEE

[30]. Dupont, S., Lejeune, J., Alvares, F., & Ledoux, T. (2015, September). Experimental analysis on autonomic strategies for cloud elasticity. In Cloud and Autonomic Computing (ICCAC), 2015 International Conference on(pp. 81-92). IEEE

[31]. Zhu, Q., & Agrawal, G. (2010, June). Resource provisioning with budget constraints for adaptive applications in cloud environments. In Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing(pp. 304-307). ACM

[32]. Al-Dhuraibi, Y., Paraiso, F., Djarallah, N., & Merle, P. (2017). Elasticity in Cloud Computing: State of the Art and Research Challenges. IEEE Transactions on Services Computing

[33]. Bankole, A. A., & Ajila, S. A. (2013, May). Predicting cloud resource provisioning using machine learning techniques. In Electrical and Computer Engineering (CCECE), 2013 26th Annual IEEE Canadian Conference on(pp. 1-4). IEEE

[34]. Gong, Z., Gu, X., & Wilkes, J. (2010, October). Press: Predictive elastic resource scaling for cloud systems. In Network and Service Management (CNSM), 2010 International Conference on(pp. 9-16). IEEE

[35]. Islam, S., Keung, J., Lee, K., & Liu, A. (2012). Empirical prediction models for adaptive resource provisioning in the cloud. Future Generation Computer Systems, 28(1), 155-162

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

A Survey on the State of Art Approaches Used in Intrusion Detection System

Satish Kumar[1], Dr. Sunanda[2], Dr. Sakshi Arora[2]
[1]Research Scholar, Department of CSE, SMVD University, Katra, Jammu and Kashmir, India
[2]Assistant Professor, Department of CSE, SMVD University, Katra, Jammu and Kashmir, India

ABSTRACT

The security has always been the prime issue for a user as well as for the network system. Intrusion detection is being used as security other than the first line of security like firewall in which malicious packets are prevented from being penetration to the target. Within the development of the technologies and system resources, there have always been intrusion detection systems which are capable in detection of malicious attack in an efficient manner with less false positive instances. This paper reveals current scenarios of used technologies for the purpose of detection of intrusions.

Keywords : IDS, Anomaly, Malicious Attacks, Detection Rate, False Positive Intrusion.

I. INTRODUCTION

To promote internet security mechanism, there are large numbers of techniques used. Like firewalls, authentication and access control and data encryption are considered as the first line security among these security techniques and these first line security defences are not sufficient for covering the overall network security mechanism whereas another line of security defence is intrusion detection system (IDs). Now-a-days, use of IDs with antivirus has a significant impact on computer network security mechanism and provides a more prominent scenario for protecting computer network from unauthenticated access control services. However, there is no such technique/s or approaches that guarantee the full protection of computer network. [1, 2]

1.1 Intrusion Detection System (IDs)

According to National Institute of Standard and Technology, intrusion detection is defined as "The process of monitoring the events occurring in a computer system or network and analysing them for sign of intrusions, defined as attempts to compromise the confidentiality, integrity, availability or a bypass the security mechanism of a computer or a network."[3]

The monitoring processes can be accomplished with the help of software or hardware to secure the system from malicious activity or from the violation of the policies of the system for integrity. Intrusion detection system usually does not provide prevention of the system from intrusion attack rather than it merely generates an alarm after detection of an attack in the system in real time or in efficient time. It is equally important to generate an alarm for an attack after it happened in the system because an IDS maintains and update the profile of an intrusion in the log.

The information generated by IDS goes to either SIEM (Security Information Evolution and

Management System) or to the network administrator. The wide spectrums of IDS are antivirus, traffic monitoring and host based Intrusion detection system (HIDS). Systems with response capabilities are called Intrusion Prevention System (IPS).

On the basis of analysis the intrusion detection system can be divided into Network based Intrusion Detection System (NIDS) and Host based Intrusion Detection System (HIDS). First, NIDS, that is based on detection of attack from the interconnection of computers and the intrusion detection approaches for NIDS can be further divided into i.) Misuse/ Known Based NIDS ii.) Anomaly/ Unknown Base NIDS. Second, HIDS, in which attacks are detected from a single computer system and these attacks, are easy to prevent. HIDS also monitor important files of operating system. The attacks in HIDS usually comes from externally connected devices like pen drive, CD, DVD, floppy etc.

Hybrid IDS system have been also introduced which can be implemented on the both on host as well as Network.[5]

1.2 Attacks

The recognition of the pattern of attacks broadly can be categorized in to Known/Misuse/Signature based attacks, Unknown/anomalies based attacks and Specification based attacks. First types of attack are of general types and simple to process, locate and implement [6]. There is requirement of continuous maintaining and updation of signature's log files that contains the list of known attacks detecting from computer or network system. The second types of attacks are detected on the observation of deviation from normal attack behaviour. There is need to establish each user's normal activity profile and marking of flag deviations from the established activity profile for the attacks. Detection of attack of this types are computationally complex and expensive and hence time consuming as because of keeping track of pattern of attacks, updating of (several) system profile matrices. Third types of attacks take regards of various features and parameter's consideration and compare these specifications with the bench marks established in the dataset.

1. Datasets

Datasets are used for the classification and establishing the benchmark [6, 7] for the intrusion and intrusion detection. The various data sets as the benchmark are C, NSL- KDD, ISCX 2012 Data set based on Data Set (KDD Cup 99) and Kyoto 2006+ etc. On the bases of KDD Cup 99, the intrusions can be classified into four Groups [6,7]. Namely, i.) DoS: Denial of Service Attack. Attacker/s makes flooding of superfluous request on the target machine and hence makes busy the memory and computing resources of the machine to avoid the fulfilment of legitimate request of users by the machine. ii.) R2L: Remote to Local Attack. Attackers access the network and penetrate into the network with unauthorized access and breach the confidentiality of the system's information. iii.) U2R: User to Root. Like a sniffer, the intruder watches on the activity and event on the network and use that information for the purpose of misuse of information. iv.) Probe. This is based on the working of surveillance and other probing, like port scanning.

2. Components of IDS[8]

Usually, an IDS consists of three components, namely

I. **Event Generator (Data Source):** this act as a monitor and the faction of working of event generator can be a HOST based Monitor, Network based Monitor, Application based Monitor or a Target based Monitor.
II. **Analysis Engine:** It takes information from the data source and examines the data for symptoms of attacks or other policy violations.
III. **Response Manager:** when susceptible intrusion attacks are found on the system, the

response manager act as informer to the administrator or generate alarm.

II. LITERATURE SURVEY

B.A. Fessi, S. Ben Abdallah, M. Hamdi and N. Boudriga[9] applied genetic algorithm approach for Intrusion Response System (IRS) which is a decision part in the NIDS. IRS takes decision on three approaches - a.) Notification Response System (NRS) b.) Manual Response System (MRS) c.) Automatic Response System (ARS). NRS just generate alarm or response on anomalies detection whereas MRS works based on human intervention with high degree of automation than NRS. ARS generate an immediate response through an automated decision making tool. This paper also reveals the combined work in the field of artificial intelligence and computer security. As this work was based on GAs and GA's require high resource consumption involved.

S. Devaraju and Dr. S. Ramakrishnan[10] presents the analysis of performance of intrusion Detection system using neural network classifier has been explored in this paper. NN classifier like PNN (Probabilistic Neural Network) and Radial Basis Neural Network are used in MATLab for the analysis of performance of IDS applied on KDD Cup 99 dataset. The performance of full dataset and reduced dataset is analysed. The use of neural network gives the better result in learning of the weight and parameters for optimisation. More better results comes in the hybrid form of IDs.

Mohammed A. Ambusaidi, Xiangjian He, Priyadarsi Nanda and Zhiyuan Tan[11], this paper reveals the problem of redundant and irrelevant features in data that are the cause of network traffic classification. The problem of slow down of process of classification and problem of accurate decision by classifier also described. Here mutual information based algorithm that analytically selects the optimal feature for classification also been revealed. An algorithm MIBFS, Mutual Information Based Feature Selection Algorithm's ability to handle linearly and non linearly dependent data features are also described along with an IDS, Least Squared Support Vector Machine Based IDS (LSSVM-IDS) building by using feature selection techniques. The performance evolution is also been performed using the data set: KDDCup 99, NSL-KDD and Kyoto 2006+ dataset. Least Square and SVM approaches used and hence complexity with time consumption comes arises.

Xu Yang and Zhao Hui[12], an IPSO- RBF model (improved particle swarm optimization- radial basis function network) for intrusion detection has been proposed which decrease the feature dimension in feature selection and obtains the better RBF neural Network parametric values in a network. No doubt that IPSO-RBF reduces the feature dimensions. But complexity increases with the population of data.

Audrey A. Gendreau[4], Survey of Intrusion Detection Systems towards an End to End Secure Internet of Things (IoT) and this survey of the Intrusion Detection Systems (IDS) use the most recent ideas and methods to propose the present IoT. To understand and illustrate IDS platform differences and the current research trend towards a universal, cross-platform distributed approach has been taken in the consideration.

A. Gupta and O. J. Pandey[13], a proposal for Computational Intelligence (CI) based systems have been discussed which is adaptable and react to new situations by applying reasoning without relying on users. A 3-Tier architecture for monitoring intrusion by applying computational intelligence and reporting to administrator has proposed. The IP addresses of the source messages have been tracked in IDS and store it against their network or system patterns. Tier architecture induces complexity in IDs and hence time consumption also increases.

Fatemeh Kavousi and Behzad Akbari's paper[14] helps to find out the identification of the attack behaviour patterns. This paper reveals the attack strategies through automatic analysis of intrusion alerts. A new algorithm to mine attack behaviour patterns from a large number of intrusion alerts without specific prior knowledge about attacks. A neural network (Bayesian) mechanism for automatically generation of correlation rules from previously observed alerts. This new introduced approach helps us to predict forthcoming attack in a real time system.

Richard Zuech et al[15], this paper explores a survey on IDS and 'Big Data' with the illustration of monitoring heterogeneous source. Deep packet inspections along with 'Big Data' challenges over heterogeneous data of host log events data handling issues are discussed. With dealing with Big Data more resources are required for maintaining log events.

Chun Guo et. al.[16], proposed a clustering based hybrid approach, ADBCC (anomaly Detection Method Based on the Changing of cluster Centre) and calculating the distance form cluster head. Chun Guo et al proposed his two level hybrid model based on K-NN with 96% accuracy. But for misuse attacks, there is no or very less applicability.

Chi-Ho Tsang, Sam Kwong and Hanli Wang[17] present a novel intrusion detection approach to extract both accurate and interpretable fuzzy IF–THEN rules from network traffic data for classification. The proposed fuzzy rule-based system is evolved from an agent-based evolutionary framework and multi-objective optimization. In addition, the proposed system can also act as a genetic feature selection wrapper to search for an optimal feature subset for dimensionality reduction. To evaluate the classification and feature selection performance of the proposed approach, it is compared with some well-known classifiers as well as feature selection filters and wrappers were used. The extensive experimental results on the KDD-Cup99 intrusion detection benchmark data set demonstrate that the proposed approach produces interpretable fuzzy systems, and outperforms other classifiers and wrappers by providing the highest detection accuracy for intrusion attacks and low false alarm rate for normal network traffic with minimized number of features. Due to If-Then-Else approach, complexity increase.

Sumaiya Thaseen Ikram, Aswani Kumar Cherukuri[18], proposed a model using fusion of chi-square feature selection and multi class SVM. A parameter alteration technique is used for optimization of Radial Basis Function kernel parameter namely gamma represented by'γ'and over fitting constant C'. These are two important parameters required for the SVM model. This model use the idea of building a multi class SVM which is not so far used for IDS to reduce the training and testing time and also increase accuracy for classification of the network attacks.

E. Biermann, E. Cloete, L. M. Venter [19], Compares the various IDS systems and provide the 'Best Fit' norms for selection of an IDS to system. The proposed work tries to assist for the selection of a single appropriate IDS or combined approaches that may be suitable for a particular computer or network system. This approach does not support for introduction of a general purpose IDS.

TIAN Xin- Guang et. al. [20], introduced a machine learning based method for anomaly detection of user behaviour in host based intrusion detection. The methodology is based on the shell command patterns of user's behaviours profile. The drawback of this implementation is of being large user's profile and updating of log/profile file of users.

Enamul Kabir et. al.[21] proposed Least Square Support Vector Machine as a novel statistical technique for intrusion detection systems. Which is based on the idea of sampling and this is referred as the optimum allocation based least square support vector machine (OA-LS-SVM). This approach divide the training and testing dataset into some predetermined subgroups of arbitrary instances and these instances are used as input set in LS-SVM to detect different intrusions. The results of this research show that the used method is effective for detecting intrusions for static not so for dynamic.

3. IDS Technologies

The Various approaches used in IDS are basically depends on the following approaches [23, 24].

A.) **Statistical Based[21] (Stochastic Behaviour and well defined)**

The merits and demerits of the statistical approaches used in the process of intrusion detection are shown in Table 1.

B.) **Machine Learning Based(Categorization of Patterns)**

Along with the merits and demerits of machine learning, the other combined method are also described in Table 2 for the purpose of classification of attacks and features of intrusion.

C.) **Bio- inspired Algorithms Based**

Whereas Table 3's contents show the various types of bio-inspired approaches likes ACO, BFO, BAT etc. are used with other types of approaches in intrusion detection systems

D.) **Fuzzy Logic Based**

Table 4 shows the fuzzy logic's use with merits and demerits in the field of intrusion detection

It is difficult to maintain the record of various computational task and the used algorithms that were developed to solve these complex problem. Classical problem solving methodologies involve two branches: Exact methods (logical, mathematical programming) and Heuristics. Heuristic approach seems to be superior in solving hard and complex optimization problems, particularly where the traditional methods fail.

Comprehensive approaches like heuristics along with the formal structure like algorithms, probabilistic, statistical or rationalistic reasoning provide improved approaches in monitoring events generated from various heterogeneous sources and generate more realistic awareness to intrusive attack

Table 1. Statistical Based IDs Approach

Used Approach		Merits	Demerits
Statistical Based	**chi-square feature selection and multi class SVM**[18]	Accurate notification of malicious activities	• Difficult setting for parameters and metrics • Statically based approaches are Susceptible to be trained by attackers. • Difficult setting for parameters and
	Optimum Allocation-based least square support vector machine (OA-SSVM) for IDS[21]	• The CRF, Naïve Bays and Decision Tress are not at all suitable for detecting R2L attack but OA-SSVM is Suitable for R2L attacks.[21]	

	Optimization of Feature Selection[25] using correlation analysis and association impact scale	• Detecting the association of each feature and canonical correlation of the features. • High classification • accuracy, and meanwhile reduce the complexity of the • Rules that are extracted from training data.	metrics. • Unrealistic quasi-stationary

Table 2. Machine Learning Based IDs Approach

Used Approach		Merits	Demerits
Machine Learning Based	**KNN based Classifier Systems for Intrusion Detection[22]**	Classifying network traffic using SVM (support Vector Machine)	• Complex and More stage • High Dependency on the assumption about the behaviour accepted for the system. • High resource consuming.
	Novel KPCA-GA-SVM[26]	• selected samples from the subset of KDD • The subset was randomly divided into two subsets viz. Normal and Abnormal class	
	Confusion matrix[5] feature selection analysis and building hybrid efficient model	• Building the hybrid model • Representing the dataset and choosing the important features • Training classifier and classification	
	Graph based machine learning for Classification[24,28]	• Optimal Classification • Reduced False Positive Alarm generation • Clustering algorithm for grouping the training set into k clusters as the training subsets	

Table 3. Bio- Inspired Algorithm Based IDs Approach

Used Approach			Merits	Demerits
Bio-Inspired	GA (Easy	NSGA-III	• NSGA-III starts with a random population	• Despite from global search heuristics

Algorithm Based	Training. So adding new rules. But the Crossover rate is low.) Another type of machine learning-based technique		• Solve many-objective optimization problems • Higher classification accuracy and lower computational complexity	and particular class of evolutionary algorithms with converges to a solution from multiple directions, GA's require high resource consumption involved.
		fuzzy association rule mining classifier[29]	• Use of Genetic Fuzzy Systems • Classification	
		feature selection analysis[5]	improved accuracy, high false negative rate, and low false positive rule	
	ACO	ACO with SVM[30]	Reduce Mis-Classification and Clustering	• Complexity increases with the population of data. • Unlike the formal structure like algorithms, bio-inspired heuristic do not guarantee of optimal or even feasible solution and are often used with no theoretical guarantee.
		Alarm Filtering[11]	• Reduce False Positive Alarm Generation • False alarm rate is reduced • Convergence rate is higher	
	PSO	IPSO-RBF[12]	• Improve the accuracy rate of NIDS. • Highest accuracy rate of intrusion detection, feature subset selection or optimizing the neural network parameter only can optimize one aspect, without considering the memory contacts between them. • IPSO-RBF reduces the feature dimensions, and obtains better RBF neural network parameter, and improves the network intrusion detection effects	
	PCO, Fuzzy with Machine Learning	particle swarm optimization Clustering	• Divide and conquer • MCLP/**SVM optimized by time-varying chaos particle swarm**	

		between normal and attacks[31]	**optimization** • Better separability between a "normal activity' and the different attack types.	
	BAT	• Classification Model • Select the best features for detecting intrusions • performance of a Neural Network-RIPPER (Repeated Incremental Pruning to Produce Error Reduction)		
	Cuttlefish	Remove the Redundant and irrelevant features		

Table 4. Fuzzy Logic Based

Used Approach	**Merits**	**Demerits**
Fuzzy logic[219] Effective, for port scans and probes.	• Deals with uncertainty and complexity. • Easy feature selection and decision of degree of maliciousness of intrusion instead of 'yes'/ 'No'. • 'If-then-else' rules are easily defined.	Reasoning is approximate rather than precise

4. Data sets used for Experiments

In most of the simulation studies of intrusion detection system for computer security, the KDD'99 dataset have been used. Some changes have been introduced in the KDD'99 to introduce a new dataset called NSL-KDD that consist of selected records of the complete KDD Datasets.[76]

[76] Shows that the performance of learning machines on the KDD'99 data set are not reliable and cannot be used as good indicators of the ability of the classifier to serve as a discriminative tool in network-based anomaly detection. On the contrary, KDDTrain+, KDDTest+ and KDDTest–21 test set provide more accurate information about the capability of the classifiers.

Although beyond the limitations of the KDD'99 data like poor evolution of anomaly detection approaches and affect on the performance of evaluated system, it still remains a standard and benchmark dataset that is widely used in the design of network intrusion detection due to the free availability and wide testing and training data available as labelled and unlabelled in KDD'99[16, 76].

KDD'99 data contain three labelled classes

a. Full training set,
b. The 10% training set
c. The test set.

Each record in these datasets contains 41 features, and a label provides its type. All of the attack records in the KDD'99 data are mapped to four basic attack classes, namely DoS, Prb, U2R and R2L.

Kyoto 2006+, Kyoto university benchmark dataset (KUBD)[27] and ISCX 2012 Data set are also used for setting benchmark [31, 27]

5. Measurement Metrics [27][33]

Performance of an IDS system can be calculating by detection accuracy, precision, and recall percentage. On the basis of these matrices, we can calculate the relevant usefulness of the IDS. The equation for these metrics are given below

i. Accuracy(also for classification)

$$Accuracy = \frac{\mathrm{TP + TN}}{TP + TN + FP + FN} \text{ X } 100 \quad ..(1)$$ [33]

ii. True Negative Rate (TNR) or Recall

$$TNR = \frac{TN}{TN+FP} \quad ....(2)$$ [16]

iii. False Positive Rate or False Alarm Rate (FPR) [27]

$$FPR = \frac{FP}{FP+TN} \quad ....(5)$$ [16]

iv. Detection Rate (DR)[27]

$$DR = \frac{TP}{TP+FN} \quad ....(4)$$ [16]

v. Precision[33]

PRECISION (P) is the proportion of attack cases that are correctly predicted relative to the predicted size of the attack class.

Where

True positive (TP): Number of samples correctly predicted as attack class

False Positive (FP): Number of samples incorrectly predicted as attack class

True Negative (TN): Number of samples correctly predicted as normal class

False Negative (FN): Number of samples incorrectly predicted as normal class.

III. CONCLUSION

From using of statistical method based single level IDS, various approaches based on statistical and machine learning techniques has been introduced to provide more accuracy and efficiency in the process of intrusion detection. More recently, statistical, machine learning and bio inspired algorithmic approach used and detection rate along with accuracy rate also increased tremendously in IDS. The false positive instances also reduced by using these hybrids approaches in 2-level or multi level intrusion detection system.

IV. REFERENCES

[1]. Richard Zuech, Taghi M. Khoshgoftaar and Randall Wald: "Intrusion detection and Big Heterogeneous Data: A Survey" in Journal of Big Data (2015), DOI 10.1186/s40537-015-0013-4, Springer Open Journal.

[2]. Chin-Tser Huang, Rocky K. C. Chang, and Polly Huang: "Signal Processing Applications in Network Intrusion Detection Systems"; Hindawi Publishing Corporation, EURASIP Journal on Advances in Signal Processing, Volume 2009, Article ID 527689, DOI: 10.1155/2009/527689

[3]. Praveen Lalwani, Sagnik Das: "Bacterial Foraging Optimization Algorithm for CH selection and Routing in Wireless Sensor Networks"; 3rd International Conference on Recent Advances in Information Technology, RAIT- 2016

[4]. Audrey A. Gendreau, Michael: "Survey of Intrusion Detection Systems towards an End to End Secure Internet of Things"; 4th International Conference on Future Internet of Things and Cloud, IEEE, 2016

[5]. Shadi Aljawarneha, Monther Aldwairi, Muneer Bani Yasse: "Anomaly-based intrusion detection system through feature selection

analysis and building hybrid efficient model"; Journal of Computational Science, Available online 22 March 2017, Elsevier2017. Page 1- 9

[6]. Srinivas Mukkamala, Guadalupe Janoski, Andrew Sung: "Intrusion Detection Using Neural Networks and Support Vector Machines"; Proceedings of the International Joint Conference 2002 - IJCNN'02 on Neural Networks, 2002, IEEE 2002, Pages 1702- 1707

[7]. Sannasi Ganapathy, Kanagasabai Kulothungan, Sannasy Muthurajkumar, Muthusamy Vijayalakshmi, Palanichamy Yogesh & Arputharaj Kannan: "Intelligent feature selection and classification techniques for intrusion detection in networks: A survey"; EURASIP Journal on Wireless Communications and Networking (A Springer Open journal), 2013, Volume 2013, Issue 01, Artical 271, Page 01- 16.

[8]. Mohammad Sazzadul Hoque, Md. Abdul Mukit and Md. Abu Naser Bikas: "An Implementation of Intrusion Detection System Using Genetic Algorithm"; International Journal of Network Security & Its Applications (IJNSA), Vol.4, No.2, March 2012, Pages 109- 120

[9]. B.A. Fessi, S. Ben Abdallah, M. Hamdi and N. Boudriga: A New Genetic Algorithm Approach for Intrusion Response System in Computer Networks"; Symposium on Computers and Communications, 5- 8 July 2009, IEEE Xplore 2009, Pages 342- 347, IEEE, 2009.

[10]. S. Devaraju and Dr. S. Ramakrishnan: "Performance Analysis Of Intrusion Detection System Using Various Neural Network Classifiers"; IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011, MIT, Anna University, Chennai, Tamil Nadu, India. June 3-5, 2011, IEEE 2011, Pages 1033-1038.

[11]. Mohammed A. Ambusaidi, Xiangjian He, Priyadarsi Nanda and Zhiyuan Tan: "Building An Intrusion Detection System Using A Filter-Based Feature Selection Algorithm" IEEE Transactions on Computers, Oct. 1 2016, Volume 65, Issue 10, Pages 2986 – 2998.

[12]. Xu Yang, Zhao Hui: "Improving the Particle Swarm Algorithm and Optimizing the Network Intrusion Detection of Neural Network"; Sixth International Conference on Intelligent Systems Design and Engineering Applications, 2015, Date of Conference: 18-19 Aug. 2015, Date Added to IEEE Xplore: 02 May 2016, Pages 452- 455.

[13]. A. Gupta, O. J. Pandey, M. Shukla, A. Dadhich, S. Mathur, and A. Ingle: "Computational Intelligence Based Intrusion Detection Systems for Wireless Communication and Pervasive Computing Networks": IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Date of Conference: 26-28 Dec. 2013, Enathi, India, pages 1-7. IEEE, 2013, Pages: 1 - 7

[14]. Fatemeh Kavousi and Behzad Akbari: "Automatic Learning of Attack Behaviour Patterns Using Bayesian Networks"; 6th International Symposium on Telecommunications 2012- (IST'2012), Date of Conference: 6-8 Nov. 2012, Date Added to IEEE Xplore: 21 March 2013, Pages 999-1004.

[15]. Richard Zuech, Taghi M. Khoshgoftaar and Randall Wald: "Intrusion detection and Big Heterogeneous Data: A Survey"; Journal of Big Data (2015), Journal of Big Data (2015), Volume 2, Issue 1, Article 3, December 2015, Page 1-41.

[16]. Chun Guo, Yuan Ping, Nian Liu, Shou-Shan Luo: "A Two-Level Hybrid Approach For Intrusion Detection"; Neuro computing 214 (2016), Elsevier, 2016 page 391–40

[17]. Chi-Ho Tsang, Sam Kwong and HanliWang: "Genetic-Fuzzy Rule Mining Approach And Evaluation Of Feature Selection Techniques For Anomaly Intrusion Detection"; Pattern Recognition Society, Elsevier, 2007

[18]. Sumaiya Thaseen Ikram, Aswani Kumar Cherukuri: "Intrusion detection model using fusion of chi-square feature selection and multi class SVM"; Journal of King Saud University – Computer and Information Sciences (2016), Received 7 July 2015; revised 4 October 2015; accepted 3 December 2015.

[19]. E. Biermann, E. Cloete, L. M. Venter: "A comparison of Intrusion Detection systems"; Computers & Security, Elsevier Science Ltd, 20 (2001) page 676-683.

[20]. TIAN Xin- Guang, GAO Li-zhi, SUN Chun-Lai, DUAN Mi-yi, ZHANG Er-yang: "A Method for Anomaly Detection of User Behaviours Based on Machine Learning"; The Journal Of China Universities of Posts And Telecommunications, Vol. 13, No. 2, Jun. 2006.

[21]. Enamul Kabir, Jiankun Hu, Hua Wang, Guangping Zhuod: "A novel statistical technique for intrusion detection systems"; Future Generation Computer Systems, Elsevier, 2017

[22]. Roshni Dubey, Pradeep Nandan Pathak: "KNN based Classifier Systems for Intrusion Detection"; International Journal of Advanced Computer Technology (IJACT), Volume-2 Issue-4: Published On August 25, 2013.

[23]. Ismail Butun, Salvatore D. Morgera, and Ravi Sankar: "A Survey of Intrusion Detection Systems in Wireless Sensor Networks"; IEEE COMMUNICATIONS SURVEYS & TUTORIALS, ACCEPTED FOR PUBLICATION, IEEE COMMUNICATIONS SURVEYS & TUTORIALS, 2013.

[24]. P. García-Teodoro, J. Díaz- Verdejo, G. Maciá-Fernández, E.Vázquez: "Anomaly-based network intrusion detection: Techniques, systems and challenges"; Computer & Security, Elsevier, Volume 28, Issues 1–2, February–March 2009, Pages 18-28.

[25]. V. Jyothsna, V.V. Rama Prasad: "FCAAIS: Anomaly based network intrusion detection through feature correlation analysis and association impact scale"; The Korean Institute of Communications Information Sciences (KICS), ICT Express, Elsevier (2016) Volume 2, Issue 3, September 2016, Pages 103-116.

[26]. Fangjun Kuanga, Weihong Xua, Siyang Zhang: A novel hybrid KPCA and SVM with GA model for intrusion detection; Applied Soft Computing, Elsevier, Volume 18, May 2014, Pages 178-184.

[27]. Avita Katal, Mohammad Wazid, R. H. Goudar D. P. Singh: "A Cluster Based Detection and Prevention Mechanism against Novel Datagram Chunk Dropping Attack in MANET Multimedia Transmission"; Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013), IEEE 2013, Pages 479- 484

[28]. Hamid Bostani, Mansour Sheikhan: "Modification of supervised OPF based intrusion detection systems using unsupervised learning and social network concept"; Pattern Recognition, Elsevier, Volume 62, February 2017, Pages 56–72.

[29]. Salma Elhag, Alberto Fernández, Abdullah Bawakid, Saleh Alshomrani: "On the combination of genetic fuzzy systems and pair wise learning for improving detection rates on Intrusion Detection Systems"; Expert Systems with Applications, Elsevier, Volume 42, Issue 1, January 2015, Pages 193-202.

[30]. Wenying Feng, Qinglei Zhang, Gongzhu Hu, Jimmy Xiangji Huang: "Mining network data for intrusion detection through combining SVMs with ant colony networks"; Future Generation Computer Systems, Elsevier, Volume 37, July 2014, Pages 127-140.

[31]. Seyed Mojtaba Hosseini Bamakan, Huadong Wang, Tian Yingjie, YongShi: "An effective intrusion detection framework based on MCLP/SVM optimized by time-varying chaos particle swarm optimization"; Neuro

computing, Elsevier, Volume 199, 26 July 2016, Pages 90-102.

[32]. Sejal K. Patel, Umang H. Mehta, Urmi M. Patel, Dhruv H.Bhagat, Pratik Nayak and Ankita D. Patel: "A Technical Review on Intrusion Detection System"; International Journal of Computer Science & Engineering Technology (IJCSET), Vol. 6 No. 01 Jan 2015, Pages 17- 22

[33]. Mohammed Anbar, Rosni Abdullah, Iznan H. Hasbullah, Yung-Wey Chong and Omar E. Elejla: "Comparative performance analysis of classification algorithms for intrusion detection system"; 14th Annual Conference on Privacy, Security and Trust (PST), Auckland, New Zealand, 12-14 Dec. 2016, Date Added to IEEE Xplore: 24 April 2017.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

An Optimized Dual Band Antenna for Wireless Body Area Network

Vipin Gupta[1], Mihir Narayan Mohanty*[2]
[1]SSCET,BADHANI, Pathankot, Punjab, India
[2]ITER, SOA University, Bhubaneswar, Odisha, India

ABSTRACT

Recent technological developments and advancements in wireless communication spread over different areas. Body Area Network (BAN) is a system of devices in close proximity to a person's body that cooperate for the benefit of the user. It is still an emerging technology, and as such it has a very short history. BAN technology emerges as the natural byproduct of existing sensor network technology and biomedical engineering. In this paper, a compact dual Band high gain Microstrip Patch antenna is designed for Wireless Body Area Network Applications. A reflecting Surface is added at a distance s from the antenna to enhance the gain. A significant result has been obtained to support the applications such as medical application, communication as well as wireless monitoring.

Keywords : Dual Band Antenna, Wireless Body Area Network, Return Loss, Radiation Pattern.

I. INTRODUCTION

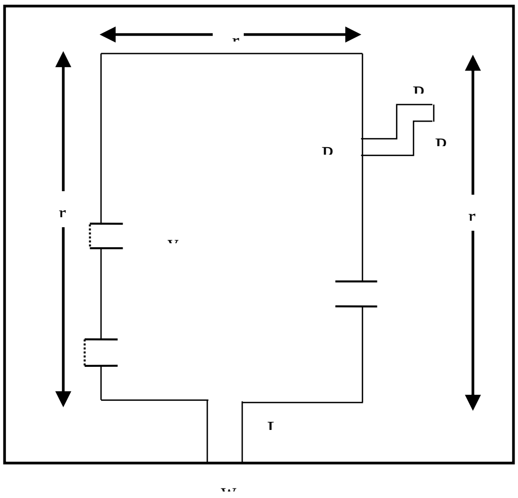
Future applications in the area of telecommunications are being driven by the concept of being connected or able to communicate anywhere and at any time. A Wireless Body Area Network consists of small, intelligent devices attached on or implanted in the body which are capable of establishing a wireless communication link (Chen *et. al.,* 1998). Wireless body area networks (WBAN) are expected to be a breakthrough technology in healthcare areas such as hospital and home care, telemedicine, and physical rehabilitation. Because the human body has a complex shape consisting of different tissues it is expected that the nature of the propagation of electromagnetic signals in the case of WBAN to be very different than the one found in other environments, e.g. offices, streets, etc. The idea is to monitor several vital signs parameters recorded by different sensors placed on the body surface, or even by implanted sensors; and that all signals are collected by a wearable receiver or wireless gateway to transmit the recordings to the doctor.

Continuous, everyday, wearable monitoring and actuating is part of this change. In this setting, sensors that monitor the heart, blood pressure, movement, brain activity, dopamine levels, and actuators that pump insulin, "pump" the heart, deliver drugs to specific organs, stimulate the brain are needed as pervasive components in and on the body. They will tend for people's need of self-monitoring and facilitate healthcare delivery (Durney *et. al.,* 1986 ; Gabriel *et. al.,* 1996 ; Pethig 1987 ; Gaandhi 1990 ; Yazdandoost *et. al.,* 2007).

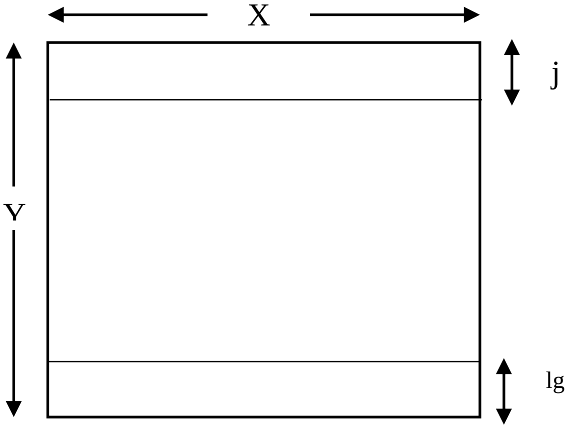
An antenna placed on the surface or inside a body will be heavily influenced by its surroundings (Chen *et. al.,* 1998). The consequent changes in antenna pattern and other characteristics need to be understood and accounted for during any propagation measurement campaign. The human body is not an ideal medium for radio frequency wave transmission. It is partially conductive and consists of materials of different dielectric constants, thickness, and characteristic impedance (Movassaghi *et. al.,* 2014 ; Baek *et. al.,* 2013 ; Sabrin *et. al.,* 2015 ; Karthik *et. al.,* 2017 ; Chamaani *et. al.,* 2015 ; Sukhija *et. al.,* 2017 ; Shakib *et. al.,* 2017). Therefore depending on the frequency of operation, the human body can lead to high losses caused by power absorption, central frequency shift, and radiation pattern destruction. The absorption effects vary in magnitude with both frequency of applied field and the characteristics of the tissue.

II. DESIGN OF PROPOSED ANTENNA

The proposed antenna is designed with the following materials and the dimensions of the antenna before optimization and after optimization is given in Table 1. The proposed geometry is shown in Figure 1. The optimization has been done for the dimensions of the antenna using particle swarm optimization which is inbuilt with the antenna simulation software.



(a)Front View



(b)Back View

Figure 1. Geometry of the proposed antenna (a) Front View (b) Back View

Table 1. Dimension of the parameters of the proposed antenna

Variable	Without Optimization (mm)	With Optimization (mm)
X	30	30
Y	38	38
t	0.035	0.035
h	0.8	0.82
lf	4	4
wf	2.8	2.8
lg	2	1.8487
r	8.5	8.7960
rl	32	32.2593
j	4	4.1005
u	2	2.0818
v	2	2.0732
s	9	8.7877
p	1	1
p1	6.5	6.9707
r1	3	2.85
p2	1	0.9873

Material = RT Duroid 5880 . Semi flexible material. Er= 2.2; loss tan = 0.0009

SAR value is calculated by placing the antenna on a heterogeneous rectangular phantom.

3 layers of phantom are skin, fat and muscle with the thickness of 2mm, 2mm, 6mm.

Dimension of phantom = 60×60 mm$^{2.}$

Resonance frequency = 5.5 Ghz and 7.5 Ghz

III. RESULTS AND DISCUSSION

The results of the proposed antenna is described in this section as it is required to be implanted in the body the SAR value has been evaluated based on the following relation.

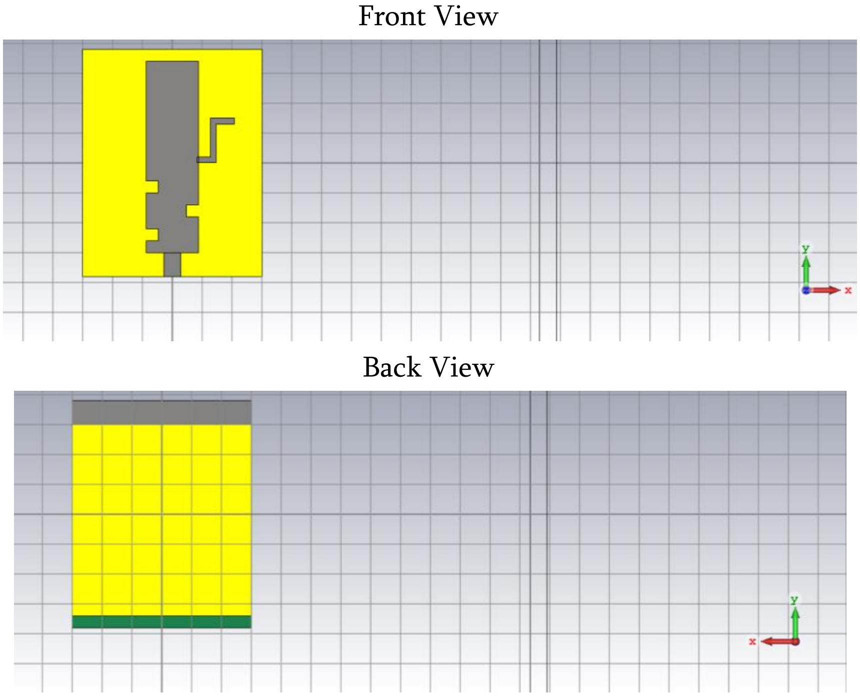
$$\mathrm{SAR} = \int \frac{\sigma(\mathrm{r})|\mathrm{E}(\mathrm{r})|^2}{\rho(\mathrm{r})}\,\mathrm{dr}$$

Where E(R)= electric Field

σ(R) = Conductivity of human tissue

ρ (R) = mass volume density of tissue.

The simulation results for the antenna, radiation pattern and the return loss are shown in Figure 2 through Figure 5. Also it satisfies two bands as 5.5 GHz and 7.5 GHz. The performance table is shown in Table 2. The Simulation software is CS.


Front View

Back View

Figure 2. The proposed Antenna (a) Front View (b) Back View

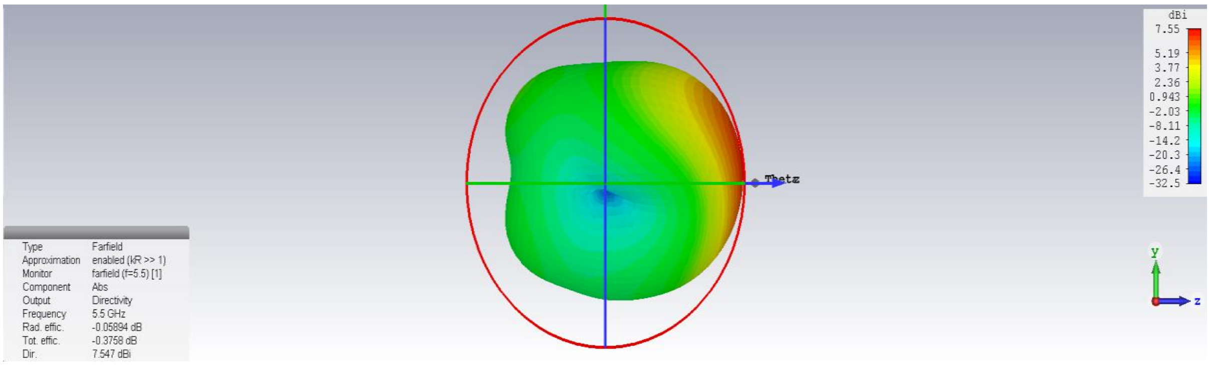

Type Farfield
Approximation enabled (kR >> 1)
Monitor farfield (f=5.5) [1]
Component Abs
Output Directivity
Frequency 5.5 GHz
Rad. effic. -0.05894 dB
Tot. effic. -0.3758 dB
Dir. 7.547 dBi

Figure 3. the 3-D radiation Pattern of the proposed antenna

S parameters

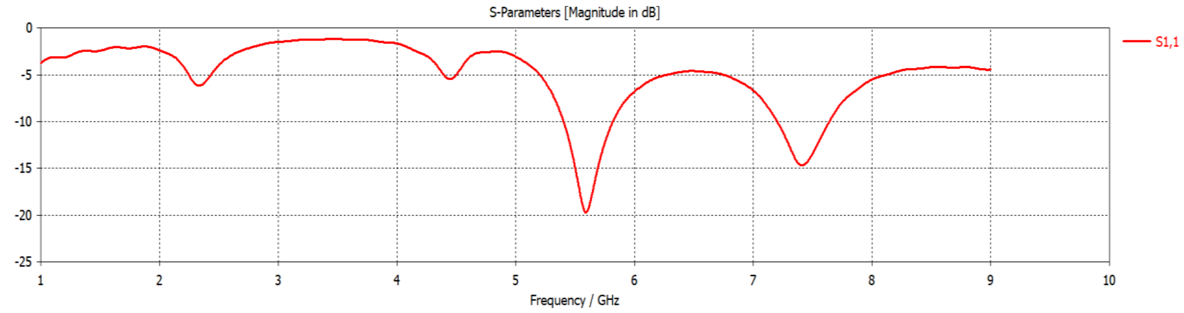


Figure 4. Return Loss of the proposed antenna

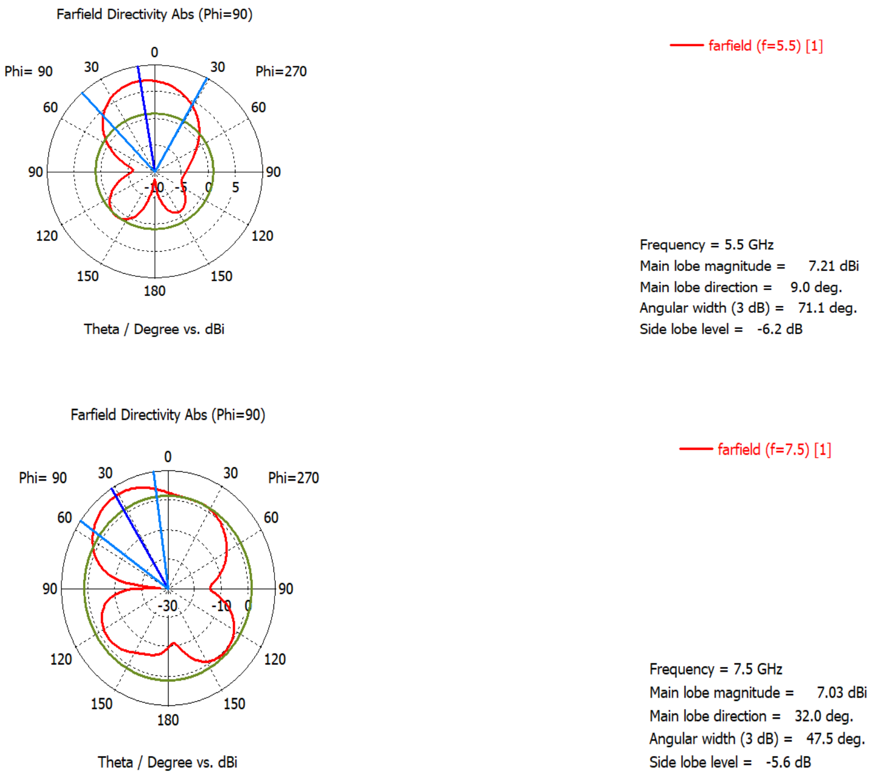


Figure 5. the 2-D radiation Pattern of the proposed antenna

Table 2. Performance Table for dual-band

Parameter	**Frequency (5.5GHZ)**	**Frequency (7.5 GHZ)**
Gain	7.21dBi	7.03dBi
Return Loss	-19.344dB	-14.667dB
Total Efficiency	90% (On Body)	80% (On Body)
VSWR	Less than 2	Less than 2

IV. CONCLUSION

In this piece of work, a dual band antenna has been designed for the use of wireless body area network to achieve two different bands. As Table 2 and Figure 4 two distinct frequency notches have been reflected where the bandwidth is 900 MHz approximately. The antenna is simulated in the platform of CST studio. The proposed antenna is an attractive candidate for wireless body area application. It can be fabricated and optimized using other algorithms to enhance the results.

V. REFERENCES

[1]. Chen, W. T., & Chuang, H. R. (1998). Numerical computation of human interaction with arbitrarily oriented superquadric loop antennas in personal communications. IEEE Transactions on Antennas and Propagation, 46(6), 821-828.

[2]. Durney, C. H., Massoudi, H., & Iskander, M. F. (1986). Radiofrequency radiation dosimetry handbook. UTAH UNIV SALT LAKE CITY DEPT OF ELECTRICAL ENGINEERING.

[3]. Gabriel, C. (1996). Compilation of the Dielectric Properties of Body Tissues at RF and Microwave Frequencies. KING'S COLL LONDON (UNITED KINGDOM) DEPT OF PHYSICS.

[4]. Pethig, R. (1987). Dielectric properties of body tissues. Clinical Physics and Physiological Measurement, 8(4A), 5.

[5]. Gaandhi, O. P. (1990). Biological effects and medical applications of electromagnetic energy.

[6]. Yazdandoost, K. Y., & Kohno, R. The Effect of Human Body on UWB BAN Antennas. IEEE802, 15-07.

[7]. Yazdandoost, K. Y., & Kohno, R. (2007, December). Wireless communications for body implanted medical device. In Microwave Conference, 2007. APMC 2007. Asia-Pacific (pp. 1-4). IEEE.

[8]. Movassaghi, S., Abolhasan, M., Lipman, J., Smith, D., & Jamalipour, A. (2014). Wireless body area networks: A survey. IEEE Communications Surveys & Tutorials, 16(3), 1658-1686.

[9]. Sabrin, S., Morshed, K. M., Rahman, M. M., & Karmokar, D. K. (2015, December). A compact, wide-beam, high gain antenna for wearable medical body-area network devices. In Electrical and Computer Engineering (WIECON-ECE), 2015 IEEE International WIE Conference on (pp. 267-270). IEEE.

[10]. Karthik, V., & Rao, T. R. (2017). Investigations on SAR and thermal effects of a body wearable microstrip antenna. Wireless Personal Communications, 1-17.

[11]. Chamaani, S., & Akbarpour, A. (2015). Miniaturized dual-band omnidirectional antenna for body area network basestations. IEEE Antennas and Wireless Propagation Letters, 14, 1722-1725.

[12]. Sukhija, S., & Sarin, R. K. (2017). Low-profile patch antennas for biomedical and wireless applications. Journal of Computational Electronics, 16(2), 354-368.

[13]. Shakib, M. N., Moghavvemi, M., & Mahadi, W. N. L. B. W. (2017). Design of a Tri-Band Off-Body Antenna for WBAN Communication. IEEE Antennas and Wireless Propagation Letters, 16, 210-213.

[14]. Baek, J. G., & Hwang, K. C. (2013). Triple-band unidirectional circularly polarized hexagonal slot antenna with multiple L-shaped slits. IEEE Transactions on Antennas and Propagation, 61(9), 4831-4835.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Optimized Pulse Shaping for Efficient UWB Communication

Jayant K. Rout, Mihir Narayan Mohanty*
Department of Electronics and Communication Engineering ITER, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India

ABSTRACT

Federal Communications Commission (FCC) guidelines on UWB spectral mask, suitable design of pulses is required to minimize the interference on other existing narrowband services due to UWB systems and vice versa. In this work, we attempt a pulse shaping method, considering the channel as a raised-cosine filter and futher optimized using genetic algorithm (GA). The pulses designed are not only meet the constraints of FCC for UWB, but also spectrally efficient. Pulses generated have a very short duration to reduce multipath interference and are mutually orthogonal. Average bit error rate (ABER) performance of time-hopping binary phase shift keying (TH-BPSK) UWB systems has been evaluated in presence of multiple access interference (MAI) and additive white Gaussian noise (AWGN) for variety of different scenarios, such as, number of users, number of slots per frame and modulation types.

Keywords : Ultra Wide Band, Impulse Radio, Pulse Shaping, GA, UWB Multiuser, Average Bit Error Rate.

I. INTRODUCTION

UWB technology has gained popularity worldwide due to its promise of providing very high bit rates in indoor wireless multimedia transmission with low complexity and low power consumption. The FCC has defined UWB communication as, any wireless scheme that occupies a fractional bandwidth greater than 0.2 or have an absolute bandwidth (-10 dB band-width) more than 500 MHz. In 2002 the FCC has introduced restrictions that the UWB systems must operate with their - 10(-20) dB bandwidth of the spectrum (the bandwidth measured between the 10 (20) dB down points of the spectrum) for indoor (outdoor) communications, exciting a maximum power spectral density (PSD) of -41.3 dBm/MHz in the range of 3.1 GHz to 10.6 GHz (FCC Document 00-163,2002). This new constraint of FCC motivates, that make an investigation into pulse shaping for UWB communication systems. In UWB systems, a train of very narrow pulses of nanosecond or 100s of picosecond is used for data transmission, whose PSD covers the entire available frequency band 3.1-10.6 GHz, for which it is also called impulse-radio(IR) system. UWB system shares the frequency band, where other co-exist narrow-band systems like GPS, PCS, Bluetooth, Wireless LAN (IEEE 802.11a) are already operating. In order to avoid interference to, and from, these co-existing narrow-band services, multi-band or-thogonal frequency division multiplexing (MB-OFDM) can be approached (Zhang, D., Fan, P. and Z. Cao, 2004). But in case of IR-UWB spectral notching is another method that can be used to limit the interference by keeping the PSD of the UWB pulse very low at the NBI occurrence frequencies (Wang, Y., Dong, X. and Fair, J., 2007; Noorodin K .S., Saeed, H., Hamidi, M. and Omali, M. G., 2010).

The PSD of commonly used Gaussian pulse violates the FCC constraints for UWB. In order to meet the FCC requirements, it requires frequency shift, which in turn increases the circuit complexity. Several pulse shaping methods include the combination of Gaussian pulse (Liu, X. , Premkumar, A. and Madhukumar, A. ,2008), modified Hermite pulse (Fan, K., Zhang, S. and Liu, X., 2008; Wenke, L., Hongsheng, S. and Gaoming, L. ,2009) and Optimal pulses (Luo,X., Yang,L. and Giannakis, G. ,2003; Rezaii, M.,2010) have been proposed for UWB communication whose spectrum fits closely to the FCC UWB spectral mask. M. Rezaii (2010) a digital FIR filter approach is proposed for synthesizing UWB pulses that satisfy the spectral mask. B. Parr *et.al* proposed a method using prolate spheroidal wave function (PSWF), which generates a set of orthogonal spectrally compliant pulses. In this work, we used the method proposed in (Parr, B., Cho, B., Wallace, K. and Ding, Z., 2003) and treating the channel as a raised-cosine filter a set of orthogonal pulses are generated. Utilizing this pulse shape, ABER performance of TH-BPSK system is evaluated.

The remaining content of this paper is organized as follows. In Section 2, we present the pulse shaping method, which generates a set of spectrally efficient and orthogonal pulses. The ABER performance evaluation is discussed in Section 3. In Section 4 simulation results are presented. Finally, the conclusions drawn from this work are presented in Section 5.

II. UWB PULSE SHAPING

For any designed pulse, we require its power to concentrate in the 3.1-10.6 GHz frequency band, for an efficient spectrum utilization. Specifically we wish to design a pulse $\Psi(t)$ that is time limited to T_m seconds while exhibiting minimal distortion as it pass through the filter with impulse response $h(t)$. $H(f)$ is a desired frequency mask, it's corresponding impulse response is $h(t)$. The output of the filter is the convolution of $\Psi(t)$ and $h(t)$, i.e.

$$\lambda\Psi(t) = \int_{-\infty}^{\infty} \psi(\tau)h(t-\tau)d\tau = \int_{-T_m/2}^{T_m/2} \psi(\tau)h(t-\tau)d\tau$$

Where λ is the attenuation factor. A discretization of the above equation by sampling N times in pulse duration T_m can be expressed as

$$\lambda\Psi(n) = \sum_{m=-N/2}^{N/2} \psi[m]h[n-m], n = -\frac{N}{2}.......\frac{N}{2}$$

where n and m take on integer values only and can be written in matrix form as

$$\lambda\Psi = H\Psi$$

where

$$H = \begin{pmatrix} h[0] & h[-1] & \cdots & h[-N] \\ h[1] & h[0] & \cdots & h[-N+1] \\ \vdots & \vdots & \cdots & \vdots \\ h\left[\frac{N}{2}\right] & h\left[\frac{N}{2}-1\right] & \cdots & h\left[-\frac{N}{2}\right] \\ \vdots & \vdots & \cdots & \vdots \\ h[N] & h[N-1] & \cdots & h[0] \end{pmatrix}$$

and

$$\Psi = \begin{pmatrix} \psi\left[-\frac{N}{2}\right] \\ \psi\left[-\frac{N}{2}+1\right] \\ \vdots \\ \psi[0] \\ \vdots \\ \psi\left[\frac{N}{2}\right] \end{pmatrix}$$

From the above equation, it is clear that Ψ is an eigenvector of H.

So every solution of $\lambda\Psi(t)$ can be found by means of the eigenvalue decomposition of $\lambda\psi$. The greater the eigenvalue, the better the power spectrum fits inside the desired frequency mask $H(f)$. Therefore, only the eigenvectors corresponding to large eigenvalues should be taken as pulse design. For the case when, $h[k]=h[-k]$ and is real for all k, then H is a Hermitian matrix. The eigenvalues of a Hermitian matrix are real, and the corresponding eigenvectors of distinct eigenvalues are orthogonal. So a set of orthogonal pulses can be generated. These multiple orthogonal pulses provides PSM for the multiple user access without interference.

B. Parr *et.al.* have taken the filter $H(f)$ as a ideal band-pass filter having a pass band between 3.1-10.6 GHz (Parr, B., Cho, B., Wallace, K. and Ding, Z., 2003). But instead of considering the channel as an ideal band-pass filter, we consider it as a raised-cosine pulse shaping filter, which is more practical. The impulse response and transfer function of the raised-cosine filter are expressed as:

$$h_{R.C}(t)=\sin c\left(\frac{\pi t}{T_m}\right)\frac{\cos\left(\frac{\pi\alpha t}{T_m}\right)}{1-\frac{4\alpha^2t^2}{T_m{}^2}},$$

$$H_{R.C}(f)=\begin{cases} T_m, & |f|\le f_l \\ \frac{T_m}{2}\left[1+\cos\frac{\pi T_m}{\alpha}\left(|f|-f_l\right)\right], & f_l<|f|\le f_h \\ 0, & f_h<|f| \end{cases}$$

where $f_l=\frac{1-\alpha}{2T_m}$, $f_h=\frac{1+\alpha}{2T_m}$ and α is the excess bandwidth parameter, it takes values from 0 to 1.

With $\alpha=0$, the raised cosine filter reduces to the classical Nyquist filter with zero excess bandwidth outside $\frac{1}{T_m}$. But in order to keep the power spectral density of the pulse zero outside the band 3.1- 10.6 GHz for different value of α, the corresponding impulse response of the raised cosine filter will be

$$h_{R.C}(t)=W\sin c(Wt)\frac{\cos(\pi\alpha Wt)}{1-(2\alpha Wt)^2}e^{j2\pi(6.8\times10^9)t},$$

Where $W=\frac{7.5\times10^9}{1+\alpha}$. The eigenvectors $(\psi_1(t),\psi_2(t))$ corresponding to two highest eigenvalues, by the genetic algorithm based optimization and are plotted in Figure 1 for $\alpha=0$ & $\alpha=0.5$ with N = 64 and $T_m=1n\sec$. The corresponding power spectra are shown in Figure 2, which shows that majority of their power is concentrated in the 3.1-10.6 GHz frequency band. Autocorrelation of $\psi_1(t)$ and the cross-correlation between $\psi_1(t)$ and $\psi_2(t)$ are shown in Figure 3. It shows that at the sampling instant autocorrelation of $\psi_1(t)$ has higher value, where as cross-correlation between $\psi_1(t)$ and $\psi_2(t)$ is zero. It is advantageous in determining multipath energy because of the higher autocorrelation value at the sampling instant. Also the multiuser interference is reduced due to zero crosscorrelation at the sampling instant.

Optimization Using GA

A genetic algorithm works like the genetic evolution process and survival for fittest procedure. The fitness function can be formulated as

O=max [□$\psi_i(t)$]

The optimal eigen values are used in the model for pulse shaping and to reduce the error simultaneously.

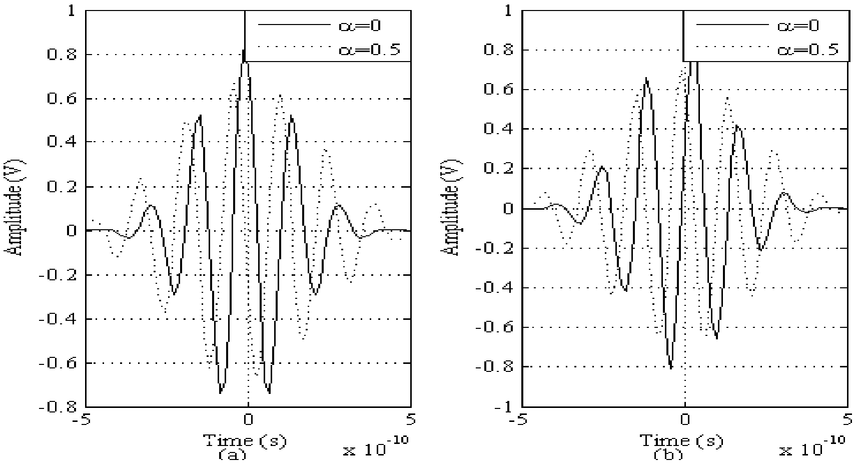


Figure 1. Pulses (a) $\psi_1(t)$ and (b) $\psi_2(t)$ obtained from pulse design method.

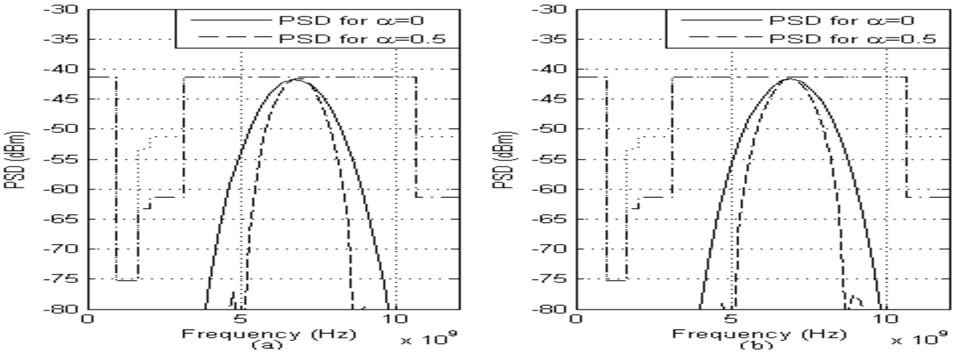


Figure 2. PSD of pulse shapes (a) $\psi_1(t)$ and (b) $\psi_2(t)$

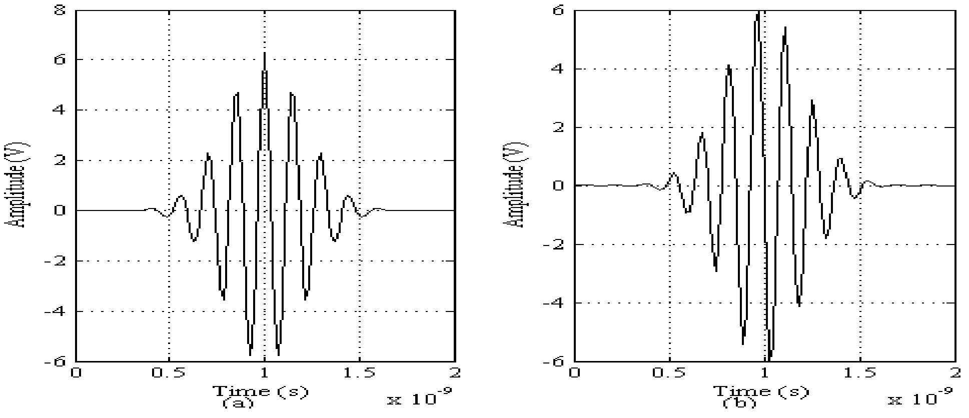


Figure 3. (a) Autocorrélation of $\psi_1(t)$ and (b) Cross-corrélation between $\psi_1(t)$ and $\psi_2(t)$

III. PERFORMANCE EVALUATION

In this section we analyzed the ABER performance for TH-BPSK UWB system, utilizing the designed pulse. The transmitted signal corresponding to the i^{th} data bit of the k^{th} user using TH-BPSK is given by (Hu, B. and Beaulieu, N. C., 2004).

$$s_{bpsk}^{(k)}(t,i) = \sqrt{\frac{E_b}{N_h}} \sum_{j=iN_h}^{(i+1)N_h-1} (-1)^{d_i^{(k)}} p\left(t - jT_f - c_j^{(k)} T_c\right),$$

Where

- *p(t)* is the reference pulse with pulse width T_m, defined in $-T_m/2 < t \le T_m/2$ and normalized so that $\int_{-\infty}^{\infty} p^2(t)dt = 1$,
- E_b is the bit energy common to all signals,
- N_h is the number of pulses required to transmit a single data bit, called the length of the repetition code,
- T_f is the time duration of a frame and thus the bit duration $T_b = N_s T_f$,
- $\{c_j^{(k)}\}$ represents the pseudorandom TH code for *k*th source whose elements take an integer value in the range $0 \le c_j^{(k)} < N_s$ where N_s is the number of hopes,
- T_c is the chip width and satisfies $N_s T_c \le T_f$,
- $d_i^{(k)} \in \{0,1\}$ represents the *i*th binary data bit transmitted by the *k*th source and different bits are assumed to be equiprobable,

Assuming N_u cochannel users are transmitting asynchronously in an AWGN channel, the received signal at the receiver input is given as

$$r(t) = \sum_{k=1}^{N_u} A_k s^{(k)}\left(t-\tau_k\right) + n(t),$$

Where

- $n(t)$ is the Gaussian noise (AWGN) with two-sided PSD $N_0/2$,
- $\{A_k\}_{k=1}^{N_u}$ is the channel attenuation suffered by the *k*th user signal,
- $\{T_k\}_{k=1}^{N_u}$ is the time shift which represents user asynchronism.

The received signal can be viewed as the desired user's signal plus interference and noise. Considering $s^{(1)}(t)$ to be the desired signal, the interference signal is due to $(N_u - 1)$ users.

$$\tau(t) = A_1 s^{(1)}(t-\tau_1) + \sum_{k=2}^{N_u} A_k s^{(k)}(t-\tau_k) + n(t)$$

A correlator receiver is used for the demodulation. In this demodulation method, correlator multiplies the received signal by a template pulse shape and integrates the output over the duration of the pulse. The output of the correlator is a measure of the relative time position and polarity of the received pulse and the template pulse. The output of the correlator is passed to sample and hold circuit, which produces the decision variable $Z(iT_f)$, which is the energy collected over N_h frames and is given as

$$Z(iT_f) = \sum_{j=1}^{N_h} y(jT_f),$$

Where $y(jT_f)$ is the output of the ideal correlator receiver for j^{th} frame and can be represented as

$$\begin{aligned} y(jT_f) &= \int_{jT_f + c_j^{(1)}T_c + \tau_1}^{jT_f + c_j^{(1)}T_c + \tau_1 + T_c} r(t)\upsilon\left(t - jT_f - c_j^{(1)}T_c - \tau_1\right)dt \\ &= y_s(jT_f) + y_{\text{int}}(jT_f) + n, \end{aligned}$$

where $\upsilon(t)$ is the template pulse used for correlation for BPSK it is same as the pulse $p(t)$. Note that *p* (*t*) is the transmitted pulse shape $\psi_1(t)$. $y_s(jT_f)$ is output of the matched filter without any noise and interference signal, $y_{\text{int}}(jT_f)$ is the total MAI due to all $(N_u - 1)$ interfering signals.

The decision block takes the decision on received symbol using the decision rule defined below. Where $\hat{d}_i^{(1)}$ is the estimate of the i^{th} information symbol sent by user 1 $d_i^{(1)}$.

$$\hat{d}_i^{(1)} = \begin{cases} 0, & Z(iT_f) > 0 \\ 1, & Z(iT_f) < 0 \end{cases}$$

Then the probability of error is

$$p(e) = \frac{1}{2} p\left(Z_n(iT_f) < -Z_s(iT_f) \middle| d_i^{(1)} = 0\right)$$

$$+\frac{1}{2}p\left(Z_n(iT_f) > -Z_s(iT_f)\middle|d_i^{(1)} = 1\right).$$

IV. SIMULATION RESULTS AND DISCUSSIONS

We generate a random message using MATLAB, which is used to modulate the pulse shape $\psi_1(t)$. Then a randomly generated $N_u - 1$ interfering pulses of equal amplitude randomly started over the frame duration is added with the desired signal along with a AWGN of proper variance. This received mixed signal is used at the correlator receiver for detection. This process has repeated for various combinations of SNR values, number of users, number of slots per frame and modulation types. The parameters used for performance evaluation are listed in Table I. For a variety of different scenarios, the ABER results are given in Figure 6 and Figure 7.

From the ABER plots it is observed that, as N_u increases, ABER also increases and the rate of increase of ABER slows down for higher values of N_u. In other words, increasing the number of users in the system has a greater impact on the system with a smaller N_u.

It is observed that, ABER is decreasing with increasing number of slots N_s. So this increases the range of SNR values unaffected by multiuser interference i.e. multiuser ABER gets very close to the single-user ABER for higher SNR values. Looking into the relation between the SNR and multiuser interference, when SNR is low, the ABER curves are packed closely together; it indicates that, multiuser interference has little effect for lower SNR values. But with increase of SNR the impact of AWGN on system performance decreases, and multiuser interference becomes more dominant on system performance. The multiuser ABER reaches a floor which is determined by the number of users and the number of slots per frame in the system. So N_s should be carefully selected to provide an acceptable ABER for the maximum number of users in the system.

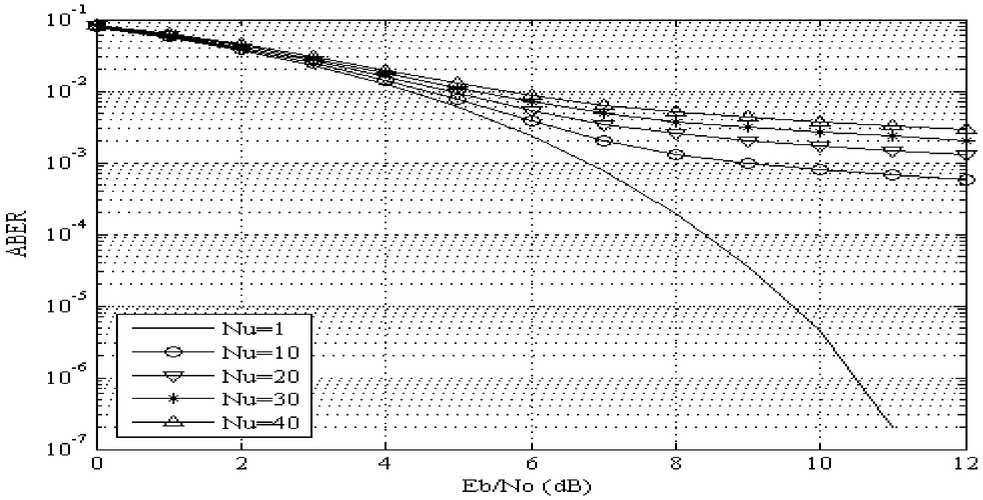


Figure 4. ABER for TH-BPSK UWB system with $N_s = 100$

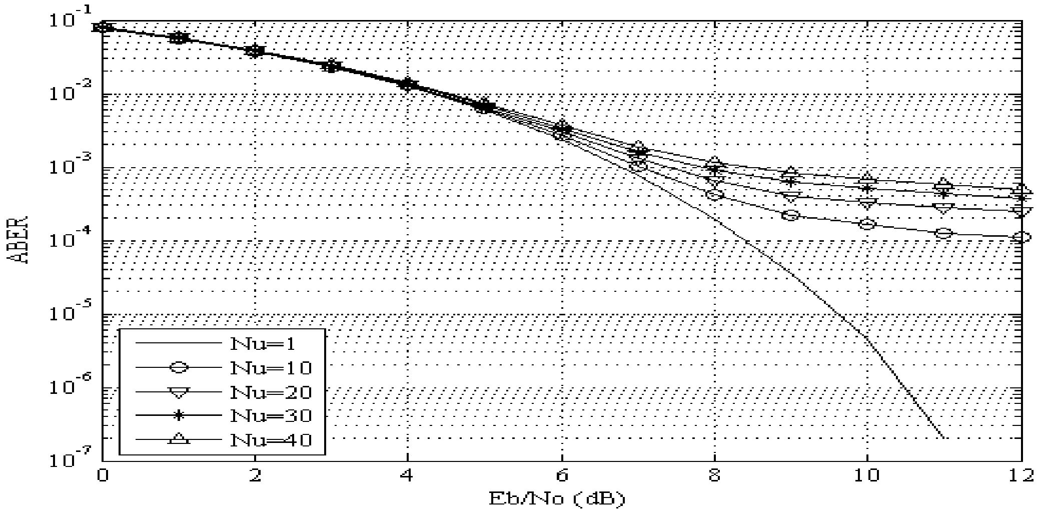


Figure 5. ABER for TH-BPSK UWB system with $N_s = 500$

V. CONCLUSION

In this work we presented a pulse shaping method that generate a set of orthogonal pulses for UWB systems. The power spectra of these pulses shows that the majority of their power is concentrated in the 3.1-10.6 GHz frequency band. The pulse designed by this method has the following advantages over the commonly used Gaussian monocycle. First, the pulses designed are orthogonal and can be used in multiple access schemes. Secondly, the algorithm provides the flexibility of designing pulses to fit frequency masks with multiple pass bands, so that the coexisting narrow band interferences can be avoided.

We presented the performance analysis in-terms of ABER for TH-BPSK multiple access systems, based on different pulse shapes and number of users. It has been observed that the MAI can be reduced by increasing the number of slots per frame at the cost of reduced data rate. This performance analysis method can be extended to any other pulse shapes.

VI. REFERENCES

[1]. FCC Document 00-163: Revision of part 15 of the Commissions Rules Regarding Ultra-Wideband Transmission Systems, Apr. 2002.

[2]. Zhang, D., Fan, P. and Z. Cao, 2004, Interference Cancellation For OFDM Systems in Presence of Overlapped Narrow Band Transmission System. IEEE Trans. Consum. Electron., 50, 108 – 114.

[3]. Wang, Y., Dong, X. and Fair, J., 2007, Spectrum Shaping And NBI Suppression In UWB Communications. IEEE Trans. Wireless Commun..6, 5, 1944 – 1952.

[4]. Noorodin K .S., Saeed, H., Hamidi, M. and Omali, M. G., 2010, A Novel UWB Pulse Waveform Design Method. Int. Conf. on Next Generation Mobile Appl., Services and Tech., 168 – 173.

[5]. Liu, X. , Premkumar, A. and Madhukumar, A. ,2008, Pulse Shaping Functions For UWB Systems. IEEE Trans. Wireless Commun., 7, 5, 1512–1516.

[6]. Fan, K., Zhang, S. and Liu, X., 2008, A UWB Pulse Shapes Modulation Scheme Based On Modified Hermite Polynomials. Conf. on

Wireless Commun., Net. and Mobile Computing. 1 – 4.

[7]. Wenke, L., Hongsheng, S. and Gaoming, L. ,2009, A Novel Pulse Shaping Method For Ultra-Wideband Communications. Conf. on Wireless Commun., Net. and Mobile Computing.1 – 4.

[8]. Luo,X., Yang,L. and Giannakis, G. ,2003, Designing Optimal Pulse-Shapers for Ultra-Wideband Radios. J. Commun. Netw.. 5, 4, 349–353.

[9]. Rezaii, M., 2010, UWB Pulse Shaping by FIR Filter to Enhance Power Efficiency. IEEE Int. Symp. Wireless Pervasive Computing (ISWPC). 522 –527.

[10]. Parr, B., Cho, B., Wallace, K. and Ding, Z., 2003, A Novel Ultra-Wideband Pulse Design Algorithm. IEEE Commun. Letter. 7, 5, 219–221.

[11]. Hu, B. and Beaulieu, N. C., 2004, Accurate Evaluation of Multiple-Access Performance in TH-PPM and TH-BPSK UWB Systems. IEEE Trans. Commun.. 52,10, 1758–1766.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 4 | Issue 4 | ISSN : 2456-3307

Implementation of Fuzzy Logic Measures for Mobile Robot Control

B. M. Bhairat[1],M. R. Gosavi[2], V. M. Thakare[3]

[1]Department of Mathematics, Br. Balasaheb Khardekar College, Vengurla, Sindhudurg, Maharashtra, India
[2]Department of Mathematics, Maharashtra Mahavidyalaya, Nilanga, Latur, Maharashtra, India
[3]Department of Computer Science and Technology, Sant Gadage Baba Amaravati University, Amaravati, Andhra Pradesh, India

ABSTRACT

Mobile robots are required to navigate in unknown and dynamic environments and in the present years the use of mobile robots in material handling has increased. In this work, on-line navigation for autonomous mobile robot in dynamic and unknown indoor environment using fuzzy logic measures is investigated. Four fuzzy logic measures are developed and used to navigate robot to its target. Goal Seeking Measure (GSM), Static and Dynamic Obstacles Avoidance Measure (SDOAM), Emergency Measure (EM), and Robot Setting Measure (RSM) are used to navigate mobile robot and obstacle avoidance. The target of this work is to use the autonomous mobile robot in warehouse with dynamic unknown environment.

Keywords : Autonomous Mobile Robot, Fuzzy Logic, Wheeled Mobile Robot, Robot Navigation

I. INTRODUCTION

Many researchers have anticipated that mobile robots will take charge in various tasks in manufacturing plants, warehouses and construction sites. Recently, the use of mobile robots in material handling applications has considerably increased. For instance, in the warehouses, the mobile robots are used for material handling from stockrooms and also monitoring the inventory of different items. The dynamic environment and the insufficient information on the environment are the main challenge in the navigation operation of the WMR. The conventional mobile robot planning approaches remain not strong and unable to overcome these challenges. As a result, many reactive approaches were introduced allowing the use of artificial intelligence techniques, where problem solving, learning and reasoning are the main issues. Within this scope, fuzzy logic [1], neural networks and other artificial intelligence techniques [2], became the basis of navigation systems in mobile robots.

Even though existing of these approaches, navigation of autonomous mobile robot is still an open area. The challenges are in the unstructured, dynamic and unknown environment. All approaches tried to solve the problems of navigation within one or two complicated measures. In order to overcome the navigation problems and challenges, we distribute the behaviors of mobile robot navigation and dynamic obstacle avoidance between various fuzzy logic measures. In this work, the author used the fuzzy logic technique with four measures in order to navigate an autonomous mobile robot in unstructured, dynamic and unknown environment. The contribution of this paper based on distributing the behaviors of mobile robot navigation and dynamic obstacle avoidance between four fuzzy logic measures (GSM, SDOAM, EM,RSM) and switching the control between them. The author used Powerbot robot as a mobile robot platform to check the effectiveness of the proposed algorithms. This paper is organized as follows. In Section II, a previous work done is presented. Existing Methodology is explained in section III. Fuzzy logic description is presented in section IV. Section V explains the proposed methodology. Experimental results are presented in section VI. Conclusion is given in section VII.

II. PREVIOUS WORK DONE

Various self-control techniques, such as fuzzy logic, neural network, and genetic algorithm are used to deal with dynamic and unknown environment. M. Cao et. al. [1] describes multiple types of inputs: sonar, camera and stored map with fuzzy logic system which is used to navigate the mobile robot. R. Rashid et. al. [2] explains fuzzy logic for indoor navigation. Nabeel K. Abid et. al.[3] describes how fuzzy logic control FLC can be applied to sonar of mobile robot. The fuzzy logic approach has effects on the navigation of mobile robots in a partially known environment that are used in different industrial and society applications. The fuzzy logic provides a mechanism for combining sensor data from all sonar sensors which present different information.

Antonio Gómez Skarmetaet. al.[4] describes the application of fuzzy logic to the navigational component of an indoor autonomous system implemented by means of intelligent agents. Oscar Castilloet. al.[5]addresses the problem of trajectory tracking control in an autonomous, wheeled, mobile robot of unicycle type using Fuzzy Logic. The Fuzzy Logic Control (FLC) is based on a back stepping approach to ensure asymptotic stabilization of the robot’s position and orientation around the desired trajectory, taking into account the kinematics and dynamics of the vehicle.

Abraham L. Howellet. al.[6] explains fuzzy logic which is a topic traditionally taught in artificial intelligence, machine learning, and robotics courses. Students receive the necessary mathematical and theoretical foundation in lecture format. The final learning experience may require that students create and code their own fuzzy logic application that solves a real world problem. Mester, Gyula [7] presents the sensor-based fuzzy logic navigation of autonomous wheeled mobile robots in the greenhouse environments. This paper deals with the fuzzy control of autonomous mobile robot motion in unknown environment with obstacles and gives the wireless sensor-based remote control of autonomous mobile robot motion in greenhouse environments using the Sun SPOT technology. K.S. Senthilkumar– [8] presents an Autonomous Mobile Robot (AMR) is a machine able to extract information from its environment and use knowledge about its world to move safely in a meaningful and purposeful manner.

III. EXISTING METHODOLOGY

Multiple types of inputs: sonar, camera and stored map with fuzzy logic system is used to navigate the mobile robot. The authors proposed how to use fuzzy logic control for target tracking control of Wheeled Mobile Robot (WMR). The authors also focused on the navigation without caring about the avoiding the obstacles; they just use FLC for motion the WMR. Tracking Fuzzy Logic Controller (TFLC) is used to navigate the WMR to its target and Obstacles Avoiding Fuzzy Logic Controller (OAFLC) is used to avoid the obstacles. The author’s used the camera and the fuzzy logic to move the robot to its goal. In additional to fuzzy logic control, genetic algorithm and neural network have been used to improve the control scheme. Fuzzy logic control and genetic algorithm are also used to find the optimal parameters for the fuzzy logic.

Navigation system for mobile robot using fuzzy-neural network which explains learning abilities for navigation system using the fuzzy–neural network in dynamic environment. A neuro-fuzzy approach for real time mobile robot navigation is used to tune the membership function parameters. A dynamic neuro-fuzzy system for obstacle avoiding is presented in which robot can reach its target, but it does not show a good interaction behavior.

Genetic algorithm and fuzzy logic control are used to navigate mobile robot. This Genetic is used to tune the fuzzy logic by modifying the shape of membership function. The experimentation result showed that this Geno-Fuzzy system improved the navigation in many case but not for all case of the navigational. Geno-Fuzzy system for mobile robot navigation is used to improve the quality of control system and by finding the optimal parameters of control system.

IV. ANALYSIS AND DISCUSSIONS

Kinematics Model of WMR:

In this research, Powerbot robot is used. Powerbot is aWMR with differential wheels are used. WMR has two driving wheels, which are mounted on forward of the chassis on the same axis and one castor wheel, which is mounted on backward of the chassis. The castor wheel uses to balance the mobile robot during the motion. The kinematic model of this kind of mobile robot described by the following nonlinear equation:

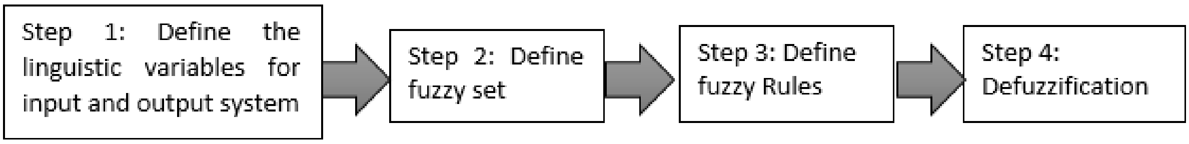
$$\dot{x} = v\cos(\theta)$$

$$\dot{y} = v\sin(\theta)$$

$$\dot{\theta} = \omega$$

where x and y are coordinates of the position of the mobile robot, θ is the orientation of the mobile robot, i.e. the angle between the positive direction X-axis, v is the linear velocity and ω is the angular velocity.

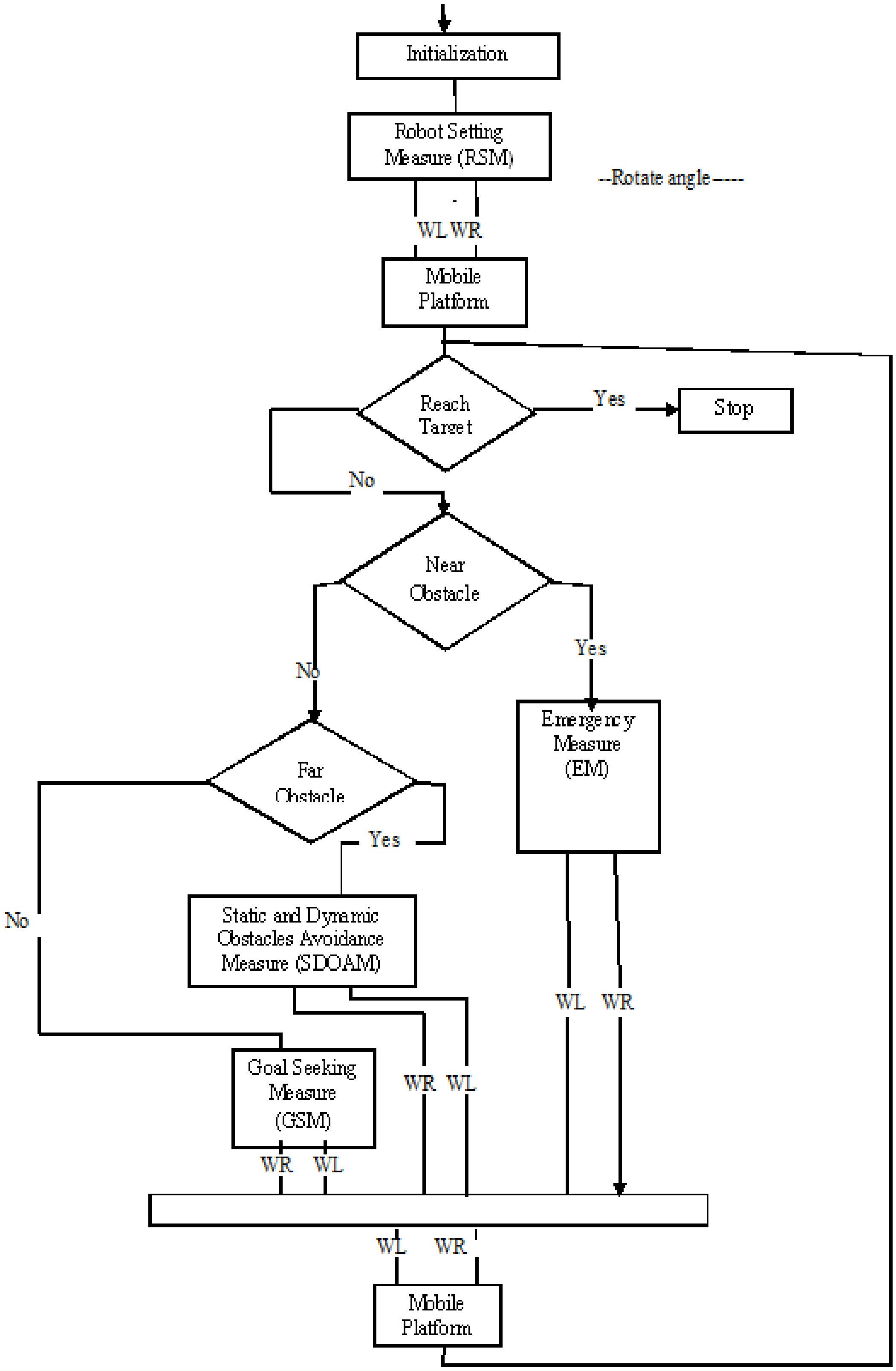
Fuzzy Logic control: -L.A. Zadeh is the father of the fuzzy logic. He introduced the fuzzy logic in 1965, in University of California. Fuzzy logic control has been become an important technique in many areas. In this paper, we use the fuzzy logic technique to implement reaching the target and static and dynamic obstacle avoidance behaviors with mobile robot. In the Fuzzy Logic Process, there are four main steps. First step defines the linguistic variables for input and output system, second step defines the fuzzy set, the third step defines the fuzzy rules and the last step is about of defuzzification. The fuzzy logic process is shown in the following flowchart 1.


Step 1: Define the linguistic variables for input and output system
Step 2: Define fuzzy set
Step 3: Define fuzzy Rules
Step 4: Defuzzification

Flowchart 1

V. PROPOSED METHODOLOGY

Four fuzzy logic measures are developed and used to navigate Powerbot to its target. Goal Seeking Measure (GSM), Static and Dynamic Obstacles Avoidance Measure (SDOAM), Emergency Measure (EM), Robot Setting Measure (RSM) are combined to perform the behaviors of reaching the target and static and dynamic obstacle avoidance.



Initialization
Robot Setting Measure (RSM)
--Rotate angle----
WL WR
Mobile Platform
Reach Target
Yes
Stop
No
Near Obstacle
No
Yes
Emergency Measure (EM)
Far Obstacle
Yes
No
Static and Dynamic Obstacles Avoidance Measure (SDOAM)
Goal Seeking Measure (GSM)
WR WL
WR WL
WL WR
WL WR
Mobile Platform

Flowchart 2

As in the flowchart 2, the algorithm starts with RSM measure. If the Powerbot reaches its target, it stops otherwise the control move to the emergency checking state. If the laser/ultrasonic sensors detect any near (distance between the obstacle and the robot is less than30 cm) movement obstacle the control switches over to the EM measure, otherwise the control checks the SDOAM state. If the laser or the ultrasonic sensors detect any far static or dynamic obstacle (distance between the obstacle and the robot is less than 100 cm and larger 30) the control switches over to the SDOAM measure. The output of GSM, EM, RSM, and SDOAM are the left and right velocities of each wheels of the Powerbot.

A. Goal Seeking Measure (GSM) -GSM uses to simulate the goal seeking behaviors, so it navigates the Powerbot robot to its target. The inputs of GSM are the angle between the direction of the robot to the target and the x-axis (error angle), and the distance between the robot and the target. The outputs of GSM are the velocities of the left and right motors. GSM has been implemented using seven membership functions for both input, see figures 3 and 4 (angle error and distance error). The linguistic variables of the distance error are: Very Far: VF, Far: F, Near Far: NF, Medium: M, Near: N, Near Zero: NZ, and Zero: Z. The linguistic variables of the angle error are: Positive: P, Small Positive: SP, Near Positive Zero: NPZ, Zero: Z, Near Negative Zero: NNZ, Small Negative: SN, and Negative: N. Left Velocity LV and Right Velocity RV of the motors are the output of the GSM. LV and RV in GSM have been implemented using seven membership functions. Figure 5 illustrates the membership of LV and RV. The linguistic variables of the LV and RV are: Z: Zero, S: Slow, NM: Near Medium, M: Medium, NH: Near High, H: High, and VH: Very High.

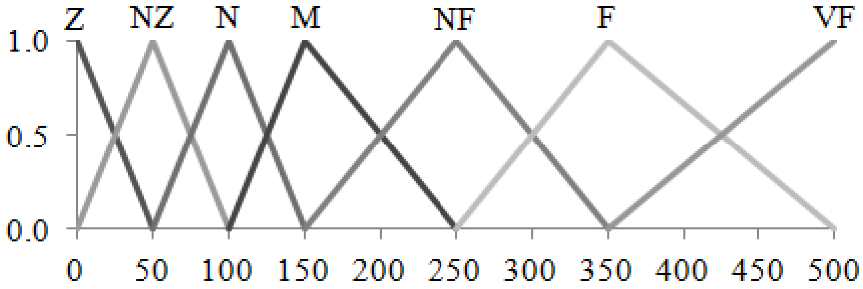


Figure 3. Membership functions for the Distance

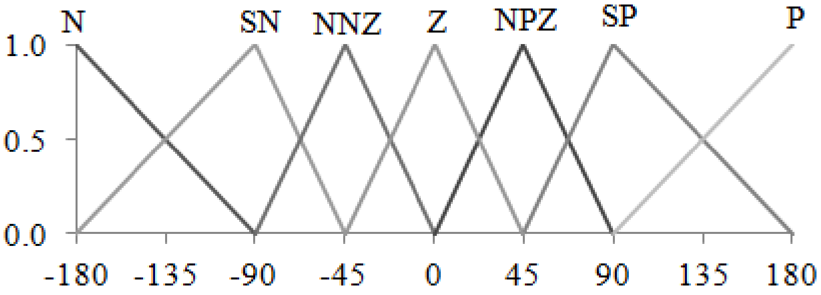


Figure 4. Membership functions for the angle error

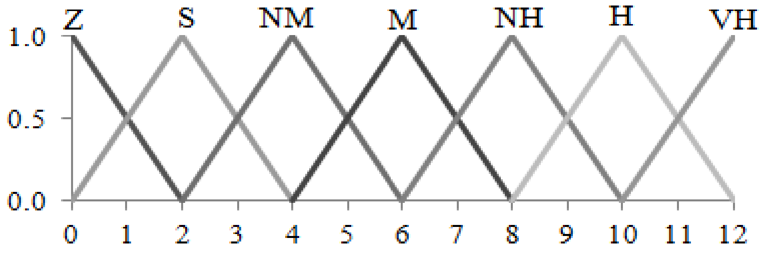


Figure 5. Membership functions of LV and RV

B. Static and Dynamic Obstacles Avoidance Measure *(SDOAM)* - SDOAM uses to simulate the avoiding far obstacles (static or dynamic) behaviors. If its location less than 100 and larger than 30 from the robot, obstacle is far. The inputs to SDOAM are: the distance between the robot and the obstacles (Dis_to_obstacle), the position of the obstacle from the view of the robot (Obs_position) and the different in angle between the target and the obstacle from the robot view (Dif_angle). These distances are acquired using laser device and ultra-sonic sensors. The researcher combines both the laser to take the advantage of its high accuracy and the ultra-sonic to take the advantage of higher coverage area for any obstacle. The linguistic variables of the Dis_to_obstacle are Near: N and Far: F. The linguistic variables of the Obs_position are Lift: L and Right: R. The linguistic variables of the Dif_angle are Small: S and Far: F}.The outputs of SDOAM are the velocities of the left LV and the right RV of the motors. LV and RV in SDOAM have been implemented using three membership functions.

C. Emergency Measure (EM) -EM uses to simulate the avoiding emergency movement behavior (distance between the obstacle and robot is less than 30 cm). The inputs of EM are the distance between the left, front, and right sides of the robot and the obstacles (LD, RD, and FD). These distances, acquired using laser device and ultra-sonic sensors. The author uses the laser to take the advantage of its high accuracy and the ultra-sonic to take the advantage of higher coverage area for any obstacle. The notations for the LD, RD, and FD are: {N: Near and F: Far}. The outputs of EM measure are the velocities of the left LV and the right RV of the motors. LV and RV in EM have been implemented using three membership functions. The linguistic variables of the LV and RV in EM are High Negative: HN, Negative: N, and High:

D. Robot Setting Measure (RSM) - RSM is used to overcome the problem of existence of close intermediate points (if the distance between the robot and the point <50 cm). RSM used to rotate the robot before the motion. The input of RSM is the rotate angle that the robot should rotate. The linguistic variables of the angle are Negative: N and Positive: P. The outputs of RSM are the velocities of the left and right motors. The linguistic variables of the LV and RV in RSM are Forward: FW, Backward: BW.

VI. POSSIBLE OUTCOME AND RESULT

In this work, the author is going to test our proposed method in a real environment with different scenarios. These experimental results will determine the effectiveness and the robustness of the proposed method. In the experimentation part of our work, the author use the Powerbot mobile robot platform, which is developed by Adept Mobile Robots Inc. Powerbot is a differential drive robot for research, which uses C++ as a programming platform language. The proposed methods have been tested using three different environments. In the first scenario, the robot is examined in unknown environment without obstacles. In the second scenario, the robot is examined in unknown environment with static obstacles. In the third scenario, the robot is examined in unknown environment with dynamic obstacles. The author moved the robot from the initial point (inside the robotics laboratory room)to the target point, which is outside of the laboratory room via intermediate points.

VII. CONCLUSION

In this work, author have distributed navigation and obstacle avoidance behaviors between four fuzzy logic measures and switching the control between them. The proposed work aims to use mobile robots for hospital, library and materials handling in instructed warehouse. In this work proposed method is able to navigate the mobile robots in such environment. The proposed method has been tested on PowerBot mobile robot and with three different scenarios, which are close to the scene in warehouse. Depending on the experimental results; the proposed method is effective and robust under varying obstacles scenarios. For the future work, the proposed motion method of robot has to be extended to a swarm of mobile robots.

VIII. REFERENCES

[1]. M. Cao and E. L. Hall, "Fuzzy logic control for anautomated guided vehicle," Intelligent Robots andComputer Vision XVII:Algorithms, Techniques, andActve Vision, vol. 3522, no. 1, pp. 303–312, 1998.

[2]. R. Rashid, I. Elamvazuthi, M. Begam, and M.Arrofiq, "Differential Drive Wheeled Mobile Robot (WMR) Control Using Fuzzy Logic Techniques,"2010, pp. 51–55.

[3]. Nabeel K. Abid Al- Sahib* Ahmed RahmanJasim**, "Guiding Mobile RobotbyApplying Fuzzy Approach on Sonar Sensors", Department of Mechatronics Engineering/Al–Khwarizmi College of Engineering, University of Baghdad * Email: dr_nabeelalsahab@hotmail.com** Email: engmktron@gmail.com Al-Khwarizmi ngineering Journal, Vol. 6, No. 3, PP 36 - 44 (2010)

[4]. Antonio Gómez Skarmeta and HumbertoMartínezBarberá, "Fuzzy Logic Based Intelligent Agents for ReactiveNavigation in Autonomous Systems", Dep. de Informática y Sistemas. Universidad de Murcia

[5]. Oscar Castillo, Luis T. Aguilar, and S´eleneC´ardenas," Fuzzy Logic Tracking Control for Unicycle Mobile Robots", Engineering Letters, 13:2, EL_13_2_4 (Advance online publication: 4 August 2006)

[6]. Abraham L. Howell, Roy T.R. McGrann, and Richard R.Eckert, "Teaching Concepts inFuzzy Logic Using Low Cost Robots", PDAs, and Custom Software, Binghamton, NY 13902 abe@abotics, mcgrann@binghamton.edu, reckert@binghamton.edu, 978-1-4244-1970-8/08/$25.00 ©2008 IEEE October 22-25, 2008, Saratoga Springs, NY, 38th ASEE/IEEE Frontiers in Education Conference T3H-7 t0 T3H-11.

[7]. Mester, Gyula,"Fuzzy-Logic Sensor-Based Navigation of Autonomous Wheeled Mobile Robots in the Greenhouse Environments", GyulaMester is with the Robotics Laboratory, Institute of Informatics, Department of Technical Informatics, and University of Szeged, Hungary Manuscript received June 15th, 2011.

[8]. K.S. Senthilkumar1 K.K. Bharadwaj2, "Hybrid Genetic-Fuzzy Approach to Autonomous Mobile Robot", 1 College of Arts and Science, King Saud University, Wadi Al Dawasir, Riyadh, KSA, 2 School of Computer & Systems Sciences, Jawaharlal Nehru University, New Delhi, India, ksskumar16@gmail.com

[9]. George J. Klir and Bo Yuan, 'Fuzzy Sets and Fuzzy Logic-Theory and Applications' PHI Learning Private Limited, New Delhi (2009).

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Facial Expressions Detection and Recognition Using Neural Networks

Er. Navleen Kour, Dr. Naveen Kumar Gondhi
Department of Computer Science Shri Mata Vaishno Devi University Katra, India

ABSTRACT

Facial Expressions are one of the most robust way of non verbal information exchange in day-to-day life. Changes occurring in emotions of a human being directly efficate the behavior of a person. In this progressive world, a numerous biometric have been evolved, each having its own purport. However, all these statistics play vital job to convey communication from one individual to another but 55% i.e a major part of communication transpire by facial expressions. Facial Expression recognition process concentrates on discerning the changes in expressions of facial muscles that automatically reflect switching of one's mind from one state to another. Humans can recognize Expressions without any effort and almost instantaneously but that are not the case with a machine since its challenges are very dynamic like orientation, lightening, pose, facial expressions, etc.

So, the process of wrenching out or extricating the facial feature points or landmarks is often very challenging. To recognize the fiducial points on the facial features and drawing out these points, that generally lie on eyes corners, chin, eyebrows, etc, facial landmarking is done. Our landmarking technique combines Viola-Jones detection algorithm for feature detection with Harris corner detection and then coarse to fine strategy is implemented using an efficient algorithm. Using the Haar like features reduces the cost of brute force search, also provides advantage of speed. Additional selection of sub-regions is also exploited using anthropometric constraints, to limit the search region. This further reduces false detection rate and improves accuracy significantly. A sub- algorithm named Iterative best fit algorithm is used find a land mark exploiting its commonality and geometric configuration and can be used in other contexts as well. This method is then tested on JAFEE database, Yale database, AT&T database and the database constructed using my own images named as Smart database and this method provides the satisfactory accuracy.

Keywords : Component, Biometrics, Facial Land Marking, Facial Emotions Recognition, Viola-Jones Algorithm, Harris Corner Detection.

I. INTRODUCTION

What we feel, the various situations we are going through, all are reflected efficiently by facial expressions. A facial expression often represents the emotional state that can be seen, visualize and reflect character of a person [20]. In this modern world of science and technology, computer vision, pattern recognition, fingerprinting recognition, bio-metrics, image processing, security, Artificial Intelligence is leading to intelligent machines which has entirely captured our lives in many areas such as surgical assistant [33], offices [34]. Recognizing the facial moods of a person using Machine Learning is gaining high popularity now-a-days. It is the process of detection and recognition of human face and facial expressions or emotions in order to ensure security in many fields. Face detection and expressions recognition have a heavy demand in this era because it is most universal, user friendly and easy accessible system. For the purpose of facial expressions recognition, emotions play a vital role. Emotions represent those internal feelings of a human mind which can't be expressed verbly or by writing. By visualizing the face of a person, one person can easily recognize the mood of another person so that the mood of person can be enhanced easily accordingly using various ways. It is well known concept that each and every technology has its root from some earlier concept. As in the same way, the foundation of facial expression detection is FACS (Facial Action Coding System), that was introduced by Ekman and his friends. Before this system was developed, researchers often relay on human observers for the information but they were not reliable. So this system was developed to improve the robustness and performance of the system. This system generally focuses on determining the alterations that occur in facial muscles. The various muscles associated with human face are as shown in fig1 [1], typically known as Action Units (AU). An action unit the basic unit which is responsible for causing all these internal changes to the facial muscles that in turn cause changes in expressions.

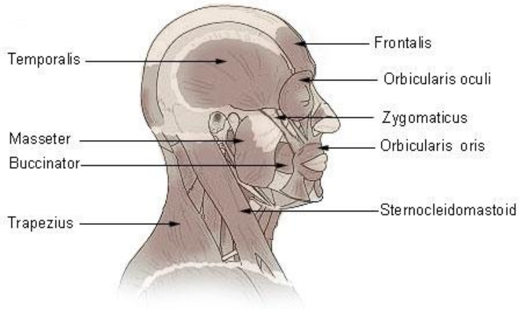


Figure 1. Various types of facial muscles

Facial muscles are divided into two parts. Some of them are related with upper portion of the face and some with lower portion of the face. Depending on different situations, there occur fluctuations in facial muscles. These facial movements are called facial action units (AU's) and the corresponding code is known as Facial Action Coding system [17]. The FACS is composed of 44 facial action units that are the essential part of FACS [1]. Out of them, mainly 30 AUs of them are affiliated to movement of a particular deck of facial muscles that includes: 12 for upper face and 18 for lower face. Facial action units can occur in combinations and vary in intensity. An action unit is in the numerical form used to give detail of all the movements of facial muscles [1, 2].

Recognizing the expressions of a person automatically provides greater support to those applications or areas that very much focuses on face such as coding field, face recollection, expressions/ moods or gesture comprehension, gaze diagnosis, animated applications, face tracing[3,4], face animation[5,6,7], registration[8,9,10], video tracing[27]. Face markers are considered as the renowned hallmark that can play a crucial role and can be treated as bulwark on the entire graph of face. Effects on different landmarks are distinct from one another. Some landmarks such as eyes end points, nose tip are affected on little- scale. So, these are cited as fundamental landmarks. The facial landmarks which are found with the help of the fiducial points are known as ancillary points.

Approaches present today for facial landmark detection can be grouped in two main classes: local and global methods [11]. The global methods [11] are more capable of detecting more landmarks than the local ones, which can mostly detect landmarks quickly. Almost all global methods use either ASM (Active Shape Models) [12] or AAM method (Active Appearance Models) [13]. In the ASM technique, the algorithm make searches in order to obtain the best match using a shape model while as in AAM, the main goal of the algorithm is to obtain the best match with a combined model using texture and shape. In local methods [6], the algorithms are used to detect landmarks such as we can take the example of the eyes end points or the nose peak point, without using information from other parts of the face. There are also some situations in which a combination of global and local method is used where classifiers are trained for different landmark with the Viola & Jones object detection approach.

In this paper, we tender a new facial landmark discernment system to detect landmarks in human faces more frequently and easily. The proposed system is a combination of local and global methods to get a mix of robustness and speed without any complicated pre-processing. It is a coarse to fine strategy which uses haar-like features and corner detection along with geometric constraints to get facial landmarks. The technique is simple and easy in implementation.

II. LITERATURE SURVEY

Facial Expressions recognition is a three phase process such as Face Acquisition, Feature Extraction and Classification as depicted below:-

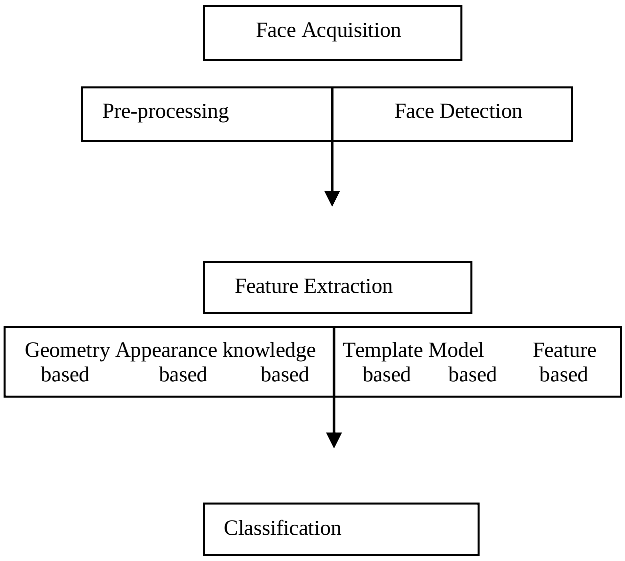


Figure 2. Process Model of Facial Expression Detection

(a) **Face Acquisition:** - is the process of acquiring image. The first step in face acquisition is image pre- processing which involves various techniques such as image cropping, resizing, padding, histogram equalization, etc.

(b) **Feature Extraction:** - Feature extraction is a method in facial expression recognition and involves various methods like dimensionality chopping, trademark extrication and landmark assortment. Feature extraction uses various methods like geomertry based, appearance based, knowledge based, template based, model based and feature based. These methods are planned chiefly for face identification and fiducial points extraction.

(c) **Classification:** - At this stage the extracted feature are used for training the classifier and then testing is done. The various classifiers used are Hidden Markov Model, Neural Networks, Support Vector Machines, Adaboost and Genetic Algorithms.

Thus, over the last few years, there have been numerous face recognition techniques that has been developed and used for the purpose of facial expressions recognition. The whole literature survey is in [35].

III. RESEARCH METHODOLOGY

Facial Expression Detection is a topic with a lot of width and depth, in spite of the amount of research that has been directed

Towards it, there is a lot of work on. Our methodology was to do a literature survey and seek out the areas of improvement.

Our main aim is to make the process more simple and efficient by combining different existing processes. Along with Feature Extraction other parts of the project also had to be developed to get a Facial Moods Recollection System. The basic building blocks of emotion identification system are enlisted below:-

1. Image acquisition: -Acquiring the input image for processing.

2. Image Resizing: - Making the image a square matrix for improved results

3. Image Cropping: - Only keeping the relevant information in the image.

4. Face Detection: - Locating position of face in the given image using pixels of skin.

5. Corner Detection: - Finding corners in the image using sudden change in pixel values.

6. Eyes Detection: - Identification of position of eyes in the face image frame.

7. Mouth Detection: - Determined lip coordinates on the face.

8. Feature extraction: - Extracting token from the image.

9. Emotion Detection: - Emotion is detected by feeding values from database to neural network.

10. Database: - stores values derived from feature extraction.

11. Output Display: - Displays the performance of system on screen.

IV. PROPOSED SCHEME

The proposed method involves the system to detect a set of landmarks in frontal faces. This system is comprised of four major parts. In the first part, there is the pre-processing step and second part has the methodology of detecting features, that is, eyes and mouth in the face. The third part detects the corner in the face and fourth part is the core of the system, where the results of two previous stages is combined and landmark is detected using proposed efficient algorithm called 'Landmarks' and the iterative-best – fit algorithm is also explained there which selects the landmark from a no. of possibilities. A detailed block diagram of the suggested method is depicted by Figure 3 given below

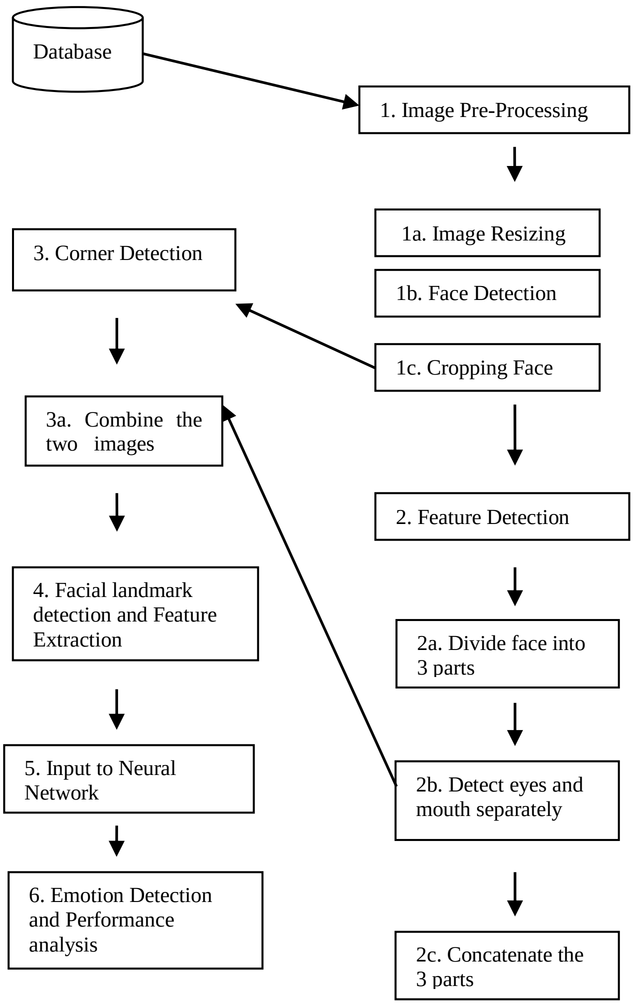

Database
1. Image Pre-Processing
1a. Image Resizing
1b. Face Detection
1c. Cropping Face
2. Feature Detection
2a. Divide face into 3 parts
2b. Detect eyes and mouth separately
2c. Concatenate the 3 parts
3. Corner Detection
3a. Combine the two images
4. Facial landmark detection and Feature Extraction
5. Input to Neural Network
6. Emotion Detection and Performance analysis

Figure 3. Workflow of Proposed Scheme

A. Image Pre-Processing:-

Almost in every system, a database is the most important information storage unit. A number of databases subsist of variety of facsimile that have sundry resolutions, backgrounds and are captured under diverge radiance. Thus it is necessary to perform preprocessing of image. Preprocessing module is induced to forge the images more comparable. In our system, complex pre-processing module is not required and comprises of successive components: image resizing, identification of face and image cropping. The image is resized to get a square image. From this image face area is discovered by employing the Viola-Jones method [15], ground on Haar-like hallmarks and AdaBoost learning technique. We assumed that there is only one face per image. The Viola and Jones practice has giant diagnostic rates and 15% faster than all supplementary approaches. It is principally planned for the muddle of face identification. It requires full view frontal upright faces. It is very robust framework. Its purpose is to provide the differentiation between faces and non- faces. Adaboost is an effective boosting algorithm required to speed up all the types of learning approaches. It predominantly make use of the combination of weak classifiers in order to create the strong one. So, Adaboost is a greedy algorithm and associates large weight with each good features and small weight with poor features after each round examples are re-weighted. The rearmost part of image pre- processing module is cropping the facsimile to get the area enclosed by rectangle of the Viola-Jones object detector for further processing.

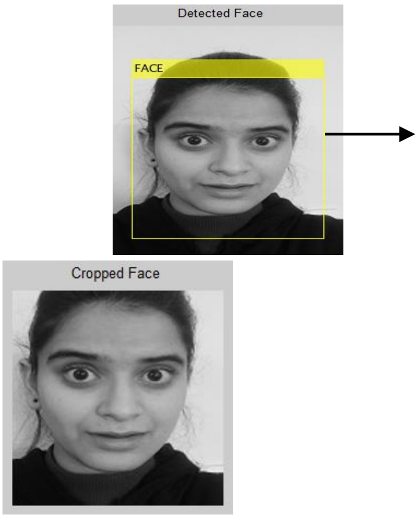


Figure 4. Cropping of face detected by Viola- Jones algorithm

B. Feature Detection:-

This is the second step to reduce the region of interest first being the cropping of face image. In this step, we desire to intuit the position of eyes as well as mouth. The Viola-Jones algorithmic concept is utilized for feature identification. The detected face attained from pre-processing is first divided into three parts such that one part contains the left eye, another contains the right eye and the third part contains the mouth. This is done to overcome the limitation of cascade object detector based on Viola Jones algorithm, which sometimes detects mouth as an eye or vice versa. This method does a multi-scale search and chooses the facial landmark candidates through thresholding [15]. A composite image of the three parts is achieved at the end with detected features. There is the possibility of multiple candidates (multi-size and or different position) for which we choose the candidate with the largest size.

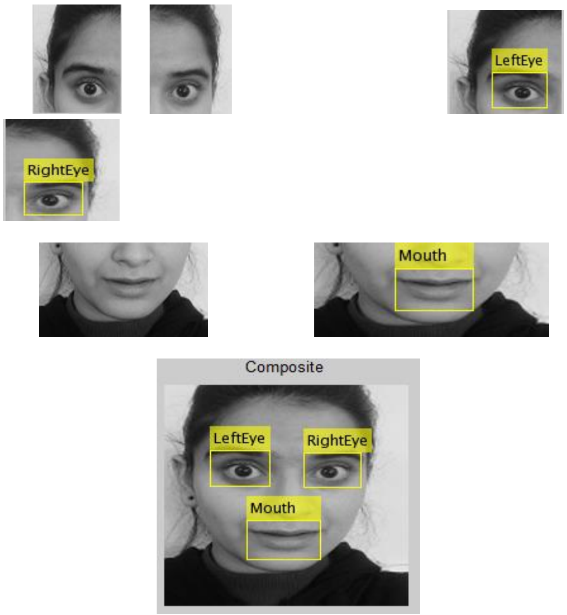


Figure 5. (a) Division of Face (b) feature detection in separate parts (c) Combining the parts of face after detection

C. Corner Detection:-

Identification of corners is an approach employed to wrench out the irrefutable hallmarks and conjecture the contents of an image. Some researchers had applied canny edge detector to calibrate amount and alignment of groove [16] but we have used corner detection. Corner is the intersection of two edges. It is the junction of contours at which intensity changes to larger extent. The significance of this step is that corner detection gives us plausible landmarks as many facial landmarks are corners and by having all corners in an image, landmarks can be chosen from them by further computation. Corner detection projects on the principle that if we spot a small window over an image, if that window is situated at the corner then there is a giant fluctuation in intensity in all the directions and consequently we realize that it must be a corner. If the window is atop the shallow area of the image then there will be no fluctuations in intensity.

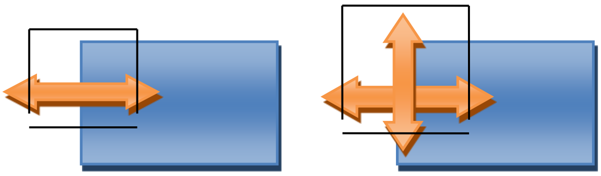


Figure 6. Working of corner detection

Harris Corner Detection algorithm, which is used by us, detects the corners in any given image and is an improvement over Moravec's corner .It is employed to the absolute pre-processed photograph and calculates the points of change in all directions. . Change of intensity for the shift [u, v]:

$$E(u,v)=\sum_{x,y} w(x,y)[I(x+u,y+v)-I(x,y)]^2$$

The function w is the window function, the function I(x + u, y + v) is the switched intensity and the last is the actual intensity. The window function is Gaussian which reduces the noise.

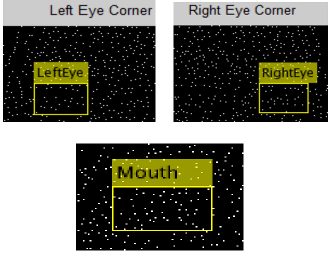


Figure 7. Detected Eyes and mouth corners

D. Facial Landmark Detection:-

This is the final step which is done to select the fiducial points. In this section the prime benefaction of the paper are presented. The results of the feature diagnosing and corner detection are combined to give a resultant image in which the corners detected in the eyes and mouth region are enclosed by a rectangle.

Figure 8. Combination of feature and corner detection

From the corners lying inside the rectangle we have to choose the facial Landmarks. We detect a total of 12 facial landmarks out of which half are primary (left and right landmarks) and (top and bottom landmarks) half are secondary. The procedure to find landmarks are recapitulated for all rectangles demarcating features in the image (w is the width and h is height of the rectangle). The values of x and y have been determined after trying different values and picking the ones which gave best result.

- ALGORITHM:-

<u>Algorithm 1: Landmarks</u>

```
i.   Input the cropped eye image from step 1, dimensions w and h of boundary bounding the feature and values a and b.
ii.  Crop image according to mentioned values of ROI.
iii. Find the positions of corners detected in the image.
iv.  Subtract a from the column value y and b from the row value x of the corner position to get find points close to the desired position.
v.   Call Iterative_Best_Fit_Algorithm to find  best fit right facial landmark from current corner points; // see sub-algorithm 1.
vi.  Return the position (column, row) of chosen Landmark.
```

<u>Sub-Algorithm: Iterative_Best_Fit_Algorithm</u>

```
Step 1: Input x [ ] and y [ ] (from step 5 in algorithm 1).
Step 2: Find value mr = minimum in x [ ].
    Step 3: index_X:= positions in x where x = mr
Step 4: Find value mc = minimum in y [ ].
Step 5: index_Y:= positions in y where y = mc.
Step 6: If common: = intersect (index_X, index_Y) is equal to 0.
          Min_X:= x [index_X (1)] +y [index_Y (1)];
            Index1:= index_X (1)
            Repeat: K =1 to SIZE (index_X)
            TEMP: = x [index_X (K)] +y [index_Y (K)];
            If (MIN_X > TEMP)
                MIN_X: = TEMP; Index1 = K;
            Min_Y:= x [index_X (1)] +y [index_Y (1)];
            Index2:= index_Y (1)
Repeat: K =1 to SIZE (index_Y)
            TEMP: = x [index_X (K)] +y [index_Y (K)];
            If (MIN_Y > TEMP)
                MIN_Y: = TEMP; Index2 = K;

            If (Min_X >= Min_Y)
                  Common: = Index2;
            Else
                  Common: = Index1;
             Step 7: row: = x [Common]; column: = y [Common];
Step 8: RETURN (column, row)
```

E. Input to neural network:

Distortation mechanisms are employed to facial emotions identification include DCT [37], Gabor wavelets [24] [25], Neural Networks [23, 29, 36] and Active Appearance Models. In our thesis, After landmark detection, input is given to the neural network in the form of training images. A label is assigned to each image such as for sad expression, it is taken 0000, for neutral- 0001, for happy- 1111, etc.

F. Emotion detection and Performance Analysis:-

Once the neural network is trained properly, the images are then tested using different databases one by one and the expressions are detected in the form of performance rate or accuracy.

V. IMPLEMENTATION

MATLAB is a multi feature programming language, emerged from Math Works in 1984. In our research work, we used Matlab for the implementation of various types of databases having different properties. Our proposed algorithm for the detection of facial landmarks was first tested using the Smart database. This expressions database contained 28 images of six primer facial expressions i.e. happy, sad, neutral, surprise, angry and fear.

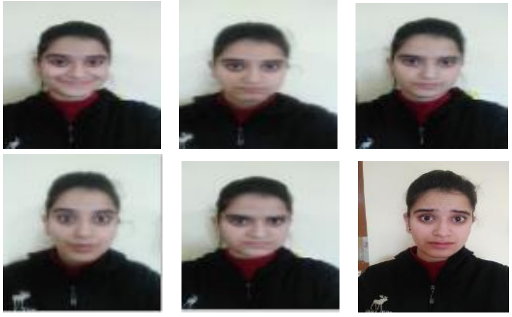


Figure 9. Smart Database

Each image was of size 256*256 pixels and in the jpg format. The images of this database are taken with an OPPO camera (CPH1701). The camera was positioned straight facing the motif. The subject performed different facial displays and these displays are based on descriptions of prototypic emotions. We have detected about 6 facial expressions, so we had used 20 images of this database for training our neural network and 8 images for testing our neural network .We had follow the above proposed algorithm and 12 facial landmarks are detected and then selected as depicted by following figures:-

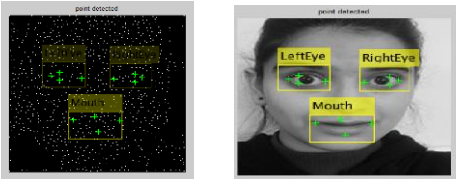


Figure 10. (a)Landmarks detected (b)Detected Landmarks mapped to the face.

Following this, the same algorithm was tested on three more databases i.e JAFFE database, Yale database and AT&T database. The Jaffee database is a Japanese Female Facial Expression database [31] encompassing 213 images of 7 facial expressions in distinct poses. 13 images of this database are used to train the neural network and 7 images are used to test the neural network. Each image in this database is of size 256*256 pixels. Yale face database embody 165 greyscale images in GIF format of 15 persons. 21 images are used for training and 17 photographs for testing the neural network. In AT&T database, there are 10 pictures of 40 dissimilar subjects. All the images are in PGM format and the size of every individual image is 92*112 pixels.

VI. RESULTS AND ANALYSIS

To determine the facial landmarks in Smart database, every image is checked and each landmark is given one of the three labels- right, bearable or wrong. Right is given if landmark is detected at the correct position. Bearable is given if there is slight change in the position Wrong is for the landmark which is entirely at the wrong position and leads to incorrect results when used. We have gone through each image and keenly observed the corner points available from which the landmarks are chosen, to label them into a category.

The results are as follows:

Table1. Analysis of different landmarks

S.No.	LANDMARKS	CORRECT	BEARABLE	WRONG
1.	Left (eyes)	86%(L)+ 89%(R)	6%(L)+ 7%(R)	8%(L)+ 4%(R)
2.	Left (mouth)	85%	15%	0%
3.	Right (eyes)	88%(L)+ 86%(R)	10%(L)+ 12%(R)	2%(L)+ 2%(R)
4.	Right (mouth)	90%	6%	4%
5.	Top (eyes)	88%(L)+ 80%(R)	9%(L)+ 14%(R)	3%(L)+ 6%(R)
6.	Top (mouth)	75%	14%	11%
7.	Bottom (eyes)	84% (L)+ 80%(R)	12%(L)+ 12%(R)	4%(L)+ 8%(R)
8.	Bottom(mouth)	84%	9%	7%

Once the facial landmarks are correctly identified and extracted for the images of all the databases, the neural network is trained and then tested and it shows variable performance for different databases. There are various parameters which were taken into amount that affects overall accuracy of the system. Performance is the most important parameter we have considered for the comparison among databases. On testing, it is analyzed that Smart database shows remarkable performance of 96% for unalike expressions like happy, sad, neural, surprise, angry and fear. It shows reliable identification of these emotions. For Jaffee database, it gives the performance of 90% that is somewhat less than the Smart database. Further, performance rate of 78% is achieved while working on the Yale database and similarly on applying the same approach on AT&T database, performance rate of 67% is achieved. The main reasons behind the low performances and accuracy of both Yale and AT&T database is due to different lightening effects, rotations of images to certain angles. The another important reason is the presence of different concerns that includes the pictures of dissimilar persons with beard and glasses also due to which neural network can't able to identify the expressions correctly.

If we take these parameters critically into account, it is observed that illumination conditions should not be present in the image otherwise it will affect the system's performance to larger extent. As seen in the Table1. No illumination conditions are present in first two databases hence achieving high performance rate. On the other hand, there are three illumination conditions in Yale database and further these conditions are not under the control while taking pictures of AT&T database, hence showing low performances. Another important parameter is the size of the image. More is the size of the image, more is the recognition rate and lesser is the distortion of image.

Thus, to analyze the performance correctly, there are various parameters on which the performance of a facial expressions recognition system depends such as different illumination conditions, size of each image, noise level, head positions, pixel intensities [21], age factor, and gender. All the information about databases is tabulated as in table 2. One of the prime factors among all is the presence of noise in the image. For Smart database, we used high resolution front camera so that we can take clear and detailed images even when the light is very low. This is possible with the help of an exclusively large image sensor present in it. It also provides the feature of Beautify 4.0 mode due to which we can take HD images with natural color tones.

Table2. Different parameters effecting performance of the facial emotion recognition system

Database	Illumination conditions	Size of image (pixels)	Noise in image	Head positions in image	Age factor
Smart	No illumination condition	256*256	Not present	straight	young
Jaffee	No illumination condition	256*256	Little bit	straight	Middle age
Yale	3 illumination conditions	320*243	Present	Presence of glasses, beards	Middle age-old age
AT&T	Conditions was not controlled during record	192*112	Present	Frontal but tilt of head	Middle age-old age

Its main spatiality is that it uses Stacked sensor i.e. CMOS which is sensitive to light and make use of PDAF (phase Detection Autofocus) and also rich filters. Due to all these features of smart database images, we have achieved high performance. In Jaffee database, low pass filter is used that also remove noise to larger extent and in last two databases no such filters are used. An important reason behind remarkable performance of smart database and Jaffee database is the presence of only upright faces while in other databases use of glasses, scarf's, beard make the system inefficient. Similarly, as the age of the person increases, the skin becomes wrinkled and it becomes difficult to identify the person. Our database consists of young age person as compared to another databases so neural network shows reliable performance.

After analyzing all the results obtained for each database, comparison on all the databases is performed as depicted by Figure 11. X-axis typifying the type of database used and Y-axis represents or typifying the performance rate achieved by different databases.

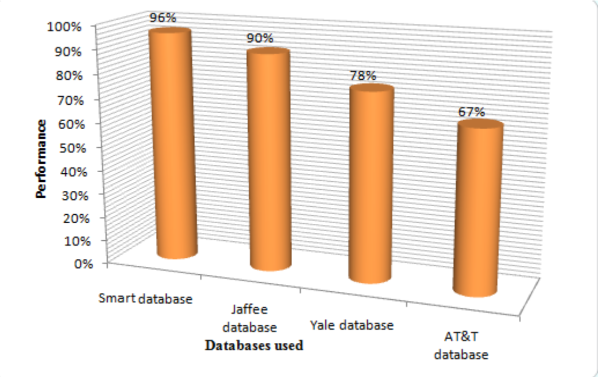


Figure 11. Performance of different databases

Besides the performance it shows the gradient for each database. The gradient descent approach works by picking the gradient of the weight space to perceive the trail of steepest or abrupt descent. The gradient descent is basically the algorithm used to minimize an error function. Smart database have gradient of 0.98, Jaffee database having 0.92 gradients. Similarly, Yale and AT&T databases have 0.87 and 0.80 gradients value.

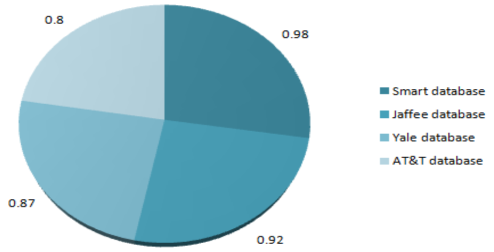


Figure 12. Gradient values for different databases

Therefore, with the increase in performance the capability of neural network to minimize the error function also increases as shown in Figure 12. A regression plot represents the network output in terms of target of training data. For a perfect fit, the training data should lie along the 45 degrees line. On testing, it is observed that for Smart database, training data satisfies the above criteria, which represents good training of the system as shown in Figure 13

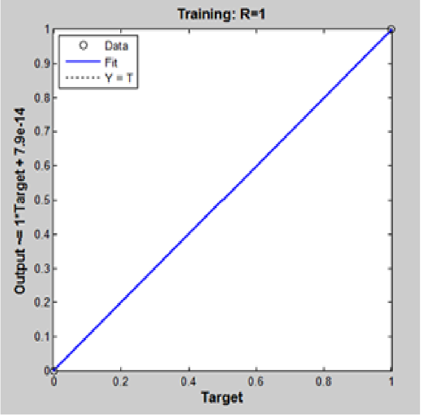


Figure 13. Regression plot

Besides all the image parameters effecting performance of a system also includes no of expressions involved, no of images used for training and testing.

Table 3. Data used for different databases

Database	Training data	Testing data	Performance/ Accuracy	No. of expressions
Smart	20	8	96%	6
Jaffee	13	7	90%	6
Yale	8	5	78%	5
AT&T	23	14	67%	5

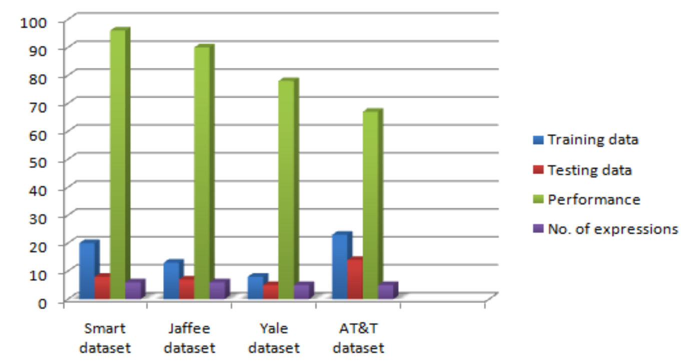


Figure 14. Image factors effecting performance of system

Accordingly these parameters, performance varies. Preferably more images should be taken for training the system. As seen In Fig13. more images are taken in AT&T database but since it involves both genders i.e. different individuals with different subjects, which results in decreased accuracy and if we compare other three datasets, Smart database has more no of training and testing images and it also depends on no of expressions. A greater variety is generally better. By taking control over all the discussed factors, one can improve the accuracy of the system in determining facial expressions of an individual.

Thus, it is analyzed that for Smart database neural network shows the highest performance in identification of different expressions effectively with low error rate and then followed by Jaffee database. Yale database has satisfactory performance and among all the databases, least performance and accuracy is shown by AT&T database and having more error rates.

VII. CONCLUSION

Facial expressions analysis can be done using both static and video images [18, 19]. There are various databases that were used by researchers such as FERAT database, have been employed for involuntary facial expression recognition by multiple researches [22, 23, 30]. In the paper we proposed a reliable landmark detection algorithm with low detection time. The usage of Haar like features reduces the search area significantly in very less time without any complex computation. Also this algorithm has exhibit the great significant performance on different databases. It shows the performances above 90% which is difficult to achieve. In spite of all these, it requires some improvement on certain areas. The designed algorithm has a big potential with the opportunity of further improvements. If the face in distorted significantly like in the mouth curled up to one side in disappointment or when the lips are parted significantly, the feature detection shows error in mouth detection which affects the land marking. Also this algorithm can't able to work much better with the images having great amount of rotations.

VIII. REFERENCES

[1]. P. Ekman et al, "Facial Action Coding System Investigator's Guide," A Human Face, Salt Lake City, UT, Consulting Psychological Press,2002.

[2]. Ekman P et al, "The Facial Action Coding System A Technique for the Measurement of Facial Movement" , San Francisco ;Consulting Psychological Press, 1978.

[3]. F Dornaika, F Davoine, ," Online appearance-based face and facial feature tracking" Washington, DC, USA, in Proc. Of Int. Conf .on Pattern Recognition, 2004, vol. 3. pp. 814–817

[4]. J Cohn, A Zlochower, JJJ Lien, T Kanade.," Feature-point tracking by optical flow discriminates subtle differences in facial expression" , in Proc. Of IEEE Int. Conf. on Automatic Face and Gesture Recognition , Nara, Japan, 1998, pp. 396–401

[5]. M Pantic, LJM Rothkrantz, "Automatic analysis of facial expressions: the state of the art" , IEEE Trans. Pattern Anal. Mach. Intell. 22(12), 1424–1445 (2000)

[6]. K Liu, A Weissenfeld, J Ostermann, X Luo," Robust AAM building for morphing in an image-based facial animation system" , in Proc .of Int. Conf. on Multimedia and Expo., Hannover, Germany, 2008, pp. 933–936

[7]. S Ioannou, G Caridakis, K Karpouzis, S Kollias," Robust feature detection for facial expression recognition" , J. Image Video Process. 2007(2), 5–5 (2007)

[8]. UPark,AKJain, "3D face reconstruction from stereo images ",in Proc. Of Int. Workshop on

Video Processing for Security., Quebec City, Canada, 2006, p. 41

[9]. AA Salah, N Alyüz, L Akarun," Registration of 3D face scans with average face models" . J. Electron. Imag. 17(1), 011006 (2008)

[10]. N Pears, T Heseltine, M Romero, "From 3D point clouds to pose-normalised depth maps" . Int. J. Comput. Vis. 89(2), 152–176 (2010)

[11]. C. Du, Q. Wu, J. Yang, and Z. Wu, "SVM based ASM for facial land- marks location," in Proc. IEEE International Conference on Computer and Information Technology (CIT'08), Nov 2008, pp. 321–326.

[12]. T. F. Cootes and C. J. Taylor, "Active shape models - smart snakes," in Proc. British Machine Vision Conference (BMVC'92), Leeds, UK, set 1992, pp. 266–275

[13]. T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in Proc. of the European Conference on Computer Vision (ECCV'98), Freiburg, DE, Jun 1998, pp. 484–498

[14]. D. Vukadinovic and M. Pantic, "Fully automatic facial feature point detection using gabor feature based boosted classifiers," in Proc. IEEE International Conference on Systems, Man and Cybernetics (SCM'05), Waikoloa, Hawaii, Dec 2005, pp. 1692–1698.

[15]. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proc. CVPR, 2001, vol. 1, pp. 511–518.

[16]. J. Canny," A Computational Approach to Edge Detection" ,IEEE Trans. Pattern Analysis Machine Intelligence, vol. 8, no. 6, June 1986.

[17]. Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 1, pp. 39–58, Jan. 2009

[18]. Essa, I.A.; Pentland, A.P., Coding, analysis, interpretation, and recognition of facial expressions, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume:19,Issue:7,July 1997, Page(s): 757 – 763

[19]. Rosenblum, M.; Yacoob, Y.; Davis, L.S., "Human expression recognition from motion using a radial basis function network architecture" , IEEE Transactions on Neural Networks, Volume: 7 Issue: 5, Sept. 1996, Page(s): 1121 –1138

[20]. G. Donato, M.S. Barlett, J.C. Hager, P. Ekman, T.J. Sejnowski, "Classifying facial actions" , IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 21, No 10,pp 974–989, 1999-2000

[21]. Gizatdinova Y, Surakka V ," Feature-based detection of facial landmarks from neutral and expressive facial image" ,. IEEE Trans Pattern Anal Mach Intell 28:135139, 2006

[22]. HC, Wu CY, Lin TM ," Facial Expression Recognition Using Image Processing Techniques and Neural Networks Advances in Intelligent Systems & Application" ,. Springer-Verlag Berlin Heidelberg, pp 259–267,2013.

[23]. Ma L, Khorasani K ," Facial expression recognition using constructive feedforward neural networks" , IEEE Trans Syst Man Cybern B: Cybern 34:1588–1595,2004.

[24]. Alessandro, L.F., "A neural network facial expression recognition system using unsupervised local processing" In: ISPA 2001. 2nd international symposium on image and signal processing and analysis, Pula, CROATIE, pp. 628–632 (2001)

[25]. Saatci, Y., Town, C. "Cascaded classification of gender and facial expression using active appearance model" ,. In: FGR 2006. Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, pp. 393–400. IEEE Computer Society Press, Washington, DC, USA (2006)

[26]. Yacoob, Y., Davis, L.S." Recognizing human facial expressions from long image sequences

using optical flow" IEEE Trans. Pattern Anal. Mach. Intell. 18(6), 636– 642 .

[27]. M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," IEEE Transactions on Affective Computing, vol. 2, no. 2, pp. 92–105, 2011.

[28]. H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 1, pp. 23–38, 1998.

[29]. P. J. Phillips, H. Moon, P. J. Rauss, and S. A. Rizvi, "The FERET evaluation methodology for face-recognition algorithms," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 10, pp. 1090–1104, 2000.

[30]. M.Lyons, J.Budynek, and S.Akamastu, "Automatic Classification of Single Facial Images" , IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.21, 1999, pp.1357-1362.

[31]. R.W. Picard, E. Vyzas, J. Healey, "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State," in IEEE Transactions Pattern Analysis and Machine Intelligence, vol. 23, no. 10, pp. 1175- 1191, 2001.

[32]. K.H. Park, H.E. Lee, Y. Kim, Z.Z. Bien, "A Steward Robot for HumanFriendly Human-Machine Interaction in a Smart House Environment," in IEEE Transactions on Automation Science and Engineering, vol. 5, no. 1, pp. 21-25, 2008.

[33]. C. Lisetti, S. Brown, K. Alvarez, A. Marpaung, "A Social Informatics Approach to Human-Robot Interaction with a Service Social Robot," in IEEE Systems, Men, and Cybernetics. Special Edition on Human-Robot Interaction, vol. 34 no. 2, 2004

[34]. M. Pantic and L.J. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," in IEEE Trans. Pattern Anal. Mach. Intell. vol. 22, no. 12, pp. 1424-1445, 2000

[35]. Y. S. Gao, M. K. H. Leung, S. C. Hui, and M. W. Tananda, "Facial expression recognition from line-based caricature," IEEE Trans. System, Man, & Cybernetics (Part A), vol. 33, no. 3, pp. 407-412(May, 2003).

[36]. L. Ma and K. Khorasani, "Facial expression recognition using constructive feedforward neural networks," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 34, no. 3, pp. 1588–1595, Jun. 2004.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Fractional-DCT ADALINE method for Speech Enhancement

R. Ram, M. N. Mohanty*
Department of Electronics and Communication Engineering Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India

ABSTRACT

Enhancement of speech is an essential task in the most of the field along with social management. The quality of the speechisdegraded mostly in the noisy environment.Also the same can be obtained from the physical disable people. To improve the quality of the speech, different enhancement algorithms can be applied. In this paper an attempt has been taken by the help of adaptive neural network based model. The fractional DCT (FrDCT) has been utilized for the input to the model. Earlier to it the discrete cosine transform (DCT) coefficients are employed to the model for the sake of verifications. That follows the coefficients of FrDCT and the results are compared. The deteriorated speech considers in this are in vehicular environment as well as summer environment with the fan.The results obtained for different noise environment that the FrDCT-ADALINE method outperforms better than the other methods.

Keywords: Adaptive Linear Neuron, Fractional Discrete Cosine Transform, Filtering, Segmental Signal-to-Noise Ratio, Mean Opinion Score, Speech Enhancement.

I. INTRODUCTION

Speech enhancement has been one of the major challenges in the speech community since a decade due to its practical applications in mobile telephony, speech recognition, hearing aids etc. The performance of these systems degraded owing to the presence of background noise, babble noise, cockpit noise, impulsive noise inducing distorted information exchange. To reduce the impact of these disturbances, many algorithms have been introduced to enhance the perceptual quality of the speech signals. It is generally a difficult task to regenerate the desired signal without affecting the speech signal and the performance is restricted between speech distortion and noise reduction (Loizou, P., 2007, Haykin,S., 2009).

Boll's spectral subtraction is one of the admired algorithm for speech enhnacement. On the other hand the presence of musical noise and the half wave rectification are the foremost drawbacks of it. Further, lots of modifications have been made in this method. Non-linear spectral subtraction, Multiband spectral subtraction, spectral over subtraction, spectral subtraction consisting of perceptual properties are some of them (Boll,S.F., 1979, Upadhyay, N., Karmakar, A., 2015). Different adaptive algorithms are also designed for this problem. Wiener filter, Least mean squares, Recursive Least Squares, State Space Recursive Least Squares are some of them (Ram, R Mohanty, M.N.,2016, Vihari,S., Murthy,A.S., Soni, P., Naik,D.C.,2016).

Neural networks can be applied to cancel and remove noise from noisy speech signal (Fah,L.B., Hussain, A., Samad,S.A., 2000). The covolutional neural network is used as a convolutional denoising autoencoder. This type of architecture reflects strong correlations between speech in time and features. The speech specttra and the ideal ratio mask are estimated and the performance are compared. Different language speech signals are trained in this network for measuring the objective quality of the speech (Kounovsky, T., Malek,J.,2017).

Youshen Xia and Jun Wang (2015) have proposed a neural network based Kalman filter for speech enhancement. The recurrent neural network is based on the noise constrained least squares estimate and it provides better performance against non gaussian noise. The neuaral network is also designed for cochlear implant users. A speaker dependent algorithm and a speaker independent algorithm are compared for speech enhancement. The intelligibilty of the speech is improved with a low computational complexity (Goehring,T., Bolner,F., Monaghan,J.J.M., Dijk, B., Zarowski,A., Bleeck,S., 2017).

The choice of fractional order is an important issue in different applications (Ozaktas, H.M., Ankan,O., Kutay,M.A., Bozdaki, G., 1996). For filtering the noisein Fractional domain, the Wigner distribution should be rotated with an appropriate angle.The filtering process is performed iteratively (Kutay, M.A., Ozaktas, H.M., Arikan, O., Onural, L., 1997). The fractional specral subtraction method is proposed by Wang Zhenli and Zhang Xiongwei (2005) for speech enhancement. To filter out the noise signal, an estimated fractional noise spectrum is subtracted from the fractional noise speech spectrum.The fractional fourier transform based adaptive filter can also be designed based on fractional Fourier transform (FrFT). Using different window and fractional order the adaptive filters are implemented (Ram, R., Mohanty, M.N., 2017). Different noisy signals are tested and FrDCT filter proves better than the FrFT filter. But DCT is much better than the DFT for removing the noise components from the speech signals (Cariolaro, G., Erseghe, T., Kraniauskas, P., 2002).

DCT based filtering is suitable on the signal dependent noise. The important issue is signal dependent and multiplicative noise present in speech signal. The DCT based filtering is found competant. Also in this case, block based processing is preferred (Jeeva, M.P.A., Nagarajan, T., Vijayalakshmi, P., 2016, Ram, R., Mohanty, M.N., 2017).

The paper is organized as follows : Section 1 provides the Introduction of the work. Section 2 deals with design of the model. The model is based on ADALINE. Section 3 explains the utilization of FrDCT in ADALINE model and the model is converted to FrDCT-ADALINE model. Section 4 discusses the result for different deterioted speech signal.Finally Section 5 concludes the piece of the work.

II. ADALINE ADAPTIVE FILTER FOR SPEECH ENHANCEMENT

ADALINE is one of the most commonly used neural networks for noise cancellation. Based on this fact, this neural network approach is widely used in the field of signal processing applications. The weights and bias of this ADALINE are adapted the LMS learning rule based on the Widrow-Hoff (Daqrouq, K., Abu-Isbeih, I.N., Alfauori, M., 2009). The weights are adjusted to minimize the error. The structure of an ADALINE is shown in Figure 1.

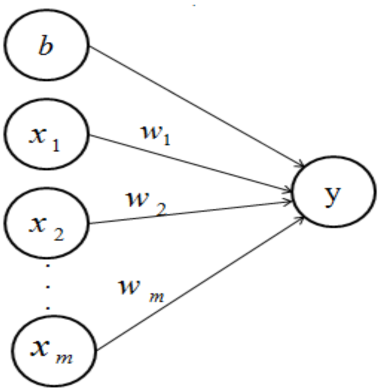


Figure 1. Structure of the ADALINE

The ADALINE has one output which receives the input from many neurons. To compute the output of each time sequence, the individual set of weight and bias are considered. The input layers $x_1, x_2,...x_m$ are interconnected to output y by weights w_1, w_2,...w_m and bias *b*. Figure 2 presents the block diagram of the speech enhancement method using ADALINE. To enhance the noisy speech signal, the clean speech signal is considered as the target signal.

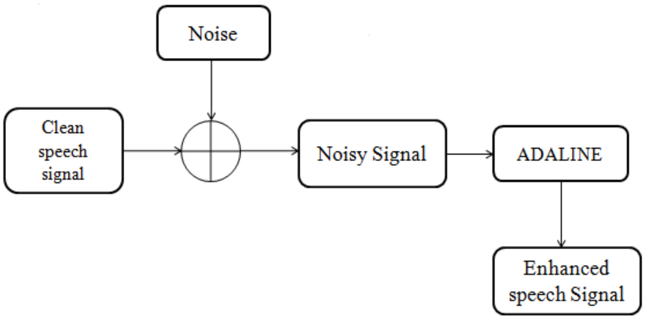


Figure 2. Speech enhancement using ADALINE

The following algorithm steps present the speech enhancement system.

1. The learning rate parameter (*l*) is set at 0.25. (experimentally)
2. The biases *(b(i))* and weights *(w(i))* are set at 0.95 and 0.35 respectively. (experimentally)
3. The clean signal is considered as target signal (t).
4. Set the noisy signal as the input signal *(x).*
5. For each time index, the output *(y)* and the error *(e)* are calculated as

   $$y_i = w_i * l_i + b_i$$

   $$e_i = t_i - y_i$$
6. The weights and the biases are adjusted as

   $$w_i(new) = w_i(old) + 0.01(e_i * x_i)$$

   $$b_i(new) = b_i(old) + 0.01 * e_i$$

An additional factor of 0.01 is considered for fine tuning. All values are set experimentally for adjusting the weights and biases. Steps 5 and 6 are repeated for the first few samples of the speech signal. The enhanced signal is obtained as the error signal yielded by the adaptive network.

III. FRACTIONAL DCT-ADALINE METHOD FOR SPEECH ENHANCEMENT

Due to the real coefficients of DCT, the spectral resolution is better for the equal size of Discrete Fourier transform. In FrDCT, both the amplitude and the phase of the noisy speech signal is enhanced and the upper bound on the maximum improvement of SNR is possible. Figure 3 presents the block diagram of the FrDCT-ADALINE enhancement method. The fractional order is selected arbitrarily to get maximum SNR.

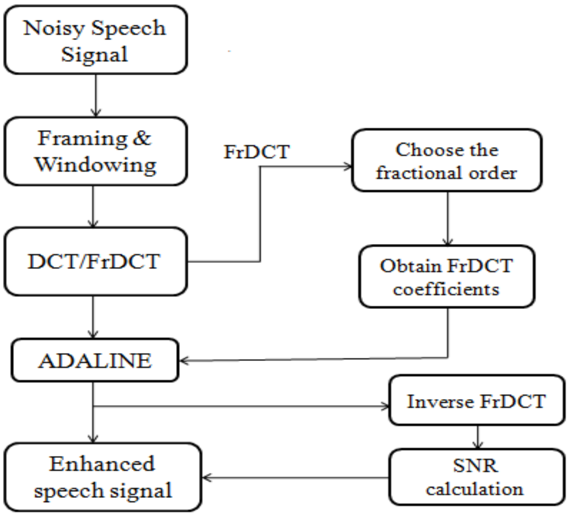


Figure 3. Block diagram of the Proposed Speech enhancement method

The forward DCT and inverse DCT of the discretesequence x_m are defined as

$$X_k = \frac{1}{\sqrt{N}} c_k \sum_{m=0}^{N-1} x_n \cos\left(2\pi \frac{(2m+1)k}{4N}\right)$$

and

$$x_m = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} c_k X_k \cos\left(2\pi \frac{(2m+1)k}{4N}\right)$$

Where k=0, 1,……, N-1, and m=0,1, ….., N-1. N =length of the signal.

$c_0 = 1$ and $c_k = \sqrt{2}$ for *k>0.* The DCT matrix (P) of size $N \times N$ for column sequence x_m and X_k are expressed as

$$P = \left\| \frac{1}{\sqrt{N}} c_k \cos\left(2\pi \frac{(2m+1)k}{4N}\right) \right\|$$

In matrix form the signals are represented as $X = Px$ and $x = P^{-1}X$.

The FrDCT is produced from the real powers of *P.* The FRDCT can be defined as $P_\alpha : X_\alpha = P_\alpha x$ where $P_\alpha = A\Lambda^\alpha A^*$. To obtain the real values of the DCT matrix, the eigen values λ_m of the matrix is replaced by $\lambda_m{}^a$. Where a is the order of the fractional transform. Using the additive and orthogonality property, the inverse FrDCT is obtained as $X_\alpha = P_\alpha s$ and $s = P_{-\alpha} X_\alpha$

Where s is the enhanced signal obtained from the FrDCT which is real and orthogonal. The same steps are followed for ADALINE by considering the fractional DCT coefficients. For reconstruction of original signal, the output of the neural network will be back to the Inverse FrDCT. In nonstationary noisy cases, fractional DCT transform is better than the standard DCT. It can circumstantially care for the low frequency cotents also. Therefore the information will not be contaminated and the error estimation is less. This method is verified by using discrete FrDCT with ADALINE and the enhancement results are exhibited and compared in the result section.

IV. RESULTS AND DISCUSSION

The objective of this proposed method is to obtain the clean signal from the noisy signal. To verify, 'have a nice day' is recorded by a female speaker in a class room at a sampling rate 8 KHz as shown in Figure 4. For enhancement, different noise is added to the speech signal such as bus noise, street noise, train noise, fan noise and babble noise. The corrupted bus noise speech signal is shown in Figure 5. The noisy signal is then applied as the input signal and the clean signal is the target signal to the ADALINE. According to the LMS rule, the linear network adapts to cancel the noise from the noisy signal. The learning rate is set at 0.25 which is determined experimentally. The bias and the weights of the network are set at 0.95 and 0.35 respectively. Figure 6 presents the result of enhancement method using ADALINE.

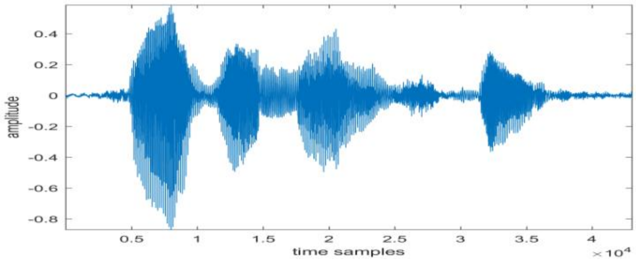


Figure 4. Clean speech signal 'Have a nice day'

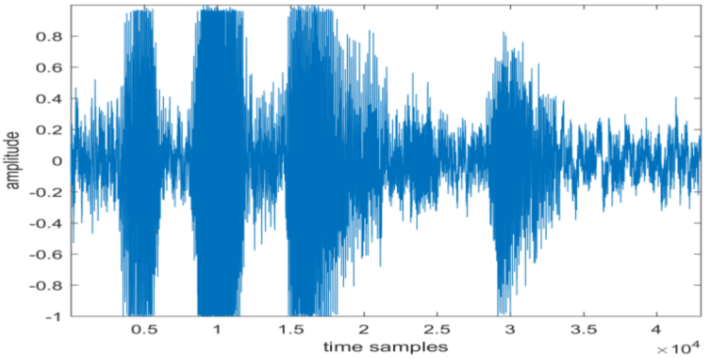


Figure 5. Noisy signal (bus noise)

To obtain better enhancement, the DCT and the FrDCT coefficients are acquired and employed to the ADALINE. Due to nonstationary nature of speech, the signal is first splitted into frames. A hamming window of length 512 is multiplied to each frame to avoid spectral leakage. The fractional order is selected arbitrarily and adaptively to achieve maximum SNR. The inverse transform is aimed to recover the enhanced signal. The enhanced signal of DCT-ADALINE method is shown in Figure 7 and Figure 8 presents the enhanced output of FrDCT_ADALINE method. The order 1.8 of FrDCT provides better enhanced signal.

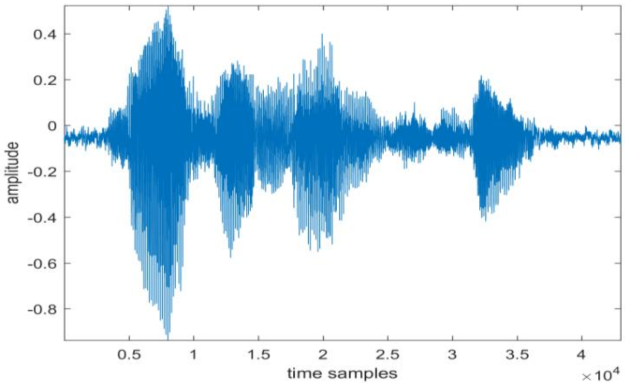


Figure 6. Filtered signal of ADALINE enhancement method

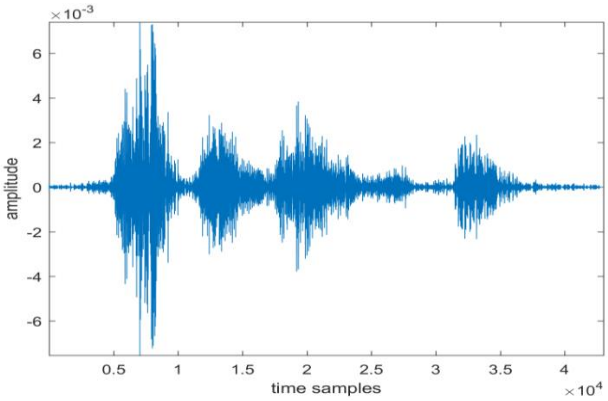


Figure 7. Filtered signal of DCT-ADALINE enhnacement method

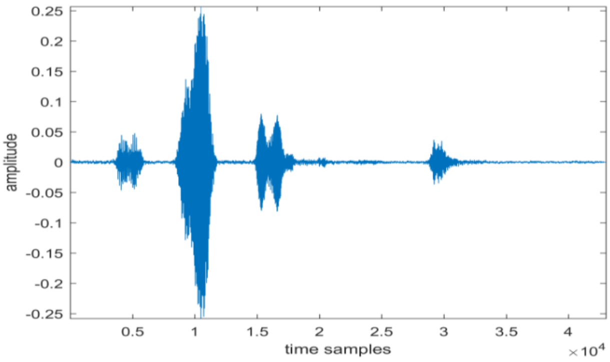


Figure 8. Filtered signal of FrDCT-ADALINE enhancement method

To measure the quality assessment of different signal, the Power Spectral Density (PSD) and spectrogram are evaluated. The PSD of clean signal, noisy signal and the ADALINE enhancement method is shown in Figure 9, Figure 10 and Figure 11 respectively. It is noticed that the PSD of ADALINE filtered signal is somehow similar to the target signal. For further improvement another methods have tested. Figure 14 and Figure 15 present the spectrogram of the clean signal and noisy signal. The spectrogram also indicates that the noise is still remains in the ADALINE enhancement method as in Figure 16.

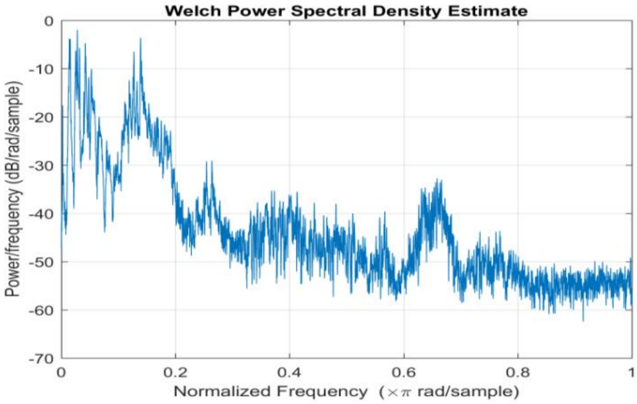


Figure 9. PSD of clean signal

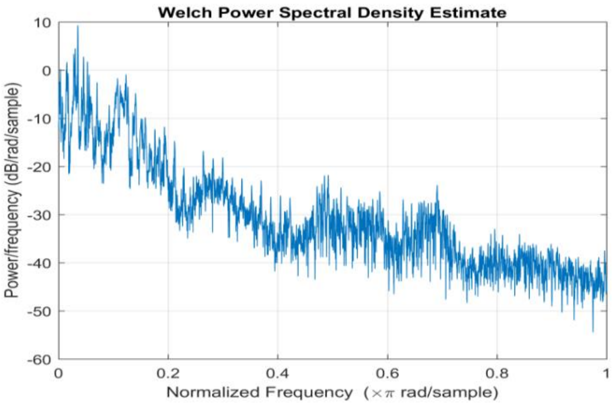


Figure 10. PSD of noisy signal

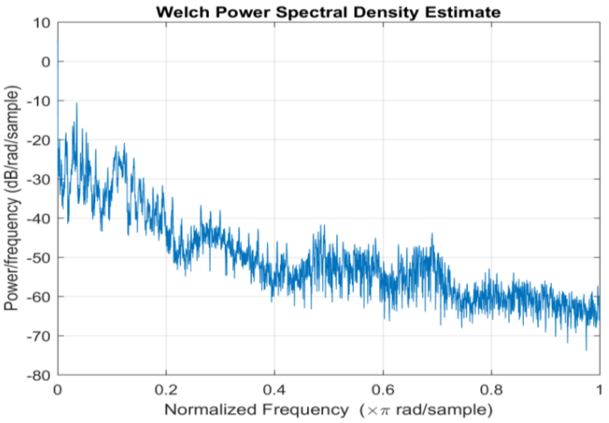


Figure 11. PSD of ADALINE enhancement method

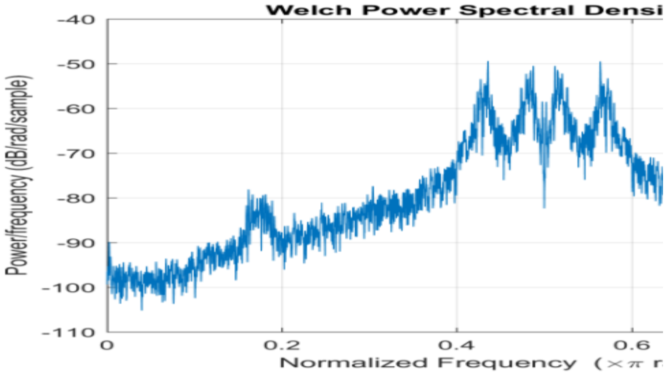


Figure 12. PSD of DCT-ADALINE enhancement method

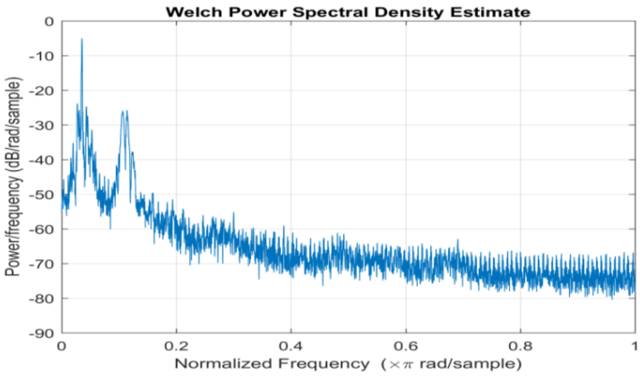


Figure 13. PSD of FrDCT-ADALINE enhancement method

The PSD of the FrDCT-ADALINE method (Figure 13) has better result than the DCT method (Figure 12). The spectrogram of the FrDCT (Figure 18) indicates the better result than the other methods. Figure 17 represents the spectrogram of the DCT-ADALINE method consisting of noise.

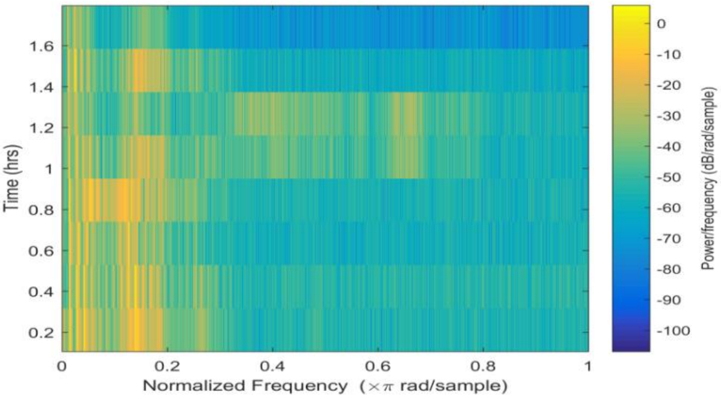


Figure 14. Spectrogram of clean speech signal

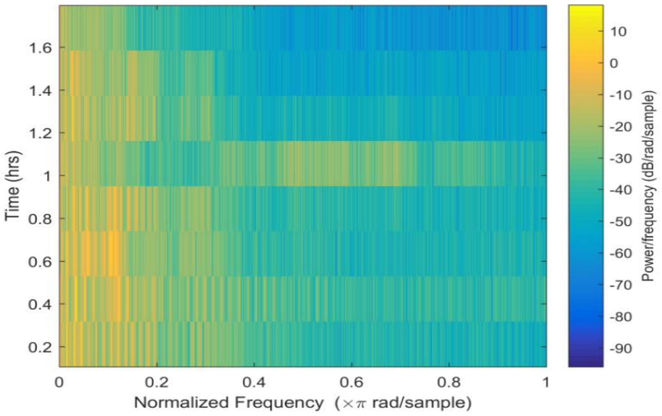


Figure 15. Spectrogram of Noisy speech signal

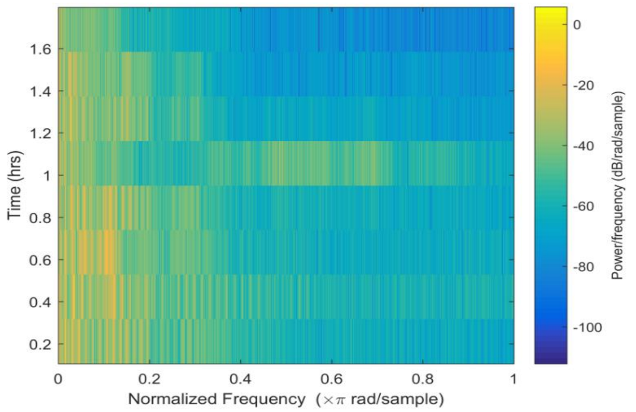


Figure 16. Spectrogram of ADALINE enhancement method

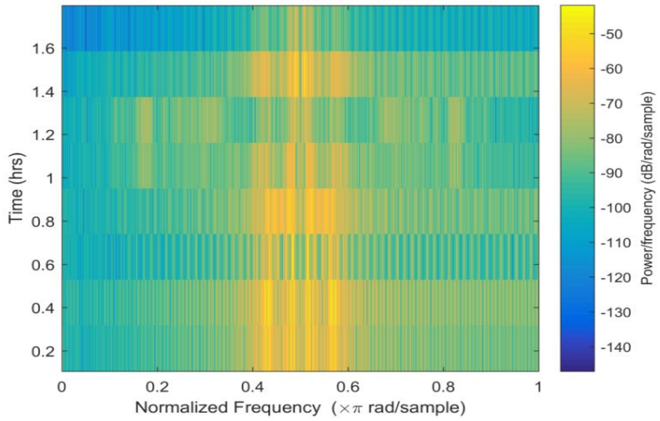


Figure 17. Spectrogram of DCT-ADALINE enhancement method

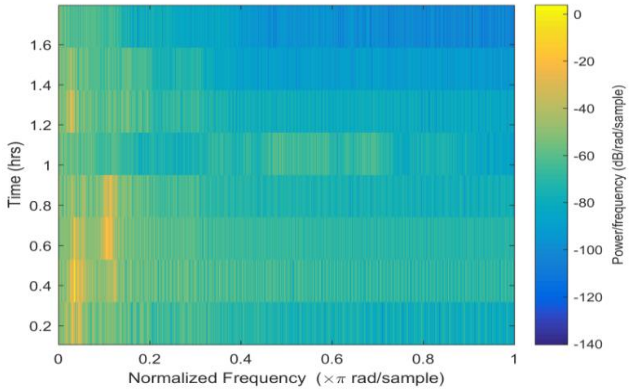


Figure 18. Spectrogram of FrDCT-ADALINE enhancement method

Different objective and subjective measures are there to test the speech quality and intelligibilty (Hu, Y., Loizou, P., 2008). In this experiment, the signal-to-noise ratio (SNR) and the Mean-Opinion-Score (MOS) are considered for evaluation of the enhancement methods. Table 1 shows the SNR of different types of noisy signals and Table 2 shows the MOS of noisy signals. The maximum SNR improvement is 6.1042 dB achieved in FrDCT-ADALINE for train noise. But the highest MOS is 4.55 obtained for babble noise.

Table 1. SNR Improvement for different types of noise signal

	SNR before Enhancement (dB)	SNR after Enhancement (dB)	SNR Improvement (dB)
ADALINE			
Bus Noise	4.5346	6.3652	1.8306
Street Noise	3.4565	5.3492	1.8927
Train Noise	6.8760	8.5032	1.6272
Fan Noise	3.7659	4.8116	1.0457
Babble Noise	2.8762	4.6234	1.7472
DCT_ADALINE	4.5346	7.6547	3.1201
Bus Noise	3.4565	7.8734	4.4169
Street Noise	6.8760	9.2560	2.3800
Train Noise	3.7659	7.2454	3.4795
Fan Noise	2.8762	6.4582	3.5820
Babble Noise			
FrDCT_ADALINE	4.5346	10.0874	5.5528
Bus Noise	3.4565	8.9565	5.5000
Street Noise	6.8760	12.9802	6.1042
Train Noise	3.7659	9.5434	5.7775
Fan Noise	2.8762	8.5653	5.6891
Babble Noise			

Table 2. Mean Opinion Score for different types of noise signal

	ADALINE	DCT_ADALINE	FrDCT_ADALINE
bus noise	2.63	3.34	4.35
street noise	2.83	3.67	4.33
train noise	2.90	3.52	4.08
fan noise	2.87	3.89	4.53
babble noise	2.44	3.55	4.55

V. CONCLUSION

The DCT-ADALINE and the FrDCT-ADALINE speech enhancement algorithms are proposed in this work and the results are compared with the ADALINE. Results show that the proposed method is better than the other methods.Due to higher energy compaction and better spectral resolution property, DCT and the FrDCT provide better enhanced signal. The speech signal corrupted by different noise are tested to show the results. The SNR improvement is more perceptible in case of train noise i.e. 6.1042 dB. And the maximum MOS is 4.55 is obtained from the babble noise affected speech signal. Furthermore the listening test is done for all the enhanced signals. FrDCT-ADALINE enhancement method proved to be better for all the tests compared to other algorithms. Better enhanced signals can be achieved in other transforms as well, and thiswillbescoped in future work.

VI. REFERENCES

[1]. Loizou, P., 2007, Speech Enhancement: Theory and Practice. CRC Press.

[2]. Haykin, S. S., & Haykin, S. S. , 2009, Neural networks and learning machines. New York: Prentice Hall/Pearson.

[3]. Boll,S.F., 1979, Suppression of Acoustic Noise in Speech using Spectral Subtraction.IEEE Transaction ASSP. 113-120.

[4]. Upadhyay, N., Karmakar, A., 2015, Speech Enhancement using Spectral Subtraction-type Algorithms: A Comparison and Simulation Study. Elsevier Procedia Computer Science. 55, 574-584.

[5]. Ram, R Mohanty, M.N.,2016, Performance Analysis of Adaptive Algorithms for Speech Enhancement Applications. Indian Journal of Science and Technology. 9(44).

[6]. Vihari,S., Murthy,A.S., Soni, P., Naik,D.C.,2016, Comparison of Speech Enhancement Algorithms. Procedia Computer Science. 89, 666 – 676.

[7]. Fah,L.B., Hussain, A., Samad,S.A., 2000, Speech Enhancement by Noise Cancellation Using Neural Network. IEEE Conf..

[8]. Kounovsky, T., Malek,J.,2017, Single Channel Speech Enhancement Using Convolutional Neural Network. IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics. 1-5.

[9]. Daqrouq, K., Abu-Isbeih, I.N., Alfauori, M., 2009, Speech Signal Enhancement Using Neural Network and Wavelet Transform. International Multi-Conference on Systems, Signals and Devices.

[10]. Xia,Y., Wang,J., 2015, Low-Dimensional Recurrent Neural Network-based Kalman Filter for Speech Enhancement.Neural Networks 67, 131–139.

[11]. Goehring,T., Bolner,F., Monaghan,J.J.M., Dijk, B., Zarowski,A., Bleeck,S., 2017, Speech enhancement based on neural networks improves speech intelligibility in noise for

cochlear implant users.Hearing Research 344, 183-194.

[12]. Ozaktas, H.M., Ankan,O., Kutay,M.A., Bozdaki, G., 1996, Digital Computation of the Fractional Fourier Transform. IEEE Transactions on Signal Processing, 44(9).

[13]. Kutay, M.A., Ozaktas, H.M., Arikan, O., Onural, L., 1997, Optimal Filtering in Fractional Fourier Domains. IEEE Transactions on Signal Processing, 45(5).

[14]. Zhenli, W., Xiongwei, Z., 2005, On the application of fractional Fourier transform for enhancing noisy speech. IEEE International Symposium on Microwave, Antenna, Propagation and EMC Technologies for Wireless Communications Proceedings.

[15]. Cariolaro, G., Erseghe, T., Kraniauskas, P., 2002, The Fractional Discrete Cosine Transform. IEEE Transactions on Signal Processing. 50(4).

[16]. Ram, R., Mohanty, M.N., 2017, Design of Fractional Fourier Transform based Filter for Speech Enhancement. IJCTA. 10(07), 235-243.

[17]. Jeeva, M.P.A., Nagarajan, T., Vijayalakshmi, P., 2016, Discrete Cosine Transform-Derived Spectrum based Speech Enhancement Algorithm using Temporal-Domain Multiband Filtering.IET Journal of Signal Processing.

[18]. Ram, R., Mohanty, M.N., 2017, Design of Filter using Fractional-DCTfor Speech Enhancement. Int. Conf on Sustain.able Computing Techniques in Engineering, Science and Management.

[19]. Hu, Y., Loizou, P., 2008, Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang. Process., 16, 229–238.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Survey on Liver Segmentation Schemes in CT Images

Sakshi Thapa
Sam Higginbottom University of Agriculture, Technology and Sciences,Allahabad, SMVDU, Allahabad, Uttar Pradesh, India

ABSTRACT

In the field of medical image processing, the segmentation of liver in computed tomography images are of enormous significance. Dividing schemes into two categories that are semi-automatic and fully automatic schemes. Both classes have some techniques, approximation, related queries; some drawbacks will be described and clarified. To obtain a liver segmentation, there is an analysis on methods for segmentation of liver as well as techniques using computed tomography images are shown. Following the relative study for liver segmentation schemes various measurements, scoring for liver segmentation are given; advantages and disadvantages of techniques will be emphasized carefully. Several faults and difficulties of the suggested methods are still to be focused.

I. INTRODUCTION

Now a day, in an area of medical image processing, the segmentation of liver in computed tomography images have enormous importance. It is the start and an important action for detection of liver diseases, liver volume measurements and 3D liver volume rendering. To bring out the liver data or information, a manual process and visual inspection are used, which is very time consuming process and process of ideas to fix a problems. Dividing the methods of liver segmentation into two categories that is methods of semi-automatic and methods of fully-automatic segmentation of liver. The image processing and machine learning theories gives the more knowledge about these two methods of liver segmentation. Furthermore, it is not an easy task because of low level contrast and indistinct boundaries which are used to identify the computed tomography images. The above features are generated by the partial volume effects because of spatial averaging, patient movement, and beam hardening. In addition, same types of gray levels may be used by neighbor organs in the body like spleen, liver, and stomach. For now, same type of gray levels cannot use the same organ related to the same topic. All these characteristics, difficulty with huge diversity of liver shapes enhances the problem in the liver segmentation task.

II. LIVER VOLUME SEGMENTS

Basically, segmentation of liver by CT images is divided into two different classes that are partially/semi-automatic method and automatic liver segmentation method.

2.1 Semi-automatic scheme for segmentation of liver

In these schemes, it needs a little user involvement which is used to complete the task. The involvement for this task is changes from chosen of seed pixels manually to a manual refinement of a binary mask for the liver. The latest Semi-automatic methods for segmentation of liver are obtainable, and according to image processing techniques, these methods are predetermined.

2.1.1 Graph based semi-automatic schemes

Images are handling by weighted and undirected graphs, where pixels called the vertices, and neighboring pixels are view as connected vertices. The weights of the edges in the graph calculate the likeness among two connected vertices. The involvement of user used in the methods of segmentation of liver via the operation for the selection of seed points and via steps for modification. Normally, the live wire algorithms as well as the graph-cut segmentation algorithms are used under this class.

Barrett and Mortensen (1997) to remove edges in medical images, they planned an algorithm for live wire segmentation. To find the least cost paths among seed points which are already by the user is calculated by the algorithm of live wire segmentation. The weighted sum of Image features such as the gradient value, gray value, gradient direction, and Laplacian zero-crossing are used to compute the cost of the path. Firstly, initial seed point will be selected by the user which lies on the boundary of the organ, after that the opening from the elected seed point (already classified by the user in the image), then Dijkstra's search algorithm or dynamic programming algorithms are used to calculate all possible least-cost paths. User will select the boundary of the image.

Schenk et al. (2001) expand the above liver-wire method for segmentation of liver in computed tomography images, also helps to decrease the users communication and calculation period. The cost function is calculated via determining the liver shape from the nearest adjacent slice in the body which is already segmented. User has capability to manage the process of segmentation which is supported by algorithm of liver-wire segmentation. Job of the user will be restricted by choosing the seed points and selecting the most wanted edges where as the processor will manage the details

Beichel et al. (2007) used in their research the graph cut segmentation algorithm. They anticipated in computed tomography images depending on the method of graph cut segmentation, 3D interactive liver segmentation approach.

2.1.2 Region-growing based semi-automatic schemes

This technique is based on reality in which the common gray values are shared by close pixels. Generally, this method is used in an iterative or replication manner in which the whole organ is segmented inside the liver at distinct areas. Manually recent pixels are added to the seed area as intensity of a surrounding area is below that of seed intensity under given limited value. Beck and Aurich (2007) engaged in their approach region- growing algorithm of interaction liver segmentation. They anticipated three dimension region-growers through nonlinear coupling criterion. User manually corrects the leaked regions or missing parts. By calculating the convex hull within restricted local regions around the boundary, the segmentation proceeds. This process is called post processing step.

2.1.3 Level sets based semi-automatic schemes

In this technique, user illustrates a rough contour from inside or outside the object, and then the contour will contract/enlarge. This algorithm comes under image segmentation problem. The process of contracting/enlarging will be terminated, when contour meets the object boundary. The major function managing the way of contour contracting/enlarging also determining the terminate point of this process is done by speed function. Liver segmentation methods under semi-automatic are classified into two groups, that are 2D level sets methods and 3D level sets methods.

2.1.4 Atlas matching semi-automatic schemes

Probabilistic atlases are established from huge number of anatomical images by a manual segmentation. By using affine transformations, pictures have been submitted into a standard space. These images as well as corresponding segmentations are then averaged and engaged into a Bayesian frame for constructing a probabilistic Atlas,. For every pixel, the randomness for a particular organ is calculated. At last, to bring out the needed organ which depends on the later probability a simple thresholding or conditional mode algorithm is used. The probabilistic atlas requires a lot of training data can be gathered and physically fragmented which is its main disadvantage.

2.2 Fully-automatic liver segmentation schemes

By "fully automated" we mean that without any user involvement, segmentation process of liver will be applied. Normally, fully-automatic liver segmentation methods are highly valued by radiologists and also release by udders faults and partiality, as well as there is difficult and wastage of time plus save the operator from this drawbacks.

2.2.1 Deformable model based automatic schemes

Gao et al. (1998) proposed a liver and right kidney parameterized 3D models as well as explain the technique which adjusts them to abdominal computed tomography images. To calculate the matching between the image gradient direction and the deformable model unit surface, they will identify the term energy function. When the energy function reaches the least value, an optimal match is attained. Outcome of the segmentation of liver can be calculated via a radiologist, whereas to calculate the segmentation of right kidney, objective measurements will be used. Some Researchers in Montagnat and Delingette (1996) and Soler et al. (2001) intended to overcome these drawbacks; they merged this method and the method of an elastic registration by using a hybrid method.

2.2.2 Statistical shape model based automatic schemes

This model is constructed which gives a more instances of a contour. Every contour is characterized by set of n labeled landmark points. The instances of the labeled training are line up into a shared co-ordinate frame by the using reproduce analysis. For reducing the amount of squared distances to the mean of the set, it rotates, translates, and scales every training shape.

Lamecker et al. (2004) to attain a grand robustness to noise and outliers, he planned schemes for the segmentation of liver which depends on a SSM. This model is developed by using a semi automatic mapping procedure. Client or user involvement wants to spot the matching points on all liver training data. To confine the liver shape variations, the principle component analysis (PCA) is used. The statistical shape method is permitted after that to collapse inside the captured space of variation via best-matching profile technique expressed by Cootes et al. (1994).

2.2.3 Probability atlas based automatic schemes

Rikxoort et al. (2007) proposed a liver segmentation method which is based on pixels classification in grouping with a multi atlas registration. To present every pixel inside an automatically perceived area as liver tissue or background, K nearest neighbor (kNN) classifier is used. Firstly, every image is resampled to isotropic pixels in case of preprocessing. For detecting as well as correcting the rotations regarding the Z-axis, bones are noticed by thresholding and by applying different rotations X-direction is maximized. Then by thresholding, lungs are detected as well as the area of potential liver is restricted to a fixed height about the lower lung rim. (Rueckert et al. 1999) By using an affine transform followed by B-spines methods in multiple resolutions, the twelve selected training scans are listed to the current image. For this principle, a stochastic gradient optimizer optimizes a negative mutual information cost function which is presented by Mattes et al. (2003). To plan individual training segmentations to the current image, the resulting transformation fields are used. (Rohlfing et al. 2005) the three spatial features are based: they represent the percentage of the probabilistic segmentation above, left, and behind the pixels, this results probabilistic atlas segmentation. By using smoothing and morphological operations the outcomes are post-processed, when classifying each pixel in the area of the mask with a 15-nearest-neighbor classifier.

2.2.4 Rule based automatic methods

Chi et al. (2007) use dedicated scripting language: describe protocols which are utilized to remove dissimilar organization from that images which are already tested. Order of extraction is: background air, lungs and other intra body air, subcutaneous fat and muscle layer, bones within muscle layer, aorta, spine, heart, and liver. Image analysis is done when organization uses the previous detected structure during each removal step. These protocols can also include information regarding neighborhood relations, intensity distributions, geometric features, etc. An area of seed is selected by threshold the right side of the CT slice below the heart after containing the removed listed organization up to the heart, until an item conforming definite size criteria is perceived. A process related to region growing is initiated by using this seed region. With no use of supplied training information and factors which have not been systematically evaluated, all rules are defined.

2.2.5 Gray-level based automatic schemes

Gray-level based automatic methods depend on a statistical analysis for computed tomography segment that are physically segmented to calculate the liver gray levels. Several schemes make use of histogram analysis which depend on a previous data regarding the liver intensity range for the calculating the liver gray levels. The calculated values are used with a straightforward or cyclic process of thresholding to construct a binary map which characterized the liver and then this image processed morphologically to remove connected organs. The existing segmented image gives the information employed as a support for segment the present image or picture. Lastly, make use of active contours or B spines it assists to smoothing the edges of every computed tomography images

Seo and Park (2005) they proposed a scheme for segmentation of liver in contrast enhanced computed tomography images which depends upon algorithm of left partial histogram threshold (LPHT). The left partial histogram threshold removes other neighboring organs apart from the pixels variations. A multi-modal threshold follows histogram transformations that are used to find the ranges of gray level. Lastly, morphological filtering is managed to smoothing the edges of the image and removes the unwanted things.

The extraction of liver in computed tomography images and also used in computer aided liver analysis system; this scheme is planned by Pil et al. (2006). Measured the liver distribution, also employed to choose the region of interesting. Once the probability crossed 50%, the window will allocated as region of interesting when compared to the liver's value of existing probability. Then, to mine the liver regions, the watershed segmentation algorithm is used. The areas which are fragmented can be combined into the momentous areas which are used for optimal segmentation. Lastly, area of liver is chosen by previous data regarding the anatomic information of the liver.

III. COMPARATIVE STUDIES

Automatically fragmentation of liver, troubles are still there. The techniques for liver shape model repeatedly fail. When there is a complex shaped liver, the techniques for liver shape model always failed. Other methods also go through the common fault, the organs which are attached to the liver are failed to split. In computed tomography slice, the connected organs have an alike intensity exterior. A relation between the different techniques is not important because of the need of a common data and a distinctive calculation. Also distinct techniques and methods are experienced on little data sets also techniques performance is calculated which depends on self selected fault functions.

Heimann et al. (2009) calculated sixteen automatic scheme and interactive scheme for segmentation of liver. A much larger standard deviation of the ending scores can be examined by automatic methods when linking to the automatic and the interactive segmentation approaches. The large standard deviation occurs because faults form on outliers. Although, in the comparison of automatic method and interactive have same several successful outcomes in the comparison study of interactive methods, normally the consistency of automatic methods is yet poorer. Troubles occur at distinct test images and areas. Even though several regions cause more fault than additional regions, all methods are fail not even in a single region. When evaluating performance, this observation together with the great variation of results over different test images supports the call for a large and diverse collection of test. With the exactness of created outcome, methods are calculated. Segmentation with great exactness will be observed.

IV. CONCLUSION

The two schemes for the segmentation of liver that are semi-automatic and automatic liver segmentation using computed tomography images related techniques and suggestions have been examined. Even though, many methods for segmentation are tested, troubles always lie there. In reviewed methods, the level of gray based methods are used to achieve the optimistic results, but for database variation they are not that much strong. The high variability of CT intensity values does not examined by gray level estimation. When multifaceted and large data sets are used, the performance could decreases significantly. In addition, several methods needs physical involvements and also need some serious parameters to be experimentally estimated, robustness method is affected by all these facts. Methods of learning are based on the training set, and should select watchfully. There is requirement for plenty of information can be accurately gathered, also can be physically fragmented to create structure, the model based techniques and probabilistic atlases go through the some difficulty; this is because the training set as well as faults of users and unfairness are strongly affects the obtained model. Starting assignment changes the outcome of the segmentation. The algorithms will be unsuccessful while dealing with non standard liver shapes. It is hard to describe an accurate speed function and its factors are the main limitation. In addition, a relation between the different techniques is not important because of the need of a common data and a distinctive calculation. Furthermore, used datasets in mostly investigators are extremely little.

V. REFERENCES

[1]. Arya S, Mount DM, Netanyahu NS, Silverman R, Wu A (1998) an optimal algorithm for approximate nearest neighbor searching. J ACM 45(6):891–923

[2]. Barrett W, Mortensen EN (1997) Interactive live-wire boundary extraction. Med Imaging Anal 1(4):331–341

[3]. Beck A, Aurich V (2007) HepaTux-a semiautomatic liver segmentation system. In: Proceedings of MICCAI Workshop on 3D segmentation in the clinic: a grand challenge pp225-234

[4]. Beichel R, Bauer C, Bornik A, Sorantin E, Bischof H (2007) Liver segmentation in CT data: a segmentation refinement approach. In: Proceedings of MICCAI workshop on 3D

segmentation in the clinic: a grand Challenge, pp 235–245

[5]. Campadelli P, Casiraghi E, Esposito A (2009) Liver segmentation from computed tomography scans: a survey and a new algorithm. Artif Intell Med 45(2–3):185–196

[6]. Carr JC, Beatson RK, Cherrie JB, Mitchell TJ, Fright WR, McCallum BC, Evans TR (2001) Reconstruction and representation of 3-D objects with radial basis functions. In: Proceedings of SIGGRAPH, pp 67–76

[7]. Chi Y, Cashman PMM, Bello F, Kitney RI,(2007) A discussion on the evaluation of a new automatic liver volume segmentation method for specified CT image datasets. In: Proceedings of MICCAI workshop on 3D segmentation in the clinic: a grand challenge, pp 167–175

[8]. Cootes TF, Hill A, Taylor CJ, Haslam J (1994) Use of active shape models for locating structures in medical images. Imag Vis Comput 12(6):355–366

[9]. Heimann T et al (2009) Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans Med Imaging 28(8):1251–1265

[10]. Kainmüller D, Lange T, Lamecker H (2007) Shape constrained automatic segmentation of the liver based on a heuristic intensity model. In: Proceedings of MICCAI workshop on 3D segmentation in the clinic: a grand challenge, pp 109–116

[11]. Koss JE,Newman FD, Johnson TK,Kirch DL (1999) Abdominal organ segmentation using texture transforms and a hopfield neural network. IEEE Trans Med Imaging 18(7):640–648

[12]. Lamecker H, Lange R, SeebaM (2004) Segmentation of the liver using a 3d statistical shape model. Technical report Zuse Institue, Berlin, pp 1–25

[13]. Lee CC, Chung PC, Tsa H (2003) Identifying multiple abdominal organs from CT image series using a multimodule contextual neural network and spatial fuzzy rules. IEEE Trans Inf Technol Biomed 7(3):208–217

[14]. Lim SJ, Jeong, YY, Ho YS (2004) Automatic segmentation of the liver in ct images using the watershed algorithm based on morphological filtering. In: Proceedings of SPIE, pp 1658–1666

[15]. Lim SJ, Jeong, YY, Ho YS (2005) Segmentation of the liver using the deformable contour method on CT images. In: Proceedings of SPIE medical imaging, pp 570–581

[16]. Lim SJ, Jeong YY, Ho YS (2006) Automatic liver segmentation for volume measurement in CT Images. JVCIR 17(4):860–875

[17]. Mattes D, Haynor DR, Vesselle H, Lewellen TK, Eubank W (2003) PET-CT image registration in the chest using free-form deformations. IEEE Trans Med Imaging 22(1):120–128

[18]. Montagnat J, Delingette H (1996) Volumetric medical images segmentation using shape constrained deformable models. In: Proceedings of CVRMed-MRCAS, pp 13–22

[19]. Pil UK, Yun JL, Youngjin J, Jin HC, Myoung NK, (2006) Liver extraction in the abdominal CT image by watershed segmentation algorithm. World congress of medical physics and biomedical engineering, pp 2563–2566

[20]. Rikxoort E, Arzhaeva Y, Ginneken B (2007) Automatic segmentation of the liver in computed tomography scans with voxel classification and atlas matching. In: Proceedings of MICCAI workshop on 3D segmentation in the clinic: a grand challenge, pp 101–108

[21]. Rohlfing T, Brandt R, Menzel R, Russakoff DB, Maurer CR (2005) Quo vadis, atlas-based segmentation?

[22]. Handbook of medical image analysis—Volume III: Registration models. Kluwer Academic, Norwell MA, pp 435–486

[23]. Rousson M, Cremers D (2005) Efficient kernel density estimation of shape and intensity priors for level set segmentation. In: Proceedings of MICCAI, pp 757–764

[24]. RueckertD, Sonoda LI,Hayes C, Hill DL,LeachMO,HawkesDJ (1999) Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans Med Imaging 18(8):712–721

[25]. Schenk A, Prause GP, Peitgen H (2001) Local cost computation for efficient segmentation of 3d objects with live wire. In: Proceedings of SPIE on medical imaging, pp 1357–1364

[26]. Seo KS, Park JA (2005) Improved automatic liver segmentation of a contrast enhanced CT image. Advances in multimedia information process—PCM, pp 899–909

[27]. Slagmolen P, Elen A, Seghers D, Loeckx D, Maes F, Haustermans, K (2007) Atlas based liver segmentation using no rigid registration with a B-spline transformation model. In: Proceedings of MICCAI workshop on 3D segmentation in the clinic: a grand challenge, pp 197–206

[28]. Soler L, Delingette H, Malandain G, Montagnat J, Ayache N, Koehl C, Dourthe O, Malassagne B, Smith M,

[29]. Mutter D, Marescaux J (2001) Fully automatic anatomical, pathological, and functional segmentation from ct scans for hepatic surgery. Comput Aided Surg 6(3):131–142

[30]. Sonka M, Hlavac V, Boyle R (2007) Mathematical morphology in image processing, analysis, and machine vision. Thomson, Newyork

[31]. Susomboon R, Raicu DS, Furst J (2007) Ahybrid approach for liver segmentation. In: Proceedings of MICCAI Workshop on 3D segmentation in the clinic: a grand challenge, pp 151–160

[32]. Tsai D, Tanahashi N (1994) Neural-network-based boundary detection of liver structure in ct images for 3-d visualization. In: Proceedings of IEEE international conference neural networks, pp 3484–3489

[33]. Tsai A, Yezzi A, Wells W, Tempany C, Tucker D, Fan A, Grimson W, Willsky A (2003) A shape- based approach to the segmentation of medical imagery using level sets. IEEE TransMed Imaging 22(2):137–154

[34]. Weickert J, Romeny BMTH, Viergever MA (1998) Efficient and reliable schemes for nonlinear diffusion filtering. IEEE Trans Imaging Process 7(3):398–410

[35]. Wimmer A, Soza G, Hornegger J (2007) Two-stage semi-automatic organ segmentation framework using radial basis functions and level sets: In: Proceedings of MICCAI workshop on 3D segmentation in the clinic: a grand challenge, pp 179–188

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Flower Pollination Algorithm for the Orienteering Problem

Madhushi Verma*1, Tanvi Bothra2, Surabhi Agarwal2, K. K. Shukla2
1Department of Computer Science Engineering, Bennett University, Greater Noida, UP, India
2Department of Computer Science and Engineering, IIT(BHU), Varanasi, UP, India

ABSTRACT

The orienteering problem is an NP-Hard combinatorial optimization problem where the aim is to determine a Hamiltonian path that connects the stated source and target and includes a subset of the vertex set V such that the total collected score is maximized within the given time bound (T_max). Orienteering problem finds application in logistics, transportation, tourism industry etc. We have proposed an algorithm FPA_OP that can be implemented on complete graphs and its performance has been evaluated using standard benchmarks. Also, the results thus obtained have been compared against the latest heuristic for OP i.e. GRASP and it has been shown that for larger T_max, FPA_OP outperforms GRASP. Therefore, the decision maker can implement FPA_OP if he is willing to achieve a larger total collected score at the cost of time delay.

Keywords : Flower pollination algorithm, Metaheuristic, Orienteering problem, NP-Hard problems.

I. INTRODUCTION

Most of the optimization problems are NP-Hard and to solve these problems, we need efficient algorithms that can generate optimal solutions in polynomial time. However, it is difficult to obtain an algorithm that simultaneously possesses three properties: (1) computes optimal solutions, (2) for any instance and (3) in polynomial time [1]. To tackle the NP-Hard optimization problems, the decision maker needs to compromise with at least one of the above stated requirements. The problems that have large inputs and which cannot be solved in polynomial time can be dealt with using heuristic algorithms. Heuristic algorithms have the advantage of computing a solution for NP-Hard or NP-Complete problems with tolerable time and space complexity at the cost of optimality of the solution i.e., one needs to compromise with the quality of the solution to generate one with acceptable time and space complexity. However, the solution computed using a heuristic is most of the times a near to optimal solution and for real life applications it is sufficient to have an approximate or partial solution.

A metaheuristic is a combination of some local improvement methods and higher level techniques or strategies. It is basically an iterative generation process that helps in searching, generating or finding a heuristic (partial search algorithm) that explores and exploits the search space efficiently, avoids getting trapped in the local optima and performs a robust search to determine the near to optimal solution for the optimization problem at hand from the solution space [2, 3]. The advantage of metaheuristics is that it can also tackle the optimization problems with nonlinearity and multimodality. These days in industry and engineering applications, the problem under consideration is extremely complex and to generate an optimal solution for such problems is a challenging task. It has been found by several researchers that metaheuristic algorithms are quite efficient for solving such problems and the latest trend is to apply nature-inspired metaheuristics for solving the critical optimization problems. Several nature-inspired metaheuristics have been developed by the researchers after studying the complex biological systems. These metaheuristics include the genetic algorithm, bat algorithm, firefly algorithm, ant colony optimization algorithm, particle swarm optimization algorithm etc. [4].

In this paper, the flower pollination metaheuristic has been implemented for the orienteering problem (OP). The OP finds a lot of practical applications especially in the fields like the tourism industry, logistics, transportation, networks etc. [5]. Few exact algorithms were suggested to solve OP [6, 7]. However, as OP is considered to be an NP-Hard problem, practically it is not possible to use exact algorithms for large instances. Therefore, the best way to tackle OP is to use some heuristic or approximation algorithms. Tsiligirides suggested the first heuristic for OP [9]. Since then several heuristics have been proposed [8, 10, 11, 12, 13, 14]. One of the latest heuristic for OP was introduced by Campos et al. [15]. Few approximation algorithms have also been proposed by Blum et al., Johnson et al., Fomin et al., etc. [16, 17, 18]. Some algorithms have also been proposed that can be applied on incomplete graphs as well [19. 20, 21]. Fuzzy and intuitionistic fuzzy versions of the orienteering problem have also been studied and a few models have been introduced to solve the problem [22, 23].

II. PREREQUISITES

2.1 Problem Definition

OP can be represented by an undirected weighted graph $G(V, E)$ where V and E signifies the set of vertices and set of edges respectively. The goal in OP is to compute a Hamiltonian path P which connects the stated source (v_1) and target (v_n), includes a subset (V') of the vertex set V such that the total collected score can be maximized with the given time budget (T_{max}). A score function $S: V \rightarrow \mathfrak{R}^+$ is associated with each vertex and a time function $t: E \rightarrow \mathfrak{R}^+$ is associated with each edge. Therefore, for a subset $E' of\ E$ and $V' of\ V$ we have $t(E') = \sum_{e \in E'} t(e)$ and $S(V') = \sum_{v \in V'} S_v$ [5]. OP can be depicted as an integer programming problem as shown below:

$$Max \sum_{i=1}^{N-1} \sum_{j=2}^{N} S_i x_{ij} \quad (1)$$

$$\sum_{j=2}^{N} x_{1j} = 1 \quad , \quad \sum_{i=1}^{N-1} x_{iN} = 1 \quad (2)$$

$$\sum_{i=1}^{N-1} x_{ik} \leq 1 \quad \forall\, k = 2, \dots \dots, N-1 \quad (3)$$

$$\sum_{j=2}^{N} x_{kj} \leq 1 \quad \forall\, k = 2, \dots \dots, N-1 \quad (4)$$

$$\sum_{i=1}^{N-1} \sum_{j=2}^{N} t_{ij} x_{ij} \leq T_{max} \quad (5)$$

$$2 \leq u_i \leq N \quad \forall\, i = 2, \dots \dots, N \quad (6)$$

$$u_i - u_j + 1 \leq (N-1)\left(1 - x_{ij}\right) \quad \forall\, i, j = 2, \dots \dots \dots, N \quad (7)$$

$$x_{ij} \in \{0,1\} \quad \forall\, i, j = 1, \dots \dots, N \quad (8)$$

Variable u_i denotes the position of vertex v_i in the path. If vertex v_j is visited after vertex v_i then x_{ij} = 1 else x_{ij} = 0. The objective function of OP which is maximization of the total collected score is denoted by Eq. 1. The constraint that a path has v_1 as its source and v_N as its target is depicted by Eq. 2. Eq. 3 – 4 ensures that no vertex is visited more than once and the path remains connected. Eq. 5 takes care of the condition that the path satisfies the given time budget (T_{max}). Eq. 6 – 7 ensures the elimination of sub-tours [5].

2.2 Flower Pollination Algorithm (FPA)

The flower pollination algorithm (FPA) is a nature-inspired metaheuristic that was introduced by Xin-She Yang in 2012 [2] and here it has been used to solve the NP-Hard orienteering problem. FPA is inspired from the pollination process of the flowers. Pollination is the reproduction process of the flowers which is carried out through agents like bees, bats, birds, insects and other animals. These agents are called pollinators. Pollination can be either abiotic or biotic and can take place in two ways, namely self-pollination and cross-pollination. When pollens are carried through some pollinators like insects then the pollination that takes place is called biotic and about 90% of the pollination activity is biotic. In 10% of the cases, pollens are carried through natural carriers like wind, water etc. and are termed as abiotic. If the pollination takes place with pollen from a flower of a different plant then it is called cross-pollination and if the fertilization happens due to the pollen coming from either the same flower or a different flower of the same plant then it is termed as self-pollination. Another term that can be associated with this process of pollination is the flower constancy. Honeybee is a pollinator that helps in implementing this phenomenon called flower constancy where the pollinators tend to visit only a specific species of flowers and ignore the other species that exist. Pollinators like birds, bees, insects etc. can fly to long distances. Therefore, biotic, cross-pollination over long distances can be termed as global pollination. Also, the behaviour of the biotic pollinators i.e., their jumps and flying distances etc. follows the Levy distribution as stated by Xin-She Yang [2].

III. ALGORITHM FPA_OP

Input: A graph $G(V, E)$ with t_{ij} (time taken to traverse) value of each edge (e_{ij}) connecting vertex v_i and $v_j \in V$, S_i (score) value of each vertex $v_i \in V$.

Output: A Hamiltonian path with the highest possible collected score such that total travel time is within the specified time budget T_{max}.

```
$Initialize\ a\ population\ of\ n\ flowers$
$/pollen\ gametes\ with\ random\ solutions$
$Find\ the\ best\ solution\ R^*\ in\ the\ initial\ population$
$Define\ a\ switch\ probability\ sp\ \in [0,1]$
$\textbf{while}\ (t < No\_of\_Iterations)$
        $\textbf{for}\ i = 1$
        $:\ n\ (all\ n\ flowers\ in\ the\ population)$
                $\textbf{if}\ rand < sp$
                        $Do\ global\ pollination\ via$
    $index2 = index1 + L(index(R^*) - index1)$
        $\textbf{else}$

        $Randomly\ choose\ j\ and\ k\ among\ all\ the\ solutions$
          $Do\ local\ pollination\ via\ index2 = index1 +$
        $\epsilon(j - k)$
        $\textbf{end if}$
        $Evaluate\ new\ solutions$

        $If\ new\ solutions\ are\ better, update\ them\ in\ the\ populat$
        $\textbf{end for}$
        $Find\ the\ current\ best\ solution\ R^*$
$\textbf{end while}$
```

In the above stated algorithm [2], initially n pollens are generated randomly where each pollen represents a possible solution i.e., a path satisfying the constraint that the total time taken by the path is less than the upper bound T_{max}. Then for the valid paths three values are evaluated: (1) the total time taken by the path, (2) the total score collected by the path and (3) the ratio of $score/time(S_i/\ t_{ij})$ for each of the valid paths. Then these paths are stored in a priority queue on the basis of the $(S_i/\ t_{ij})$ ratio. To determine the best solution R^* in the initial solution, any of the selection methods like the tournament selection, random selection, (μ, λ) selection, roulette wheel selection etc. can be implemented. The switch probability $sp \in [0,1]$ controls the type of pollination to be performed i.e., global pollination or local pollination. $rand$ is a randomly generated number and if it is less than sp then global pollination is performed else local pollination is performed. In case of global pollination, a function is used that randomly generates the value for $index1$. Then using $index1$ and Levy distribution L , the value for $index2$ is computed. The value for L is calculated using the following equation:

$$L \sim \frac{\lambda \Gamma(\lambda) sin(\Pi\lambda/2)}{\Pi} \frac{1}{s^{1+\lambda}} \quad (4.10)$$

Where, $\lambda = 1.5$ and s denotes the step size. To make the step size appropriate, so that neither it is too large nor too small, its value is calculated as

$size\ of\ the\ priority\ queue/5$. Similarly, in local pollination again the value of $index1$ is considered and two randomly generated solutions j and k are used to calculate the value for $index2$. The value for ϵ is randomly chosen from the interval [0,1]. After the two parents are determined (i.e., the paths at $index1$ and $index2$), a crossover operation is performed to compute the child paths. Two points are found randomly in the parent2 (at $index2$) and the nodes that exists between the two points of parent2 are inserted in parent1, one by one at the best possible location (which leads to minimum increment in the total time taken). This way child1 is formulated. In a similar manner, two points are selected in parent1 and the nodes lying between the two points are added one by one to parent2 at its best location and this way another path i.e., child2 is created. Then it is checked whether these new paths (child1 and child2) are valid or not i.e., their total time taken is less than T_{max} or not. If the new path is valid then it is stored in the queue else it is discarded. If the new path that has been generated has a better score than the previous best solution, then it is updated in the queue. At the end, when the algorithm terminates, the path obtained with the highest value of total collected score forms the final solution.

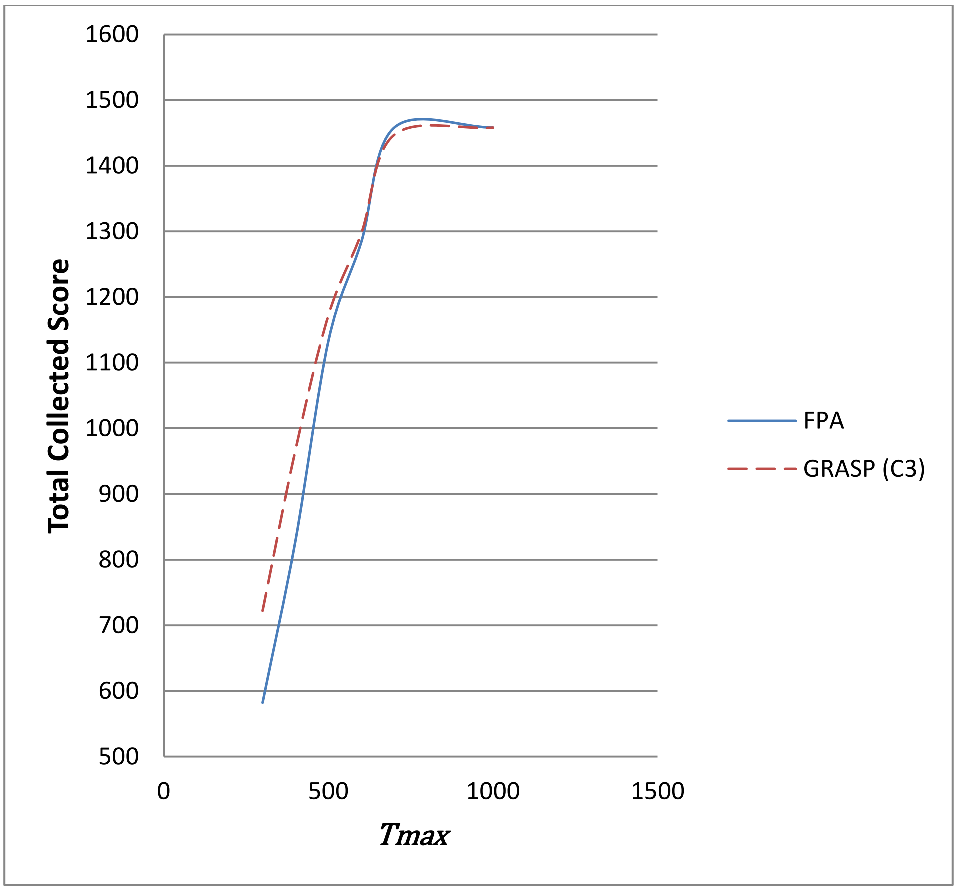


Figure 1. Comparison of the total collected score value achieved by GRASP and FPA algorithms for different T_{max} values when applied on a graph with 102 nodes, source=1, destination=102.

IV. EXPERIMENTAL ANALYSIS

The code for FPA_OP was developed in C++ and compiled using CodeBlocks on an Intel Core i5 650 at 2.20 GHz. The code was implemented on instances with 32, 33, 64, 66, 102, 150 etc. nodes. Each instance represents a complete graph. The results obtained for FPA_OP were compared against the best known algorithm in the literature for OP viz. the GRASP algorithm suggested in [15]. The results obtained (average of 10 runs) by FPA_OP were compared with the C3 method of GRASP and it was found that FPA_OP helps in obtaining higher total collected score value for larger T_{max}. However, at lower T_{max} values the results obtained through GRASP were better than FPA_OP. Figure 1 is a plot for a complete graph with 102 nodes, source=1 and destination=102. The results for the total collected score by GRASP and FPA_OP algorithms are compared for different T_{max} values and the observation stated above can be clearly seen in the plot. Therefore, the FPA_OP algorithm can be preferred in the cases were achieving a higher total collected score is the priority and the delay in time can be tolerated.

V. CONCLUSION

A meta-heuristic called the flower pollination algorithm suggested by Yang [2] has been implemented for OP (FPA_OP) and the results thus obtained for instances with different number of nodes has been compared with those obtained by running the GRASP algorithm [15]. It was found that in situations where achieving a better score is the priority at the cost of time delay, FPA_OP algorithm can be preferred as it helps in obtaining a higher total collected score than GRASP for larger values of T_{max}.

VI. REFERENCES

[1]. Williamson, D.P., and Shmoys, D.B. The Design of Approximation Algorithms, Cambridge University Press, (2010).

[2]. Yang, X.S. Flower Pollination Algorithm for Global Optimization, LNCS 7445, (2012): 240–249.

[3]. Yang, X.S., Karamanoglu, M. and He, X. "Multi-objective Flower Algorithm for Optimization." Proceedings of the International Conference on Computational Science, ICCS 2013, (2013): 861 – 868.

[4]. Fister, Jr. I., Yang, X. S., Fister, I., Brest, J., and Fister, D. "A Brief Review of Nature-Inspired Algorithms for Optimization." Elektrotehniski Vestnik 80, no. 3 (2013): 1–7.

[5]. Vansteenwegen, P., Souffriau, W., and Oudheusden, D. V. "The orienteering problem: A survey." European Journal of Operational Research 209, (2011): 1–10.

[6]. Laporte, G., and Martello, S. "The Selective Traveling Salesman Problem." Discrete Applied Mathematics 26, (1990): 193-207.

[7]. Hayes, M., and Norman, J. M. "Dynamic Programming in Orienteering: Route Choice and the Siting of Controls." Journal of the Operational Research Society 35, no. 9 (1984): 791-796.

[8]. Fischetti, M., Salazar, J., and Toth, P. "Solving the orienteering problem through branch-and-cut." INFORMS Journal on Computing 10, (1998): 133–148.

[9]. Tsiligirides, T. "Heuristic methods applied to orienteering." Journal of the Operational Research Society 35, (1984): 797–809.

[10]. Ramesh, R., and Brown, K. "An efficient four-phase heuristic for the generalized orienteering problem." Computers and Operations Research 18, (1991): 151–165.

[11]. Golden, B., Levy, L., and Vohra, R. "The orienteering problem." Naval Research Logistics 34, (1987): 307–318.

[12]. Wang, Q., Sun, X., Golden, B. L., and Jia, J. "Using artificial neural networks to solve the orienteering problem." Annals of Operations Research 61, (1995): 111–120.

[13]. Gendreau, M., Laporte, G., and Semet, F. "A tabu search heuristic for the undirected selective travelling salesman problem." European Journal of Operational Research 106, (1998): 539–545.

[14]. Schilde, M., Doerner, K. F., Hartl, R. F., and Kiechle, G. "Metaheuristics for the bi-objective orienteering problem." Swarm Intelligence 3, (2009): 179-201.

[15]. Campos, V., Marti, R., Sanchez-Oro, J., and Duarte, A. "GRASP with Path Relinking for the Orienteering Problem." Journal of the Operational Research Society 2013, (2013): 1–14.

[16]. Blum, A., Chawla, S., Karger, D. R., Lane, T., Meyerson, A., and Minkoff, M. "Approximation Algorithms for Orienteering and Discounted-Reward TSP." Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS'03), (2003): 1-10.

[17]. Johnson, D., Minkoff, M., and Phillips, S. "The prize collecting steiner tree problem: Theory and practice." Proceedings of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms, (2000): 760–769.

[18]. Fomin, F. V., and Lingas, A. "Approximation algorithms for time-dependent orienteering." Information Processing Letters 83, (2002): 57–62.

[19]. Ostrowski, K., and Koszelew, J. "The Comparison of Genetic Algorithms which Solve Orienteering Problem using Complete and Incomplete Graph." Informatyka 8, (2011): 61-77.

[20]. Verma, M., Gupta, M., Pal, B., and Shukla, K. K. "Roulette Wheel Selection based Heuristic Algorithm for the Orienteering Problem." International Journal of Computers and Technology 13, no. 1 (2014): 4127-4145.

[21]. Verma, M., Gupta, M., Pal, B., and Shukla, K. K. "A Stochastic Greedy Heuristic Algorithm for the Orienteering Problem." Proceedings of the 5th International Conference on Computer and Communication Technology (ICCCT), (2014): 59-65.

[22]. Verma, M., and Shukla, K. K. "Application of Fuzzy Optimization to the Orienteering Problem." Advances in Fuzzy Systems 2015, (2015): 1-12.

[23]. Verma, M., and Shukla, K. K. "Fuzzy Metric Space Induced by Intuitionistic Fuzzy Points and Its Application to the Orienteering Problem." IEEE Transactions on Fuzzy Systems, 24, no. 2 (2016): 483-488

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Multiple Instances Based Emotion Detection Using Discriminant Feature Tracking

Abhishek Kilak, Namita Mittal
Malaviya National Institute of Technology, Jaipur, Rajasthan, India

ABSTRACT

Automatic recognition of emotions from the facial expressions continues to be an important aspect in the field of evolution of new age computing systems. Emotion detection capable computer systems is an ongoing research field that has a numerous applications ranging from recommender systems, human-computer interaction, robotics and affective systems. The various features that increase the complexity of emotion recognition systems include ethnicity, gender, pose, occlusion, beard, moustache etc. The type of database used for learning by systems is of crucial importance. Many databases exist for this purpose but none of them is for posed Indian faces. In this paper,we bridge this gap by providing Bharat Database which contains facial images of Indian people. The database has posed expressions of 102 participants and has 896 images. The participants were asked to pose for different emotions by showing them images eliciting those emotions as well as with the help of expert artists. The annotation was done using polling by a panel of three experts.Experiments were conducted on the database using different algorithms and results are presented for reference. This database will further help the community involved in developing of algorithms for emotion recognition. In this paper we propose two emotion detection approaches, the first one is based on Compact Local Binary Pattern and is used for construction of hybrid features which increases emotion detection accuracy. Second approach is Enhanced Feature Extraction using multiple Patches face on images from indigenously developed Bharat Database of Indian Faces, Japanese Female Facial Expression Database and Karolinska Directed Emotional Faces. The results show its applicability for construction of emotion detection systems.

Keywords : Emotion detection, facial expression, emotion recognition, facial images, Indian faces.

I. INTRODUCTION

An acceptable approximation of emotion of people automatically is indispensible information for affect enabled systems. There are various approaches for emotion detection but still challenging is their availability, unintrusiveness and accuracy.

Emotional states of human beings can inferred from the gestures, text analysis, voice analysis, brain mapping, breath analysis, keyboard pressure information and face expressions. The keyboard and mouse approach for gathering log of user action is one of the simplest ways of predicting human emotions, but nonetheless it lacks inaccuracy [26].Also keyboard pressure analysis for classification of emotions suffers from having a limited scope only as people are not constantly using keyboards and also all but lab equipment are not equipped with necessary hardware to deliver the required information required for mapping them to emotional states. On the other hand there are approaches that based on sensors that are intrusive and gather physiological signals by way of EEG [27] [28] and ECG [29]. In between the two extremes are approaches that rely on breath analysis [30], gesture analysis [31], [32] and facial information [33].

Exhibition of emotions through gestures and face expressions by humans is the most common aspect. The reverse engineering of analyzing emotions from these expressions is a very common fundamental for humans. They are capable of communicating among themselves through the exchange of these mutually dependent parameters. Gestures may be controlled by people voluntarily and thus may not be coherent with the emotional state. Text analysis is generally a binary analysis of either positive or negative state and thus may provide information only about valence. While emotional state comprises of not only valence but orthogonally arousal also. Brain mapping is an important technique to predict emotional states based on surges of neurons of human brain. This technique is more lab based and not practical as the subjects under consideration are always aware of their examination and thus actual emotional state may be not be predicted. Also this technique involves we ring of monitoring caps which is obtrusive, limited to lab environment and impractical in real life situations. Breath analysis method for emotion detection is again an obtrusive technique and sufferers from the same limitations as of brain mapping technique. Humans are capable of reading gestures and facial features and understanding them near precision. But when it comes to computers, the research is novice and there is a great scope of improvement.

Face expressions provide an insight into the emotional state of human beings. Face expressions change in tandem with emotional states and people generally have limited or no control over their facial movements. Also if people might try to control their expressions in some specific temporal space but expressions are generally spontaneous and in tandem to their emotional state. Therefore emotion detection based on analysis of facial features is undoubtedly the best way to detect emotions as it is not limited to lab environment, unobtrusive, coherent to emotional state and practically applicable.

The real life applications of this research are enormous and include but not limited to Artificial Intelligence, Affective computing, Driver assistance systems, Robotic interactions, Surveillance systems, Mob control systems, Military applications, health care support systems, personalized recommenders for online as well as real life shopping, Content delivery for social media, Movie Target Audience analysis, sentiment analysis of public towards political and business decisions, tailored support solutions for differently abled persons, anxiety bipolar and depression detection, crowd behavior analysis and Smart City as well as IoT applications.

Detection of emotions from face expressions involves extracting mathematical information from spatial variations of the faces. This spatial information also invariably involves ethnicity, gender, scale correction as well as noise features involving occlusion submounting to beard, moustache, hairlocks, face Rotation, resolution and physical hindrances towards full face detection.

Development of systems capable of emotion classification invariably requires training to catch the strains that effectively and categorically map them to different categories. This in turn requires data for the systems to train, which has the following two aspects. Either the training data can be too personalized to serve for only particular subject involved. For eg: Personal response systems based on specific requirement of individuals. This involves catching individual data and catering tailor made solutions bases on them which cannot be applied to pro rata basis. Or, there may be robust general systems that are capable enough to provide services to the unknown masses based on rigorous training provided that involves non ideal conditions of data parity, ever unknown individuality and other parameters mentioned in the previous paragraph.

This requires creation of database that is essential to train the systems, test them, validating of applicability of algorithms for emotion classification systems and building of robust systems. Database creation is toilsome and tedious task. But the less is the fact that it is a prerequisite to build systems that are capable of detecting emotions. Creation of posed expression database requires proper guidance and training of the subjects by experts. Thereafter validation of database is an equally important concern. This involves labeling of the images with categories that they belong to.

In literature Paul Ekman [1] reported six basic emotions that are valid across human species irrespective of gender and ethnicity. These are happy, sad, angry, fear, surprise and disgust. Humans can predict the emotional state of people of different ethnicity. But such is not the case with computers as they need to be trained for different for the texture, shape and appearance of the target subjects. Because of different shape and texture of people across globe, headway in the field of affective computing requires different databases covering ethnicity. To our knowledge there is no posed expression database for Indian faces that catches all the above mentioned emotion classes.

For capturing and building of affective systems it is a prerequisite to capture these spatio-temporal displacements happening in facial features due to underlying muscles. The spatial and temporal assessment of prominent facial features may be utilized for categorization of emotions. Many facial emotion techniques have been proposed which have considered 2D images, 3D images [34], [35], expressions exhibited by infants, AAM [36] based systems and AU based systems [37].

Non deliberate and deliberate ace expressions are the two categories of face expressions defined by Battocchi et al. [38], Expression which are deliberate are expressed under the absence of speech. Whereas those expressions exhibited along with speech are termed as non-deliberate. Also Valence, Arousal and Dominance multidimensional space is sometimes used for separating the emotions categorically.

This requires creation of database that is essential to train the systems, test them, validating of applicability of algorithms for emotion recognition and classification systems for building a robust systems. Database creation is toilsome and tedious task. But the less is the fact that it is a prerequisite to build systems that are capable of detecting emotions. Creation of posed expression database requires proper guidance and training of the subjects by experts. Thereafter validation of database is an equally important concern. This involves labeling of the images with categories that they belong to.

We propose two approaches to automatic recognize six basic emotions. First is compact local binary pattern representation of the images for extraction of features which are used for construction of hybrid feature vectors for classification of emotions. The other approach is Enhanced Feature Extraction using multiple Patches face on images from indigenously developed Bharat Database of Indian Faces, Japanese Female Facial Expression Database and Karolinska Directed Emotional Faces. The results show its applicability for construction of emotion detection systems.

The main contributions of this work include:

1. Construction of Database of Indian Faces with emotion annotation.
2. Propose a novel method for Compact Local Binary Pat- tern.
3. Integrate it with Histogram of Oriented Gradients for classification.

4. Propose a novel technique for extraction of features using multiple patches face on images.
5. A challenging facial expression database of Indian faces is introduced with benchmark for comparative analysis.
6. The results from the experiments shows robust performance by macro analysis of person independent emotion detection.

The rest of the paper is organized as follows. Section 2 introduce the relevant works on existing database for facial emotion and techniques for dynamic analysis of facial expression. Section 3 described the procedure of creation of proposed database. Section 4 contains the description of dataset. Section 5 and 6 illustrated the proposed work of feature extraction and emotion detection respectively. Further, experimental evaluation and classification results are given in section 7. Section 8 conclude the paper with a summary and discussion.

II. RELATE DWORK

As presented work participated in both database creation as well as emotion detection techniques, to maintain the information related to both of the domains we have presented the related work in two fold. First we present different techniques used for dynamic facial expression detection. Thereafter we discuss about the various databases available for facial expression detection.

2.1 Existing Techniques used for Facial Expressions

Various techniques have been proposed in recent times to classify emotions based on facial expressions. The features are extracted from the images or videos using appearance based and geometric approaches.

Geometric feature based approaches dependent upon fiducial points of the facial images [14]. The classification is based upon recognition of movement of facial points in spatial domain. The temporal parameters can also be used for recognition of expressions by analyzing shapes across frames of videos. Many researches thrive upon posed facial expressions [15] for emotion recognition.

Texture in images can be captured using Local binary patterns [12]. LBP is computationally simpler and also is tolerant to illumination conditions. LBP features on orthogonal planes of space and time are applied by Zhao et al. which led to introduction of temporal features augmenting the spatial textural features [13]. Local Phase Quantization [16], Histogram of Gradients (HoG) [17], LBP and SIFT are the commonly used low level features for feature extraction and classification and may be extended to include temporal features for real time applications. Being simpler computationally, robust in extraction of textural features and having the potential extendibility to include features from temporal domain makes LBP a popular choice foe feature extraction [18]. An extended approach of LBP comprising circular neighbors of different radius and showing a further discriminating presentation has also been proposed [19].

Technique based on blocks is generally used taking into account the locations of subregions and co-occurrences [20], [21], [24]. Use of this technique was done primarily for recognition of faces [22]. Image was partitioned in regions that were non overlapping and Local Binary features were derived and weighted according to their importance in contributing to identification and then appended for formation of augmented representation. This method was further extended by Zhao et al. by considering overlapping blocks for the experiments [23]. However, person related bias and resultant over fitting was not care of in their experiments. Weight based approach was considered by Shan et al. [22], wher only expressions were considered and not identities [25]. However it lacked the leverage of using overlapping blocks.

For detection of emotions localization of mouth and eye part has been performed in many approaches [39]. This technique suffers from loss of data as the rest portion of the face is not considered. Face motion tracker method has been proposed by Shen et al [40] which is composed of combination of multiple models for construction of cost function. The constituent models of the cost function are varied according to the application scenario.

Viola Jones is widely used algorithm for detection of face part in the images [41]. Thereafter Gabor filters are applied at different scales and orientations for building of feature vector [42]. This feature vector is then used for classification of emotions in different categories. Active Shape model is used by some researchers which has the drawback of using only shape constraint information. This is overcome by a widely used approach for tracking of faces and emotion detection is the Active Appearance Model which is a geometrical approach that uses texture information also. AAM is used for the matching of a statistical model of the shape of facial features upon actual facial image. Supervised training is done using landmark coordinates which are posted on the training image set [43].

The manual location of feature points was done by Wang et al [44] at prominent features nearing eyes, mouth and eyebrows for detection of emotions. However it is not practical as it lacks positioning of these lines and dots in the impending images after training. Kernel methods based on time series have been used for recognition of emotions upon landmark data by Lorinez et al [45]. The have shown that Global Alignment Kernel or Dynamically Time warping are required for emotion categorization. A system registering facial expressions based on series of steps that detects emotions has been proposed by Sariyanidi et al. [46]. The registration of rigid, parts and point parameters is done for encoding the image sequences frame by frame. Shapes are represented using coordinate pairs of a sequence of facial points. The bulky information is then reduced dimensionally and the resultant is used for emotion recognition. Various statistical measures are used which are complex in nature and recognition is done on posed data only.

Multiscale learning based upon high and low level spatio- temporal facial expressions has been proposed by Liu et al [47]. The high level features constitute different gestural events which are assessed for different duration. Low level features are comprised of information obtained from head pose, appearance and face geometrical features. However the system is very complex to be deployable in real life situations.

2.2 Existing Databases of Facial Expressions

Automatic recognition of emotions from the facial expressions continues to be an important aspect in the field of evolution of new age computing systems. The various features that increase the complexity of emotion recognition systems include ethnicity, gender, pose, occlusion, beard, moustache etc. Therefore there had been an emergence of several databases covering various features that are available publicly emotion recognition. The different databases are discussed in brief in this section.

One of the most widely used database is Cohn-Kanade Action Unit Coded Facial Expression database. It comprises of 486 image sequences posed by 97 subjects was released in year 2000. The image sequence proceeds from neutral face image to extreme expression. The peak expression images were coded using Facial Action Coding System and annotation is provided in form of emotion labels. It had the limitation as the emotion labels were not validated. Rather the emotion labels were those which were asked from the subjects to perform.

This dataset was extended to address as Extended Cohn- Kanade (CK+) database [2]. The images of 123 subjects corresponding to 593 sequences were taken which were coded using Facial Action Coding System for the last or peak frame of the sequence. The Action Units and their intensity were provided for the peak expression images. Also the images were tracked using Active Appearance Model for 68 landmark points. Out of the total sequences only 327 were having corresponding emotion files. This was because only these sequences were validated. The emotion labels were neutral, anger, contempt, disgust, fear, happy, sadness and surprise.

Japanese Female Facial Expression (JAFFE) database has 213 images 10 female Japanese models that posed for both neutral and six basic expressions. JAFFE database was created by Lyons et. al. [3]. The images are in grayscale.

Binghamton University-3D Dynamic Facial Expression database [4] has 2500 3D face expressions of 100 subjects having the six basic expressions having four intensity levels. The different aspects considered include age, race and culture.

The MMI Database was initially conceived in 2002 with the goal of serving as a source that can be used across facial expression recognition community [5]. It has videos that has sequence from neutral to apex and back to neutral expression. It has over 2900 videos of 75 subjects as well as still images. The videos are annotated for presence of Action Units.

The Belfast Database [6] has different sets of over 250 coloured video clips depicting natural emotions at different resolutions. Multimedia Understanding Group (MUG) has 1462 posed color sequences of 86 subjects which are annotated with emotion labels. The Radboud Faces Database (RaFD) has posed color images of 67 subjects at five different camera angles and three different gaze directions for eight emotion labels.

Indian Spontaneous Expression Database (ISED) [7] has 428 spontaneous color videos of 50 subjects having emotion labels of sad, happy, surprise and disgust only.

Denver intensity of spontaneous facial action database (DIFSA) [8] is a color video database of 27 subjects whereby each video sequence is of 4845 frames of spontaneous reactions while viewing video of 4 minute duration. Six intensity level annotation of Action Units is provided for the facial expressions.

III. CREATION OF THE DATA BASE

The BDIF has still posed facial images for various emotions. Subjects were asked to pose for all the basic as well as neutral emotion. The participants were asked to pose for different emotions by showing them images eliciting those emotions as well as with the help of expert artists. The photographs were taken in well lit conditions.

The BDIF was carefully constructed by showing the subjects valid labeled emotion images from different databases and also with the help of expert artists training them how to elicit the said emotions. The effectiveness in expressing emotion was due changing the mental state of the subjects by showing them visual cues and also narrating real life situations which pertain to those particular emotions.

The annotation of images was done using polling by three annotators who were familiar with Facial action coding system. They were shown images and told to classify them as belonging to one of the classes. Only those images were labeled in which a consensus arrived among the annotators.

3.1 Experimental Setup

The subjects participated voluntarily for the posing of expressions. They were made comfortable by telling as how to pose and showing visual cues as well as real life situations which are associated with the particular emotional state. The subject being comfortable with the environment and counseled properly led to the capture of expressions effectively. These images were captured in ideal conditions which was free from noise and other distractions. The images were captured in ambient light conditions. According to experience obtained from preliminary studies, the light conditions were as subjects are generally accustomed to and not being very bright or dull. These made the subjects comfortable with the experiment environment.

Closed rooms were used for the conduction of shoot sessions. The subjects were made well aware about the experiment and also as how it will help in future scientific studies.

The subjects were asked to stand comfortably taking the sup- port of wall. They were allowed time to ease them for elicitation of emotions. Thereafter images were captured by posing for neutral, anger, contempt, disgust, happy, fear, sadness and surprise. In order to avoid disturbance the rooms were kept close during the different shooting sessions. In first part of the experiment it was found after annotation that the sample of anger and disgust emotion was the least. The statistics of first phase is shown in Figure 1.

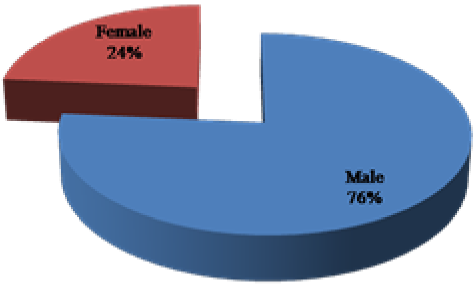


Figure 1. Male Female Ratio.

The images were taken comprising of 4320 × 3240 pixels at 300 dpi horizontal and vertical resolution using Nikon Coolpix 120. The images were taken with compulsorily no flash. Distance maintained between subject and the camera was around 1 meter. The emotion label categories for the complete database is shown in the Figure 2.

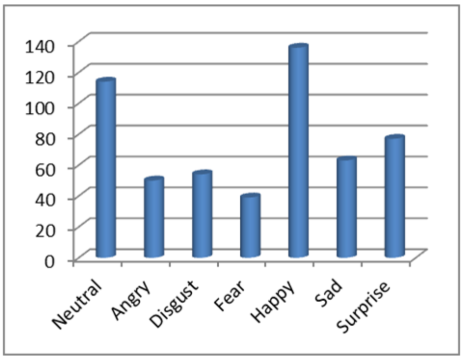


Figure 2. Graph showing the emotion label categories for the complete database.

3.2 Obstructions

Gathering information from facial expressions may be hindered by the presence of glasses, moustache, beard etc. These act as noisy features in implementation of automatic emotion detection algorithms. However they are a regular feature in the real life situations. Therefore these features were deliberately included in the formation of database so that more robust emotion detection algorithms can be built using such images also from the database.

3.3 Annotation

The images obtained for the database were subjected to labeling for different emotions. This is particularly important as it serves as ground truth for comparison with results obtained by applying automated algorithms. Therefore the technique of validating emotion labels is substantial significance. The images were labeled by a panel of three annotators who were familiar with facial action coding system. The labeling was done for neutral, happy, angry, disgust, contempt, surprise, sad, fear and invalid emotions. For validation polling was done for each image by gathering the emotion labels of all the three annotators . Only those images which had a majority for particular emotion label were validated for that particular emotion. Rest of the images were not annotation validated for the particular emotion labels.

The resultant database consisted of:-

Total Subjects-59
Images Collected-436

Total Classes of Valid Emotions: 7

1) Neutral : 71
2) Angry: 24
3) Disgust: 14
4) Fear: 3
5) Happy:91
6) Sad: 40
7) Surprise: 47

The results of the first phase of database creation led to a finding that even when the subjects were made comfortable with the environment, they could not provide fear and disgusted motions. This was because people do not feel comfortable to exhibit these emotions willingly and publically. Therefore in the next phase, the subjects were given privacy and shoots were performed in confined places so that they can comfortably exhibit those emotions. Also subjects were provided visual cues as what that expression looks like. Assistance was provided to them with the help of professional artist to help them exhibit those emotions.

The database was extended in the next phase by including new subjects and taking care of findings of the first phase. The extension led to:
New Subjects-43
Images Collected-460

Total Classes of valid Emotions: 7

a. Neutral : 43
b. Angry: 23
c. Disgust: 39
d. Fear: 36
e. Happy:41
f. Sad: 23
g. Surprise: 36

3.4 Consent from Subjects for publication

Initially the subjects were reluctant to pose for the Database. The subjects involved in the collection of dataset were given incentive like chocolate, juice and ice cream. The participants were informed that their images may be used for publication for research purpose. Ethically the subjects were asked to fill up their consent for publication of images for research purpose. Images of those subjects who did not provide consent were deleted from the final database.

IV. BHARAT DATA BASE

BDIF initially had expression images of 59 subjects for seven classes of emotions viz. neutral, happy, angry, fearsome, disgust, surprised and sadness. It initially had 436 images. Thereafter it was extended by addition of expression images of 43 new subjects exhibiting the seven classes of emotions. 460 new images were added in this phase. The complete database thus has total 102 subjects and a total of 896 images. The class wise distribution is shown below:-

1) Neutral : 114
2) Angry: 50
3) Disgust: 54
4) Fear: 39
5) Happy:136
6) Sad: 63
7) Surprise: 76

Total Male : 78
Total female : 24

Sample images of different class of emotions can be seen in the Figures 3 - 9.

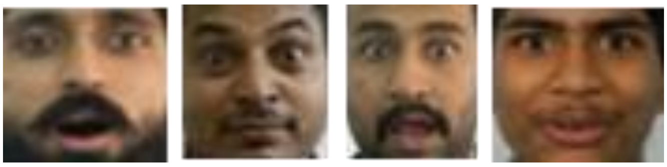


Fig. 3. Images from Bharat Database of Indian Faces exhibiting Surprise emotion.

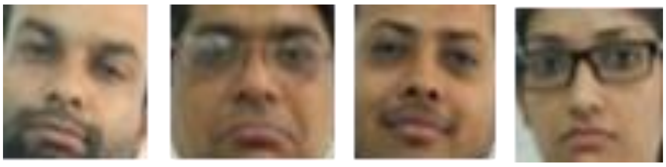


Figure 4. Images from Bharat Database of Indian Faces exhibiting Sad emotion.

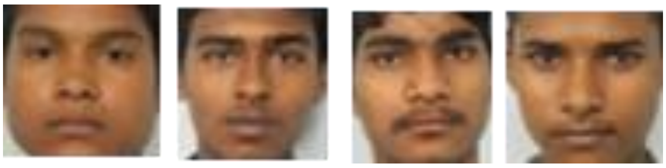


Figure 5. Images from Bharat Database of Indian Faces exhibiting Neutral face.

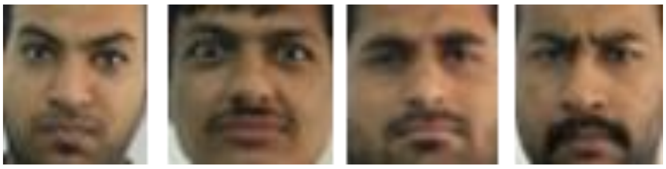


Figure 6. Images from Bharat Database of Indian Faces exhibiting Anger emotion.

Access to this database may be made available for research purpose only by sending an Email to the author at 2012RCP9516@mnit.ac.in.

V. PROPOSED WORK I

The steps involved in emotion detection involve the following steps:

1) Image Preprocessing
2) Extraction of Feature Vector
3) Feature Vector Vector Size Reduction
4) Emotion Classification
5) Compact Local Binary Pattern extraction
6) Statistical moment and other feature calculation and ap- pending
7) Classification

5.1 Preprocessing

Viola Jones [9] algorithm was used for detection of the face part from the images of Bharat Database of Indian Faces which resulted in 100% accuracy on our dataset by adjusting minimum permissible height and width to 600 pixels. Cropping of images was done so that the images may contain only the face part. The images of the database are colored and they were converted to grayscale for the conduction of experiments. Sample of images which were converted to grayscale and cropped are shown in The Figure 10. The images were resized to 1024*1024 pixel size.

5.2 Extraction of Features

Histogram of Oriented Gradients were calculated for the grayscale cropped images that had corresponding emotional tags. Edge directions distribution or gradient intensity are able to catch shape of an object and local appearance. Division of image into cells of uniform length and breadth was done in the experiment for cell sizes ranging from 10 × 10 pixels to 512 × 512 size.

Gradient directional histograms were then computed for the cells. HOG blocks of 2 ∗ 2 cells with 50% overlap were used for the construction of feature vector comprising of orientation binned cell histograms. Sample Histogram visualization is shown in the following Figure 11.

$$T_{Blocks} = floor\left(\frac{Len}{Cell} - 1\right) \mathrm{X}\, floor\left(\frac{Bre}{Cell} - 1\right)$$

Where used to denote total number of blocks in an image. Since every block is composed of 2 ∗ 2 cells, therefore

$$\mathrm{TCells} = \mathrm{TBlocks} * 4 \qquad (2)$$

Where T_{Cells} represents total number of cells that participate forthe computation. The total number of histograms per cell used in this work is given in Equation 3.

$$Hist_{Cell} = 9 \qquad (3)$$

5.3 Reduction of Feature Size

Histogram features for gradient orientations [10] were extracted for cell sizes ranging from 64*64 pixels to 512*512 pixels. Experiments were conducted on this feature vector and reduced feature vector which was done by applying principal Component Analysis [11] covering variance of 0.95.

5.4 Extraction of features using Proposed Compact LBP

A new approach is proposed as Compact LBP in which the neighbouring pixels of the previous were discarded in the resultant image. This resulted in image size reduction from 1024*1024 pixels to 341*341 pixels.

The conversion process is shown in the following figures with the help of sample pixel values for an image subsection.

The resultant Compact LBP Images are shown in the following figure.

5.4.1. Calculation of Statistical moments and other features and appending

For each Compact local pattern image mean and median were calculated for each pixel row. This provided feature vector of length 341 + 341 i.e. 682. To this 51 histogram counts were added for each image. Also mean of the whole image and standard deviation were calculated. This resulted in feature vector of length 735 for each image. The results were calculated for classification using different classifiers. In the next phase results were classified appending this feature vector to the feature vector histograms of oriented gradients.

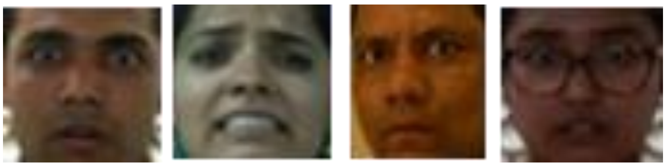


Figure 7. Images from Bharat Database of Indian Faces exhibiting Fear emotion.

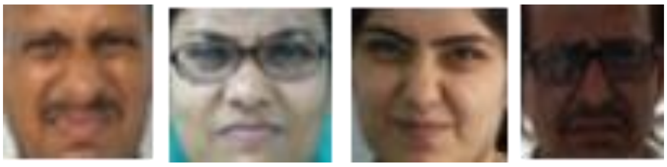


Figure 8. Images from Bharat Database of Indian Faces exhibiting Disgust emotion.

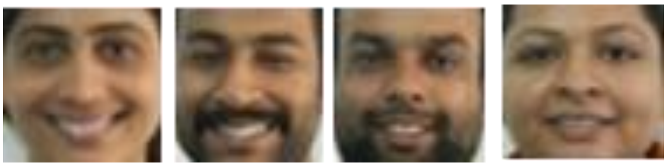


Figure 9. Images from Bharat Database of Indian Faces exhibiting Happy emotion.

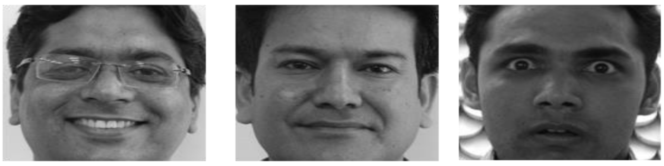


Figure 10. Grayscale Cropped Image samples.

VI. PROPOSED WORK II

The procedure for computing structural Features is illustrated in Algorithm 1.

The first database used for the experiments was Bharat Database of Indian Faces which is composed of total 102 subjects and a total of 896 images. This database has images of school going children of 11th and 12th class, subjects doing graduation, post-graduation and research scholars, staff members from different offices and random people. There are occlusions of beard, moustache and spectacles.

The class wise distribution is shown below:-

Neutral: 114; Angry: 50; Disgust: 54; Fear: 39; Happy:136; Sad: 63; Surprise: 76

Total Male : 78

Total Female : 24

6.1 Marking of points on Neutral faces

This was followed by marking of points on the neutral faces of these subjects. The neutral face in the images was subjected to tracking points manually. These points covered eyebrows, eyes, nose and lips of the subjects. The reason for this allotment was that these parts of the faces convey the most information. In literature associated to physiology these parts contemplate to Action Units. Sample image with point marked is shown in Figure 17.

6.2 Tracking of the points in corresponding images

The points established in the previous phase on the neutral images were averaged in this phase. Matrices corresponding to 5 × 5

grid around pixel points of considered. The pixel values of these points were taken and average value was calculated.

R[ ]= R(r – I R + 3, c – I C + 3)

G[ ]= G(r – I r + 3, c – I c + 3)

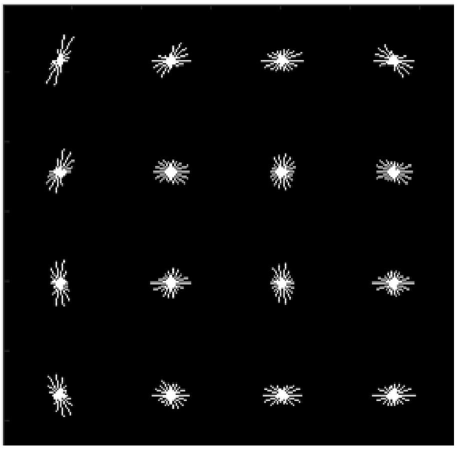


Figure 11. Cropped Image Histogram visualization for cell size 256.

B[ ]= B(r – I r + 3, c – I c +3) Where r = I R – 2 to I R + 2 and c = I C – 2 to I C + 2

Where I R and I C correspond to initial Row and Column. Average Color Channel values were calculated for each color.

These Average Color Channel values were used to track these points in the corresponding emotion images of the subjects. Maximum permissible distance of 15 pixels in both directions was used in searching of probable points within the periphery of initial points of consideration.

$$AvgR1 = \sum_{i+2;j=2}^{i-2;j-2} \frac{R1}{25}$$

For each pixel in periphery of 15 × 15.

6.3 Calculation of distance vector

In this step the Euclidean distance between the initial point positions and the tracked point positions. This is particularly important as the displacements capture information that may be

115	121	214	56	78	98
87	98	76	211	58	67
123	124	178	189	111	48
76	78	67	90	99	111
99	100	102	104	88	76
244	235	68	89	96	89

Figure 12. Initial image sample pixels.

1	1	1	0	1	1
0	98	0	1	58	1
1	1	1	1	1	0
0	0	0	1	1	1
0	100	1	1	88	0
1	1	0	1	1	1

Figure 13. Applying Binary pattern on the immediate neighbors of the pixels shown in red color.

used to categories among various emotions. The distance vector is composed of these corresponding displacements.

$$D = \sqrt{\left(X_i - X_j\right)^2 - \left(Y_i - Y_j\right)^2}$$

Algorithm 1 Mark Point Technique

```
Ensure: A – Emotion Accuracy
Require: Ii -Image i = 1......T ∗ t
    For each image Ii do
    Take a netural emotion image of the subject.
    Crop and resize
    Mark the points(N )
    Read coordinatorof all the points
    Get R, G and B channel values in windowof –2 to 2 x and y values around
    Calculate Average R, G and B values for each window around N points
    end for
    For i = 1 to T
    for For j = 1 to all images in i folder do
    Find face, Crop and Resize
    Get R, G and B channel values in widow of –9 to +9 x and y values around
    end for
    for For each N (x, y) in emotion image and in periphery of 3 to 17 pixels do
    Calculate Average R1, G1 and B1 for widow size –2 to +2 in x and y direction
    Slide the widow
    end for
    for For each Value obtained do
    Calculate the Difference=Average R - Average R1
    Repeat for other 2 color channels
    end for
    Calculate the point having minimum difference out of values obtained.
    Obtain coordinates of the pixel whose window has minimum Difference.
    Calculate signed difference in coordinates of Neutral Image and Emotion Image for all N points
    Classification
```

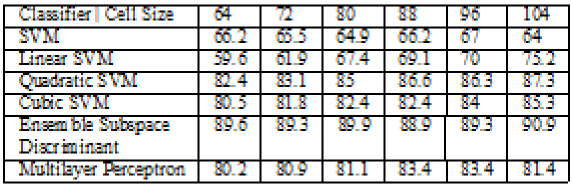
Calculation of the points having minimum difference out of values obtained was done. Thereafter coordinates of the pixel whose window had minimum difference were obtained. Also signed difference in coordinates of Neutral Image and Emotion Image for all N points was calculated in this step.

VII. CLASSIFIC AT ION AND RESULTS

The feature vector obtained in the previous step was then used for classification. The results were matched with the emotion labels provided in the database for obtaining percentage of ac- curacy. The resultant vector of Histogram gradients was subjected to classification. Carrying the experiment further this vector was subjected to principal component analysis and subsequently then subjected to classification. In the next phase resultant vectors of statistical moments of Compact LBI images were classified to obtain accuracy percentage. Further on this vectors was appended to the feature vector obtained by HOG and then classified to obtain accuracy percentage at five fold cross validation is shown in Table1 and 2 and in Figure 17.

The results in terms of obtained accuracy at 5 fold cross validation (HOG) is shown in the following Table 1 and Table 2.

Table 1. Results at various cell sizes from 64 to 104



Classifier | Cell Size	64	72	80	88	96	104
SVM	66.2	65.5	64.9	66.2	67	64
Linear SVM	59.6	61.9	67.4	69.1	70	75.2
Quadratic SVM	82.4	83.1	85	86.6	86.3	87.3
Cubic SVM	80.5	81.8	82.4	82.4	84	85.3
Ensem ble Subspace Discr im inant	89.6	89.3	89.9	88.9	89.3	90.9
Multilayer Perceptron	80.2	80.9	81.1	83.4	83.4	81.4

Table 2. Results at various cell sizes from 112 to 512

Classifier| Cell Size	112	120	128	256	512
SVM	64.7	62.8	61.9	50.8	39.9
Linear SVM	83.7	84	86	84.7	85
Quadratic SVM	88.3	85	88.3	85.7	87.3
Cubic SVM	86	84	87.6	84.4	84.7
Ensemble Subspace Discriminant	88.6	89.9	91.5	89.6	91.9
Multilayer Perceptron	85.3	85.3	87	85.3	85.3

0	0	0	0	0	0
0	1101110	0	0	01110111	0
0	0	0	0	0	0
0	0	0	0	0	0
0	00010110	0	0	11101111	0
0	0	0	0	0	0

Figure 14. Resultant image LBP shown in binary.

206	119
22	239

Figure 15. Compact LBP.

In the next phase only Quadratic SVM was used for the hybrid feature which comprised features of both HOG and Compact LBP as shown in Figure 18.

The number of points marked on the neutral images were 14 to 32 at a gap of 2 points for each subject of three databases namely

Bharat Database of Indian Faces (BDI F), The Japanese Female Facial Expression Database(J AF F E) [20] and Karolinska Directed Emotional Faces(K DEF) [21]. J AF F E has posed seven facial expressions of 10 Japanese female models and has a total of 213 images. KDEF has images of 70 amateur actors comprising of 35 males and 35 females. The total number of images in the database is 4900. The subject images have no occlusions in the form of earrings, eyeglasses, moustache, beard and also no visible make-up. It has images taken from five different angles. However only frontal images were used in the experiments.

The points were marked to be in conformity to action units of Facial Action Coding System as described by Paul Ekman. These points covered eyebrows, eyes, nose and mouth region of the images. The tracking of these points was done in the all the emotion images of those subjects.

The accuracy of emotion detection is then calculated:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

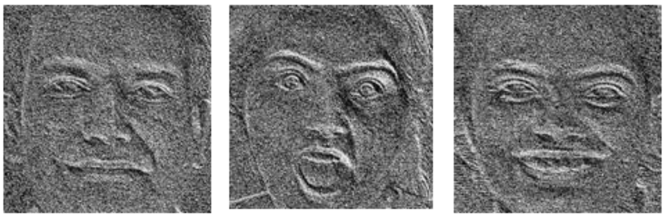


Figure 16. Sample Compact Lbp extracted from BDIF.

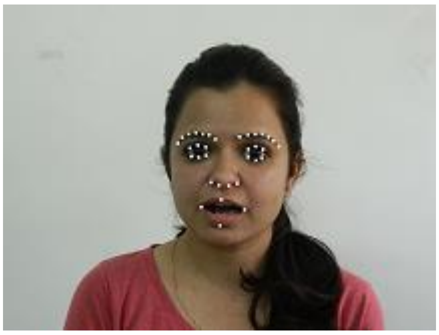


Figure 17. Marked points on sample image.

Support Vector Machines are used in machine learning as supervised learners with algorithms for data analysis for classifi- cation. Training examples are categorized into different categories based on the classification labels. SV M builds model based on the training examples as falling into different categories. After build- ing model, new examples are categorized for different classes. It is non-probabilistic classifier that represents examples as points in space. Mapping is done to categorize the examples of the different categories so that there is a clear distinction between them. After model building the test examples are mapped into the same space. Thereafter prediction is done about their class based on which side of space they come into.

Multilayer Perceptrons maps input data to different output categories based on feedforward artificial neural network. It has directed graph of nodes in multiple layers, the nodes of one layer are connected fully to nodes next layer. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. It has one or more hidden layers that have nonlinear activating nodes.

Navie Bayes provides least percentage of accuracy of the three classifiers used in the experiment. This is particularly true because it is probabilistic classifier with the assumption of independency among features. But in the feature vector used in the experiments, there is a correlation among the various features associated with the points marked. SVM provides high degree of accuracy because of its ability to construct hyperplane between the classes. Multilayer Perceptron also provide high accuracy because of limited number of input nodes, maximum of 34, mapped to total seven output nodes. The results show that SVM and Multilayer perceptron's are best suited for this approach.

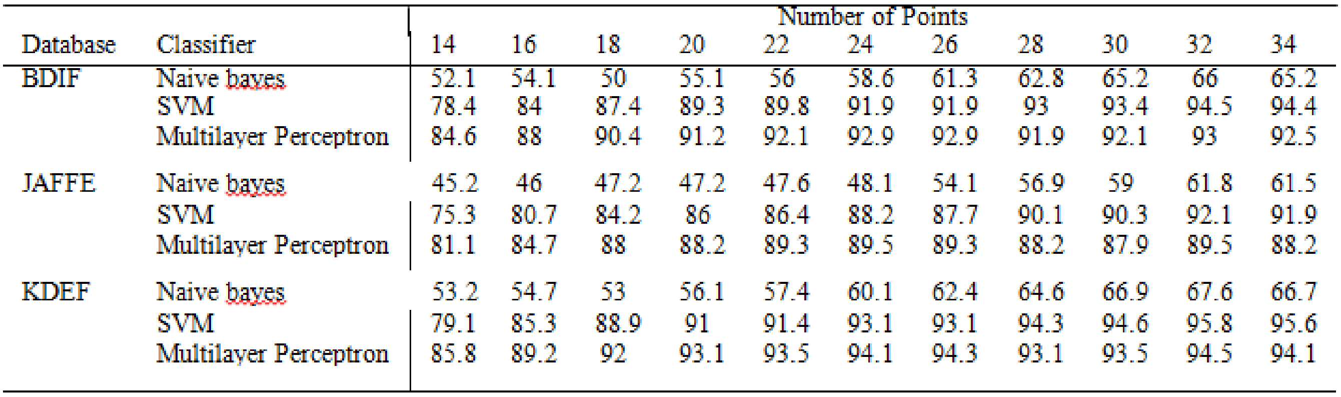
Where T P is True positive, T N is true negative, F P is False Positive and F N is False Negative.

The classifiers used for classification are NaÃrve Bayes, Sup- port Vector Machine (SVM) and Neural Network.

Naive Bayes are probabilistic classifiers based on applying Bayes' theorem with assumption of independence between features. It is used for classification of examples as falling into different categories.

Results are shown in Table 3 corresponding to the number of points considered for tracking.Across the databases, the accuracy percentage of KDEF is highest followed by BDIF and JAFFE is at last. KDEF has the highest accuracy as the images of the database.

Table 3. Accuracy of emotion detection in percentge(%)



		Number of Points										
Database	Classifier	14	16	18	20	22	24	26	28	30	32	34
BDIF	Naive bayes	52.1	54.1	50	55.1	56	58.6	61.3	62.8	65.2	66	65.2
	SVM	78.4	84	87.4	89.3	89.8	91.9	91.9	93	93.4	94.5	94.4
	Multilayer Perceptron	84.6	88	90.4	91.2	92.1	92.9	92.9	91.9	92.1	93	92.5
JAFFE	Naive bayes	45.2	46	47.2	47.2	47.6	48.1	54.1	56.9	59	61.8	61.5
	SVM	75.3	80.7	84.2	86	86.4	88.2	87.7	90.1	90.3	92.1	91.9
	Multilayer Perceptron	81.1	84.7	88	88.2	89.3	89.5	89.3	88.2	87.9	89.5	88.2
KDEF	Naive bayes	53.2	54.7	53	56.1	57.4	60.1	62.4	64.6	66.9	67.6	66.7
	SVM	79.1	85.3	88.9	91	91.4	93.1	93.1	94.3	94.6	95.8	95.6
	Multilayer Perceptron	85.8	89.2	92	93.1	93.5	94.1	94.3	93.1	93.5	94.5	94.1

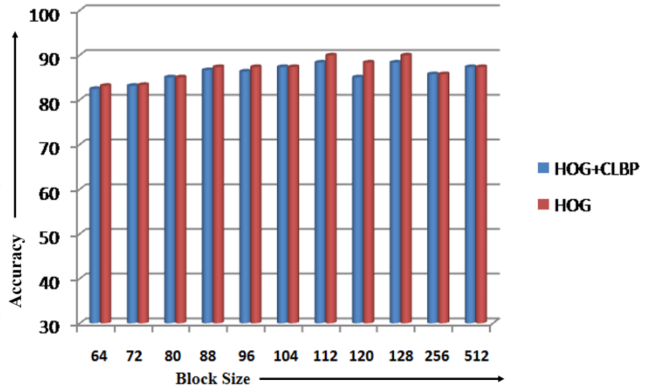


Figure 18. Comparison of result achieve using HOG and HOG+CLBP.

Gathering information from facial expressions may be hindered by the presence of glasses, mustache, beard etc. These act as noisy features in implementation of automatic emotion detection algorithms. However they are a regular feature in the real life situations. Therefore these features were deliberately included in the formation of database. In order to show its effectiveness and applicability quantitative analysis of accuracy is done.The results show that it is robust to gender, occlusions and ethnicity.

Are free from occlusions in the form of earrings, eyeglasses, moustache, beard also no make-up. BDIF stands second because of the presence of occulisons. JAFFE stands last because it is composed of low resolution images. Therefore the calculation of distance vector within the periphery of original point location that has minimum intensity difference is not always that accurate. This results in feature vector that is less accurate towards classification of emotions.

VIII. CONCLUSION

Automatic recognition of emotions from the facial expressions continues to be an important aspect in the field of evolution of new age computing systems. The various features that increase the complexity of emotion recognition systems include ethnicity, gender, pose, occlusion, beard, moustache etc.

The type of database used for learning by systems is of crucial importance. Many databases exist for this purpose but none of them is for posed Indian faces. We bridge this gap by providing Bharat Database which contains facial images of Indian people. The wide variety of subjects and their emotion labeling may help researchers in developing robust algorithms for futuristic artificially intelligent systems. Several evaluations of accuracy were done to behave as a baseline by researchers to develop more.

High degree of accuracy is obtained for all the point tracking for multiple classifications. Extensive experiments were done to show that a window size ranging from 14 to 34 points is sufficient to categorize the emotions. This may be attributed to points conveying the information inline with action units described in physiological studies.

Exhaustive experiments may be conducted involving different models for hybrid vector creation over various other datasets for increasing robustness and accuracy in categorization of emotions.

IX. ACKNOWLEDGMENT

We are grateful to all participants who have provided their time and also helped by ascertaining the advice provided for creation of this database which was a herculean task. Also photography expert and professional artist deserve a great thanks for making this project successful. Last but special thanks to the annotators who provided their time to validate the emotion labels associated with the database.

X. REFERENCES

[1]. Ekman, Paul. "An argument for basic emotions." Cognition & emotion 6, no. 3-4 (1992): 169-200.

[2]. Lucey, Patrick, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression." In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, pp. 94-101. IEEE, 2010.

[3]. Lyons, Michael J., Shigeru Akamatsu, Miyuki Kamachi, Jiro Gyoba, and Julien Budynek. "The Japanese female facial expression (JAFFE) database." In Proceedings of third international conference on automatic face and gesture recognition, pp. 14-16. 1998.

[4]. Yin, Lijun, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato. "A 3D facial expression database for facial behavior research." In Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on, pp. 211-216. IEEE, 2006.

[5]. Valstar, Michel, and Maja Pantic. "Induced disgust, happiness and surprise: an addition to the mmi facial expression database." In Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, p. 65. 2010.

[6]. Sneddon, Ian, Margaret McRorie, Gary McKeown, and Jennifer Hanratty. "The belfast induced natural emotion database." IEEE Transactions on Affective Computing 3, no. 1 (2012): 32-41.

[7]. Happy, S. L., Priyadarshi Patnaik, Aurobinda Routray, and Rajlakshmi Guha. "The Indian Spontaneous Expression Database for Emotion Recog- nition." IEEE Transactions on Affective Computing 8, no. 1 (2017): 131-142.

[8]. Mavadati, S. Mohammad, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn. "Disfa: A spontaneous facial action intensity database." IEEE Transactions on Affective Computing 4, no. 2 (2013): 151-160.

[9]. Viola, Paul, and Michael Jones. "Rapid object detection using a boosted cascade of simple features." In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol. 1, pp. I-I. IEEE, 2001.

[10]. Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1, pp. 886-893. IEEE, 2005.

[11]. Jolliffe, Ian. Principal component analysis.John Wiley & Sons, Ltd, 2002. 12Ojala, Timo, Matti Pietikainen, and Topi Maenpaa. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns." IEEE Transactions on pattern analysis and machine intelligence 24, no. 7 (2002): 971-987.

[12]. Zhao, Guoying, and Matti Pietikainen. "Dynamic texture recognition using local binary patterns with an application to facial expressions." IEEE transactions on pattern analysis and machine intelligence 29, no. 6 (2007): 915-928.

[13]. Rudovic, Ognjen, Maja Pantic, and Ioannis Patras. "Coupled Gaussian processes for pose-invariant facial expression recognition." IEEE trans- actions on pattern analysis and machine intelligence 35, no. 6 (2013): 1357-1369.

[14]. Tong, Yan, Jixu Chen, and Qiang Ji. "A unified probabilistic framework for spontaneous facial action modeling and understanding." IEEE trans- actions on pattern analysis and machine intelligence 32, no. 2 (2010): 258-273.

[15]. Ojansivu, Ville, and Janne HeikkilAd'. "Blur insensitive texture classi- fication using local phase quantization." In International conference on image and signal processing, pp. 236-243. Springer Berlin Heidelberg, 2008.

[16]. Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1, pp. 886-893. IEEE, 2005.

[17]. Huang, Di, Caifeng Shan, Mohsen Ardabilian, Yunhong Wang, and Liming Chen. "Local binary patterns and its application to facial image analysis: a survey." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 41, no. 6 (2011): 765-781.

[18]. Ojala, Timo, Matti Pietikainen, and Topi Maenpaa. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns." IEEE Transactions on pattern analysis and machine intelligence 24, no. 7 (2002): 971-987.

[19]. Valstar, Michel F., Bihan Jiang, Marc Mehu, Maja Pantic, and Klaus Scherer. "The first facial expression recognition and analysis challenge." In Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pp. 921-926. IEEE, 2011.

[20]. Jiang, Bihan, Michel F. Valstar, and Maja Pantic. "Action unit detection using sparse appearance descriptors in space-time video volumes." In Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pp. 314-321. IEEE, 2011.

[21]. Ahonen, Timo, Abdenour Hadid, and Matti Pietikainen. "Face descrip- tion with local binary patterns: Application to face recognition." IEEE transactions on pattern analysis and machine intelligence 28, no. 12 (2006): 2037-2041.

[22]. Zhao, Guoying, and Matti Pietikainen. "Dynamic texture recognition using local binary patterns with an application to facial expressions." IEEE transactions on pattern analysis and machine intelligence 29, no. 6 (2007): 915-928.

[23]. Taheri, Sima, Qiang Qiu, and Rama Chellappa. "Structure-preserving sparse decomposition for facial expression analysis." IEEE Transactions on Image Processing 23, no. 8 (2014): 3590-3603.

[24]. Shan, Caifeng, Shaogang Gong, and Peter W. McOwan. "Facial expres-sion recognition based on local binary patterns: A comprehensive study." Image and Vision Computing 27, no. 6 (2009): 803-816.

[25]. Lv, Hai-Rong, Zhong-Lin Lin, Wen-Jun Yin, and Jin Dong. "Emotion recognition based on pressure sensor keyboards." In Multimedia and Expo, 2008 IEEE International Conference on, pp. 1089-1092. IEEE, 2008.

[26]. Soleymani, Mohammad, Sadjad Asghari-Esfeden, Yun Fu, and Maja Pantic. "Analysis of EEG signals and facial expressions for continuous emotion detection." IEEE Transactions on Affective Computing 7, no. 1 (2016): 17-28.

[27]. Matlovic, Tomas, Peter Gaspar, Robert Moro, Jakub Simko, and Maria Bielikova. "Emotions detection using facial expressions recognition and EEG." In Semantic and Social Media Adaptation and Personalization (SMAP), 2016 11th International Workshop on, pp. 18-23. IEEE, 2016.

[28]. Kim, Kyung Hwan, Seok Won Bang, and Sang Ryong Kim. "Emotion recognition system using short-term monitoring of physiological signals." Medical and biological engineering and computing 42, no. 3 (2004): 419- 427.

[29]. Boiten, Frans A., Nico H. Frijda, and Cornelis JE Wientjes. "Emotions and respiratory patterns: review and critical analysis."

International Journal of Psychophysiology 17, no. 2 (1994): 103-128.

[30]. Glowinski, Donald, Antonio Camurri, Gualtiero Volpe, Nele Dael, and Klaus Scherer. "Technique for automatic emotion recognition by body ges- ture analysis." In Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on, pp. 1-6. IEEE, 2008.

[31]. Castellano, Ginevra, Santiago D. Villalba, and Antonio Camurri. "Recog- nising human emotions from body movement and gesture dynamics." In International Conference on Affective Computing and Intelligent Interaction, pp. 71-82. Springer, Berlin, Heidelberg, 2007.

[32]. Tian, Yingli, Takeo Kanade, and Jeffrey F. Cohn. "Facial expression recognition." In Handbook of face recognition, pp. 487-519. Springer London, 2011.

[33]. Bowyer, Kevin W., Kyong Chang, and Patrick Flynn. "A survey of ap- proaches and challenges in 3D and multi-modal 3D+ 2D face recognition." Computer vision and image understanding 101, no. 1 (2006): 1-15.

[34]. Jeni, LAa˛szlAs¸ A., Jeffrey F. Cohn, and Takeo Kanade. "Dense 3D face alignment from 2D video for real-time use." Image and Vision Computing 58 (2017): 13-24.

[35]. Kamarol, Siti Khairuni Amalina, Mohamed Hisham Jaward, Heikki KAd'lviAd'inen, Jussi Parkkinen, and Rajendran Parthiban. "Joint facial expression recognition and intensity estimation based on weighted votes of image sequences." Pattern Recognition Letters92 (2017): 25-32.

[36]. Valstar, Michel F., Enrique SAa˛nchez-Lozano, Jeffrey F. Cohn, Laszlo A. Jeni, Jeffrey M. Girard, Zheng Zhang, Lijun Yin, and Maja Pantic. " FERA 2017-Addressing Head Pose in the Third Facial Expression Recognition and Analysis Challenge". arXiv preprint arXiv:1702.0417 (2017).

[37]. Battocchi, Alberto, Fabio Pianesi, and Dina Goren-Bar. "A first eval- uation study of a database of kinetic facial expressions (dafex)." In Proceedings of the 7th international conference on Multimodal interfaces, pp. 214-221. ACM, 2005.

[38]. Baron-Cohen, Simon. "The Eyes as Window to the Mind." The American journal of psychiatry 174, no. 1 (2017): 1-2.

[39]. Shen, Xiaolu, Xuetao Feng, Jungbae Kim, Hui Zhang, Youngkyoo Hwang, and Ji-yeun Kim. "An extensible framework for facial motion tracking." In Consumer Electronics (ICCE), 2013 IEEE International Conference on, pp. 288-291. IEEE, 2013.

[40]. Viola, Paul, and Michael Jones. "Rapid object detection using a boosted cascade of simple features." In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol. 1, pp. I-I. IEEE, 2001.

[41]. Bashyal, Shishir, and Ganesh K. Venayagamoorthy. "Recognition of facial expressions using Gabor wavelets and learning vector quantization." Engineering Applications of Artificial Intelligence 21, no. 7 (2008): 1056-1064.

[42]. Chen, Ying, Chunjian Hua, and Ruilin Bai. "Sequentially adaptive active appearance model with regression-based online reference appearance template." Journal of Visual Communication and Image Representation 35 (2016): 198-208.

[43]. Wang, Jun, Lijun Yin, Xiaozhou Wei, and Yi Sun. "3D facial expression recognition based on primitive surface feature distribution." In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 2, pp. 1399-1406. IEEE, 2006.

[44]. Lorincz, Andras, Laszlo Jeni, Zoltan Szabo, Jeffrey Cohn, and Takeo Kanade. "Emotional expression classification using time-series kernels." In Proceedings of the IEEE Conference on computer vision and pattern recognition workshops, pp. 889-895. 2013.

[45]. Sariyanidi, Evangelos, Hatice Gunes, and Andrea Cavallaro. "Automatic analysis of facial affect: A survey of registration, representation, and recognition." IEEE transactions on pattern analysis and machine intelli- gence 37, no. 6 (2015): 1113-1133.

[46]. Liu, Jingjing, Bo Liu, Shaoting Zhang, Fei Yang, Peng Yang, Dimitris N. Metaxas, and Carol Neidle. "Non-manual grammatical marker recognition based on multi-scale, spatio-temporal analysis of head pose and facial expressions." Image and Vision Computing 32, no. 10 (2014): 671-681.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

A Sentiment Analysis of Food Review using Logistic Regression

Mayur Wankhade[1], A Chandra Sekhara Rao[1], Suresh Dara[2], Baijnath Kaushik[3]

[1]Department of Computer Science and Engineering Indian Institute of Technology Indian Institute of Technology (ISM), Dhanbad, Jharkhand, India

[2]Department of Computer Science and Engineering Indian Institute of Technology B.V. Raju Institute of Technology, Narsapur, Telangana, India

[3]Department of Computer Science and Engineering Indian Institute of Technology Shri Mata Vaishno Devi University, Katra, Jammu and Kashmir, India

ABSTRACT

Sentiment analysis of review is most popular task in text classification. Online or Offline user opinion about the product is great platform to collecting the large volume of data for sentiment analysis.so the overall user reviews about product are the task for sentimental analysis .it can categories into two parts positive and negative. We can train the data model and find the sentiment hidden in the review .provide the review either positive or negative by analyzing the performance with respective parameter accuracy, precision, recall and f-measure calculating for each of the algorithm for comparison. Text classification by using machine learning technique several models Perceptron, Naïve Bayes and Logistic regression used to compare the model. Among the different classification algorithm using logistic regression method accuracy level improved .Sentiment analysis is performs by using two different text feature selection method and three classification method . Problem statement here is analyzing the sentiment analysis over large dataset.

Keywords : Text Preprocessing, Text Classification, Sentiment Analysis

I. INTRODUCTION

Opinion analysis is natural language processing task which important to analyze the sentiment and feeling about the product. Base on polarity classification in sentence .classification of text or sentence consist of three type positive ,negative ,neutral to develop the model consist of three main approach lexicon based ,rule based approach method ,machine learning algorithm .

Dataset Analysis

The Food Reviews of user is huge dataset which comprises of around 568454 surveys reviver sustenance items composed by commentators in the vicinity of 1999 and 2012. Each survey has the accompanying 10 parameter Id, Product Id, User Id, Profile Name, Helpfulness Numerator, Helpfulness Denominator, Score,Time Summary ,Text .So among the parameter contain score and text are the ones having some more prescient esteem. Likewise 'content' is somewhat excess as synopsis is adequate to extricate the conclusion covered up in the audit. Score has an incentive in the vicinity of 1 to 5. So with the end goal of the all audits having score 3 are neutral review, below 3 score as negative review and above 3 are positive review. For the given dataset there is large number of positive

Review 77% and negative review 23 % in given dataset. This is imperative snippet of data as it of now empowers one to choose that a stratified technique should be utilized for part information for assessment.

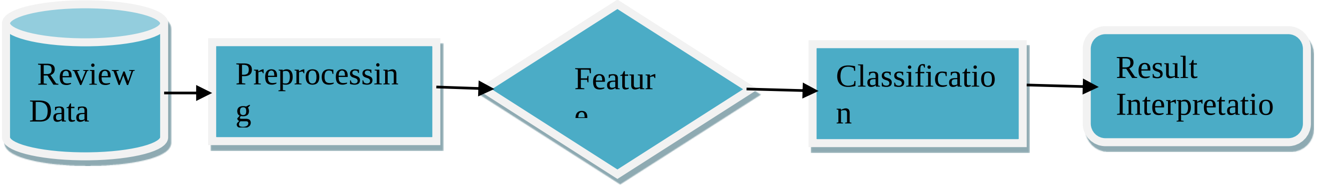


Figure 1. stepwise procedure for sentimental analysis

II. PREPROCESSING

In Pre-processing technique is more important step for text classification. We can't put sequence of symbols into these algorithms as it should have numerical values and not sequence of symbols with variable length .So there are various ways by which we can get numerical values from this sequence of symbols.

Stop word: the Stop word are common word which present in review so to remove the stop word because they don't be make any sense in predication .

Tokenizing: In these we mainly use integer token id for each string.
Counting: In this we mainly count number of tokens
Normalizing: In this we assign weight to the token that occur frequently.

The process of tokenization, counting and normalization called as bags of words

The process of converting the text document into numerical value called as a vectorization.

For comparison on text processing methods on sentiment analysis [1-2] which analysis. For the given dataset is basically a significant huge numbers to run large calculation. In this way it ends up noticeably vital to by one means or another decrease the extent of the list of capabilities. There are various ways this should be possible. The principal issue that should be handled is that a large portion of the characterization calculations expect contributions to the type of highlight vectors having numerical esteems and having settled size rather than crude content reports (surveys for this situation) which different length. which is handled utilizing BagsofWords procedure.

III. FEATURE SELECTION AND REDUCTION

Feature selection method performs the key role for sentiment analysis .for classification algorithms performance base on the feature selection method.so selecting the appropriate feature is most important task. For feature selection and reduction task we used two methods most frequent feature and principal component analysis (PCA) used to reducing the size of feature set. After the preprocessing stage we can find the unique word present in the data then we can perform the training and testing. The frequency of each token is treated as the token. The vectors of all the frequencies are taken as the sample. This is the most important preprocessing venture for feeling order. Arrangement calculations are keeping running on subset of the highlights, so choosing the correct highlights winds up noticeably critical.

Principal Component Analysis (PCA)

PCA is used for reducing the feature size .PCA statical technique which uses the orthogonal transformation to convert them into correlated variable into the linearly uncorrelated variable .the an arrangement factors to reducing the dimension of n-dimension to some small dimensions. By takin the case in which focuses are disseminated into 2-dimsnssion space containing greatest change as per x_pivot. Which is fitting the focuses on 1 -dimension on pressing every which focuses on the x The x hub is the main primary part and the information had greatest fluctuation occurred on it. Some comparative should be possible for highest measurements as well. With the end goal of the task, the list of capabilities is diminished to 200 segments utilizing Truncated Singular value decomposition (SVD) which contain a variants of principal component analysis they performs on the sparse matrices.

Most Frequently Words

Second approach for reducing the number of features most frequently word occurring use as subset in our available dataset. Here lessen the quantity of highlights which utilize subset of repeating words happening in data in the list of capabilities.

Discover the recurrence of all words in the preparation information and select the most well-known 5k words list feature set. In rationale for the approach is that all reviver use some common basic words that characterize the assessment of the surveys dataset these must happen as often as possible. 5k words are still a considerable amount of highlights yet it diminishes the list of capabilities to around 1/fifth of as per given so that it is beneficial . The recurrence circulation for the dataset looks something like underneath.

The most important 5000 words are vectorized utilizing Tf-idf transformer. Utilizing a similar transformer, the prepare and the test information are likewise vectorized . This basically implies just those expressions of the preparation and testing information, which are among the most regular 5000 words, will have numerical incentive in the created frameworks. These grids are then utilized for preparing and assessing the models.

There is critical change in every one of the models. Following is an outcome outline.

One vital thing to note about Perceptron is that it just joins when information is directly divisible. Since the quantities of highlights are so expansive one can't tell if Perceptron will focalize on this dataset. In this way confining the greatest emphases for it is essential. Following is an examination of review for negative examples.

In conclusion the models are prepared without doing any element diminishment/determination step. Choice Tree Classifier runs pretty wastefully for datasets having extensive number of highlights, so preparing the Decision Tree Classifier is maintained a strategic distance from.

Since the whole list of capabilities is being utilized, the arrangement of words (relative request) can be used to do a superior expectation. For instance : a few words when utilized together have an alternate importance contrasted with their significance when considered alone like "not great" or "not terrible".

The models are prepared for 3 techniques called Unigram, Bigram and Trigram it is perceptible that words, for instance, magnificent, incredible, best, love, tasty et cetera happen most a significant part of the time in the dataset and these are the words that conventionally have most noteworthy insightful motivating force for suspicion examination. This similarly exhibits the dataset isn't deteriorate or unimportant to the issue explanation.

IV. CLASSIFICATION METHODS

There are three different methods used for training set and testing set. Here the dataset contain large number of training set .This methods are prepared for preparation set and assessed related to the test set.

The quantity of tests in the preparation set is colossal obviously it won't be conceivable to run some wasteful grouping calculations like The 3 classifiers utilized are Naïve Bayes Classifier, Logistic Regression, and Perceptron

The models are prepared on the info grid produced previously. Test information is additionally changed in a comparative mold to get a test network. Following are the outcomes:

Note that in spite of the fact that the precision of Perceptron and Bernoulli does not look that awful but rather in the event that one considers that the dataset is skewed and contains 78% positive surveys, anticipating the larger part class will dependably give no less than 78% exactness. So contrasted with that perceptron and BernoulliNB doesn't work that well for this situation

To skewed information recall is those best measure to execution of a model. The execution about constantly on three models will be compared beneath.

Similarly as guaranteed prior Perceptron What's more Naïve bayes are foreseeing sure to Just about every last one of elements, Consequently the review Furthermore precision values would pretty low to negative tests precision/recall.

Naïve Bayes Method

Naïve Bayes is probabilistic classifier method used when size of training set has small. This method based on mathematical bayes theorem .there are two class naïve bayes variants for text .multinomial naïve bayes and benerolli naïve bayes . multinomial naïve bayes method data follows a multinomial distribution and each feature value is count . benerolli naïve bayes data follows a multivariate distribution and each feature is binary .

The conditional probability of event X occurs given the evidence Y is determined by Bayes rule by the

$$P(X/Y) = \frac{P(X)\,P(\frac{Y}{X})}{P(Y)}$$

Finding sentiment of review by using an naïve bayes as follows

P(Sentiment/Sentence) = P(Sentiment)P(Sentence/Sentiment)/P(Sentence)

P(sentence/sentiment) is calculated as the product of P (token /sentiment) ,by using formula.

Count(Thistokeninclass)+1/Count(Alltokensinclass)+ Count(Alltokens)

Here 1 and count of all tokens is called tokens Laplace smoothing or additive smoothing which used for to smooth the categorical data.

Logistic regression

Second algorithm for classification called multinomial logistic regression, sometimes referred to within language processing as maximum MaxEnt entropy modeling, MaxEnt for short. Logistic regression belongs to the family of classifiers known as the exponential or log-linear classifiers. Like naive Bayes, it log-linear classifier works by extracting some set of weighted features from the input, taking logs, and combining them linearly (meaning that each feature is multiplied by a weight and then added up). Technically, logistic regression refers to a classifier that classifies an observation into one of two classes, and multinomial logistic regression is used when classifying into more than two classes, although informally and in this chapter we sometimes use the shorthand logistic regression even when we are talking about multiple classes. The most important difference between naive Bayes and logistic regression is that logistic regression is a discriminative classifier while naive Bayes is a generative classifier. To see what this means, recall that the job of a probabilistic classifier is to choose which output label y to assign an input x, choosing the y that maximizes P(y|x). In the naive Bayes classifier, we used Bayes rule to estimate this best y indirectly from the likelihood P(x|y) (and the prior P(y).

Y* = argmaxP(Y/X)= argmaxP(X/Y)P(Y)

V. PERFORMANCE EVOLUTION

Performance evaluation used for checking the Classification result as Precision ,recall and F-measure.

True positive (TP)is correctly predicted positive values which mean that value of actual class is yes and predicated class is also yes.

True Negative (TN) is correctly predicted negative values which means that the actual class is no and value of predicated also no.

False Positive (FP) is actual class is no and predicated class is no.

False Negative (FN) is actual class is yes and predicated class is no.

Precision is the number of true positive review out of total number positively assigned review

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall is the number of true positive out of the actual positive review and it is given by

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F-measure used to calculated weighted method of precision and recall and it is calculated by

$$\text{F} - \text{measure} = \frac{2 * \text{ precision} * \text{recall}}{\text{precision } + \text{ recall}}$$

Accuracy is most important performance measure it is measure it is ratio of predicated observation to the total number of observations. We have high accuracy indicating our model is best.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN })}{(\text{TP} + \text{FP} + \text{FN} + \text{TN})}$$

Unigram

Unigram is the ordinary case, when each word is considered as a different element. The whole list of capabilities is vectorized and the model is prepared on the produced report .

Table 1

Unigram	precision	recall	f1-score
Negative	0.73	0.61	0.63
Positive	0.90	0.92	0.91
Average / total	0.86	0.86	0.86

Of course correctness's got are superior to subsequent to applying highlight diminishment or determination yet the quantity of calculations done is likewise way higher. Following are the correctness's:

Every one of the classifiers perform entirely with great exactness in review esteems for negative examples. Following demonstrates a visual correlation of review for negative examples:

Table 2

UnigramBernoulliNB	**precision**	**recall**	**f1-score**
Negative	0.73	0.67	0.70
Positive	0.91	0.93	0.92
Average / total	0.87	0.87	0.87

Table 3

Unigram Logistic	precision	recall	f1-score
Negative	0.80	0.69	0.74
Positive	0.92	0.95	0.93
Average / total	0.89	0.89	0.89

Table 4

Unigram Perceptron	precision	recall	f1-score
Negative	0.70	0.64	0.67
Positive	0.90	0.92	0.91
Average / total	0.86	0.86	0.86

Bigram

Succession in contiguous strings considered highlights separated with Unigrams. Words with "not great", "not awful", "truly terrible" and so on will likewise have a prescient esteem which wasn't there when utilizing Unigrams. The whole list of capabilities splited into vectors and modelled is prepared with created lattice.

The correctness's enhanced much more. Its calculations ran utilized run with scanty information giving arrangement in information which created during splitting into vectors. Below outcomes:

There is a change on the review of negative occurrences which may increases that numerous commentators would have utilized two word phrases like "not great" or "not awesome" to infer a negative survey. Following is the visual portrayal of the negative examples precision:

Table 5

BigramBernoulliNB	precision	recall	f1-score
Negative	0.79	0.67	0.72
Positive	0.91	0.95	0.93
Average / total	0.88	0.89	0.88

Table 6

Bigram Logistic	precision	recall	f1-score
Negative	0.86	0.80	0.83
Positive	0.95	0.96	0.95
Average / total	0.93	0.93	0.93

Table 7

Bigram Perceptron	precision	recall	f1-score
Negative	0.78	0.78	0.78
Positive	0.94	0.94	0.94
Average / total	0.90	0.90	0.90

Trigram

For successions in three neighboring characters are taken at different component separated against Bigrams and Trigrams.

This whole list of capabilities split into vectors and modelling is prepared in produced grid.

Trigrams give the best outcomes.

Calculated failure provides exactness with 93.4 % and perceptron precision is larger. Its accuracy esteems with specimens is larger with at any other time. Since calculated relapse performs best in every one of the three cases, how about we do somewhat more examination with the assistance from disarray framework. A disarray network puts the correct marks compared with anticipated names. This pictures decent approach in telling the arrangement account.

From the main framework it is obvious that countless were anticipated to be sure and their real mark was additionally positive. Through not very many negative examples which were anticipated negative were likewise really negative. Yet, this lattice isn't demonstrative of the execution on the grounds that in testing information the negative examples were less, so it is relied upon to see the anticipated name versus genuine name some portion in grid of names for softly shaded. For envisioning execution, smart thing is taking a gander in standardized perplexity grid. This standardized disarray framework speaks to the proportion of anticipated names and genuine names. Presently one can see that calculated relapse anticipated negative specimens precisely as well.

Table 8

Trigram BernoulliNB	precision	recall	f1-score
Negative	0.81	0.53	0.63
Positive	0.88	0.97	0.92
Average / total	0.87	0.87	0.86

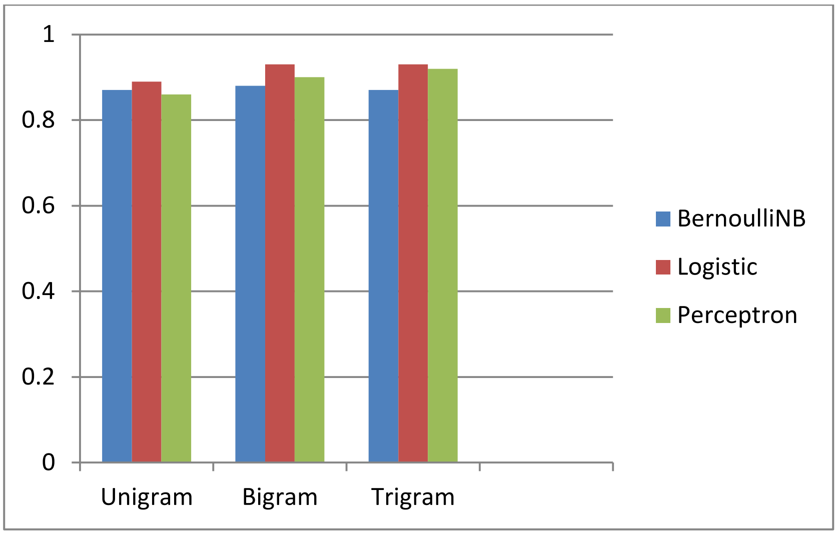
Table 9

Trigram Logistic	precision	recall	f1-score
Negative	0.87	0.82	0.84
Positive	0.95	0.96	0.96
Average / total	0.93	0.93	0.93

Table 10

Trigram Perceptron	precision	recall	f1-score
Negative	0.83	0.80	0.81
Positive	0.94	0.95	0.95
Average / total	0.92	0.92	0.92

Performance analysis of unigram ,bigram and trigram with help of chart shown in graph



Ghrap 1

K-fold Cross Validation

K-fold Cross Validation is technique which improve over the holdout method. dataset containing user review are divided into the k subsets and holdout method is repeated the k times . every time A standout amongst those k subsets is utilized Concerning illustration the test set and the other k-1 subsets are assemble to structure a preparing set. Then those Normal slip crosswise over all k trials may be registered. The preference from claiming this technique is that it matters how the information gets isolated. Each information point gets to make On An test set precisely once, and gets with be On An preparation set k-1 times. The difference of the coming about assess will be diminished Similarly as k will be expanded. Those disservice for this strategy may be that the preparation algorithm need to a chance to be rerun from scratch k times, which intends it takes k times as significantly calculation will settle on an assessment. A variant for this technique is should haphazardly gap the information under a test What's more preparing set k different times. Those playing point for finishing this is that you might freely pick how vast each test set is what's more entryway large portions trials you Normal again.

VI. FUTURE WORK

It is clear that with the end goal of supposition grouping, include diminishment and determination is critical. Aside from the techniques talked about in this paper there are different ways which can be investigated to choose includes all the more keenly. One can use POS labeling component to label words in the preparation information and concentrate the imperative words in light of the labels. For conclusion order modifiers are the basic labels. One must deal with different labels too which may have some prescient esteem.

Other propelled systems, for example, utilizing Word2Vec can likewise be used. Utilizing this would discover comparable data values and basically discover connection in names. It has different courses that will utilize Word toVector to enhance the modelling.

VII. CONCLUSION

One might say that bag of-words is an entirely proficient technique on the off chance that could bargain a with little exactness. Likewise for datasets of huge size it is cautious to utilize calculations that keep running in direct instance Classification analysis for BernoulliNB, Logistic, Perceptron technique

Table 11

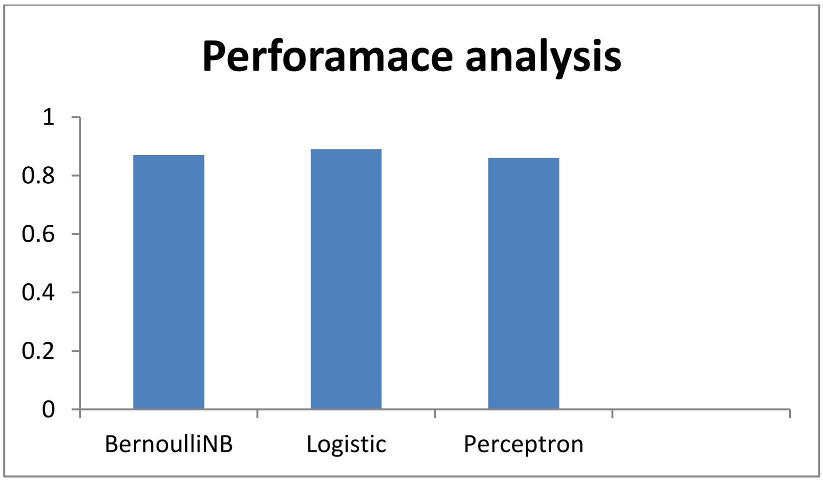
Bernoulli NB	precision	recall	f1-score
Negative	0.73	0.68	0.70
Positive	0.91	0.93	0.92
Average / total	0.87	0.87	0.87

Table 12

Logistic	precision	recall	f1-score
Negative	0.80	0.69	0.74
Positive	0.92	0.95	0.93
Average / total	0.89	0.89	0.89

Table 13

Perceptron	precision	recall	f1-score
Negative	0.70	0.64	0.67
Positive	0.90	0.92	0.91
Average / total	0.86	0.86	0.86



Ghrap 2

VIII. REFERENCES

[1]. Liu B. Sentiment analysis and opinion mining. Synth Lect Human Lang Technol 2012.

[2]. Cambria E, Schuller B, Xia Y, Havasi C. New avenues in opinion mining and sentiment analysis. IEEE Intell Syst2013;28:15–21.

[3]. Gretzel, U. and Kyung H.Y. 'Use and Impact of Online Travel Reviews', Information and Communication Technologies in Tourism(2008), pp. 35–46

[4]. E. Kouloumpis, T.Wilson, and J. Moore,"Twitter sentiment analysis: The good the bad and the

omg!" in Proc. 5th Int. AAAI Conf. Weblogs Social Media, 2011, pp. 538_541.

[5]. H. Saif, M. Fernandez, Y. He, and H. Alani,"On stopwords, _ltering and data sparsity for sentiment analysis of Twitter," in Proc. 9th Lang. Resour. Eval. Conf. (LREC), Reykjavik, Iceland, 2014, pp. 80_81.

[6]. He Yulan, Zhou Deyu. Self-training from labeled features for sentiment analysis. Inf Process Manage 2011;47:606–16.

[7]. A. Agarwal, B. Xie, and I. Vovsha,"Sentiment analysis of Twitter data," in Proc. Workshop Lang. Social Media, Assoc. Comput. Linguistics, 2011, pp. 30_38.

[8]. Tsujii K., Tsuda K. 'The attention information extraction method from a stay review using text mining' Journal of Digital Practices. 3(2012) 289–296. (Japanese)

Recent Trends in Background Subtraction Approach for Moving Object Detection

Rudrika Kalsotra, Sakshi Arora

Department of Computer Science and Engineering Shri Mata Vaishno Devi University Katra, J&K, India

ABSTRACT

Background Subtraction has attained much attentiveness in recent years due to potential growth in the field of intelligent video analytics. It is widely used technique for detecting moving objects from videos because of its flexibility and reliability. This paper presents a comprehensive survey of background subtraction approach. It highlights various applications, challenges and methods of background subtraction. The recent developments in conventional as well as in deep-learning approaches in the field of background subtraction are presented in this paper. In addition to this, future research directions in background subtraction are also outlined in the end.

Keywords : Intelligent Video Analytics; Moving Object Detection; Foreground Object; Background Subtraction; Deep-learning

I. INTRODUCTION

The problem of detecting moving objects from complex video scenes is of critical importance for the successful implementation of intelligent video analytical tasks. It is followed by object tracking, activity recognition or event analysis in high-level video analytics [1, 2]. Moving object detection is the process of extracting foreground of interests from the series of video frames based on either visual elements or motion information. There are many factors that impede the detection of complete and accurate moving objects such as dynamic video scenes, presence of shadows, video noise, motion of the camera, camouflage, challenging weather, speed and size of the object, varying light intensities and occlusion [3, 4]. Temporal differencing, Background subtraction and Optical flow are three broadly classified techniques of moving object detection from the video streams [5, 6]. The overview of moving object detection techniques are shown in Figure 1. The process of computing difference between consecutive frames based on the pixels' intensities is known as temporal differencing. Background subtraction method works by initializing a background reference frame and then each incoming frame is subtracted from the updated reference frame resulting into foreground objects. The optical flow method works by quantifying the velocities and directions of the objects. The algorithm based on the integration of different methods overcome their respective flaws and detect moving objects successfully from the video scenes. Destalem et al. [7] have presented an algorithm for moving object detection based on adaptive background subtraction and temporal differencing. The method proposed in [8] outputs complete moving object outline by integrating five frames differencing approach with background subtraction. Gang et al. [9] have improved traditional three frames differencing technique and combined it with canny edge detector followed by morphological operations to fill gaps in the foreground object. However, these algorithms do not work with complex scenarios.

Background subtraction results into accurate and complete moving object detection for the videos captured with static cameras. It does not require complex computations, has moderate time complexity and is suitable for real-time applications. It is vulnerable to environmental changes and noise interfaces but a robust background model can handle these flaws [10]. It forms a basis of almost every video analytics applications: traffic monitoring, automatic video surveillance (airport surveillance, road surveillance, and maritime surveillance), traffic flow statistics, pedestrian detection, digital composition, optical motion capture, post-event forensics, human-machine interaction and target tracking [11,12].

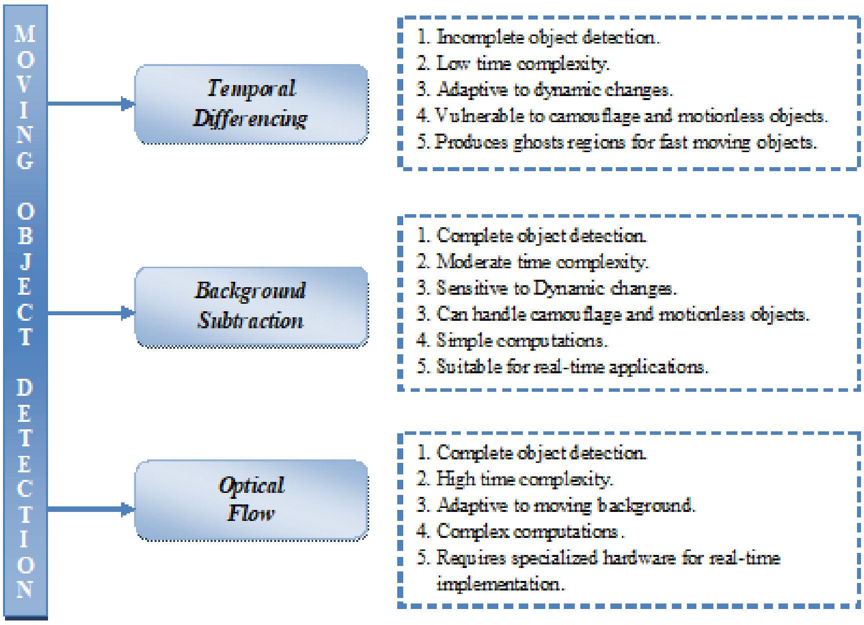

MOVING OBJECT DETECTION
Temporal Differencing
1. Incomplete object detection.
2. Low time complexity.
3. Adaptive to dynamic changes.
4. Vulnerable to camouflage and motionless objects.
5. Produces ghosts regions for fast moving objects.
Background Subtraction
1. Complete object detection.
2. Moderate time complexity.
3. Sensitive to Dynamic changes.
3. Can handle camouflage and motionless objects.
4. Simple computations.
5. Suitable for real-time applications.
Optical Flow
1. Complete object detection.
2. High time complexity.
3. Adaptive to moving background.
4. Complex computations.
5. Requires specialized hardware for real-time implementation.

Figure 1. Overview of moving object detection techniques

The selection of features plays a significant role in detection of foreground from the series of video frames. In [13], features in object detection are broadly classified into two classes: (a) human-engineering based features or hand- crafted features (color features, gradient features, pattern features and shape features) (b) learning-based features (histogram of sparse codes and deep learning features). As pointed out in [11], color features, motion features, edge features, texture features, and stereo features are widely used features and have different characteristics that can deal with complex situations. Color features are vulnerable to shadows, illumination variations and camouflage. Edges are adapted to local illumination variations. The algorithms based on texture features are robust to shadows and illumination changes [14]. The integration of different features allows us to alleviate many challenges. Conventional background subtraction algorithms are generally based on hand-crafted features and are universally adopted due to computational complexity of deep learning features [15]. The algorithms based on hand-crafted features are incapable to deal with complex video scenes [16]. Therefore, the researchers are resorting to deep-learning based background subtraction.

The rest of the paper is outlined as follows. Section II presents algorithm, different steps and challenges of background subtraction. Different background subtraction methods are explained in Section III. Recent achievements in background subtraction are discussed in section IV. Conclusions and research directions are drawn in section V.

II. BACKGROUND SUBTRACTION

The preponderance of background subtraction algorithms has been proposed by researchers for detecting moving objects from the video sequences. Figure 2 shows the background subtraction model. A general algorithm for background subtraction is shown in Figure 3. The steps of background subtraction and its challenges are explained in the following sub-sections.

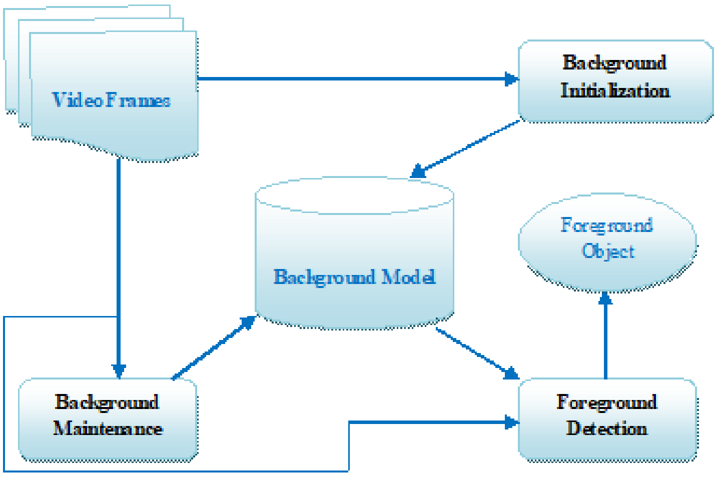
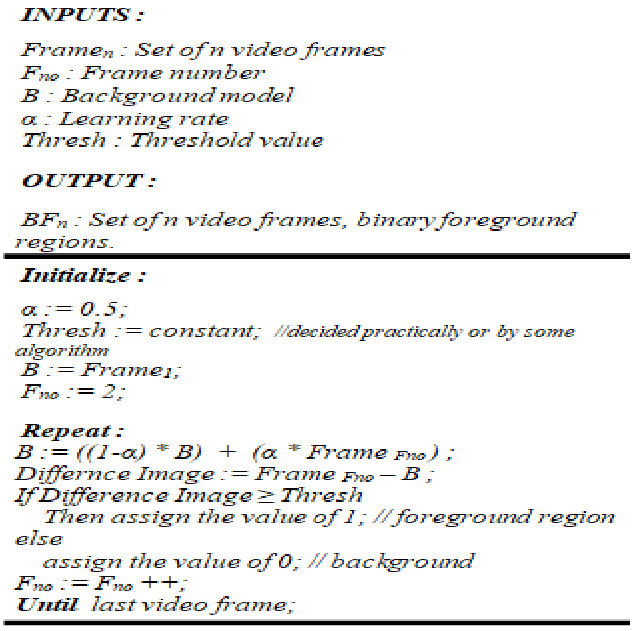


Figure 2. The background subtraction model

Steps of Background Subtraction

Based on the extensive literature study, background subtraction can be divided into three important steps: Background initialization, Foreground detection and Background maintenance [17, 18]. A graphical workflow of background subtraction is shown in Figure 4.



```
INPUTS :

$Frame_n$ : Set of n video frames
$F_{no}$ : Frame number
$B$ : Background model
$\alpha$ : Learning rate
Thresh : Threshold value

OUTPUT :

$BF_n$ : Set of n video frames, binary foreground
regions.
________________________________________________
Initialize :

$\alpha := 0.5;$
Thresh := constant;  //decided practically or by some
algorithm
$B := Frame_1;$
$F_{no} := 2;$

Repeat :
$B := ((1-\alpha) * B) + (\alpha * Frame_{F_{no}})$ ;
Differnce Image := $Frame_{F_{no}} - B$ ;
If Difference Image $\geq$ Thresh
    Then assign the value of 1 ; // foreground region
else
    assign the value of 0; // background
$F_{no} := F_{no}$ ++;
Until  last video frame;
```

Figure 3. General algorithm for background subtraction

Background Initialization: This is the first step of the background subtraction technique and the goal is to set up a background model by initializing a reference frame that is used by the other phases. There are two scenarios in a video frames while setting up a background model. First, when there is absence of foreground object in the initial video frames and second when there are one or more foreground objects present from the first video frame. Traditionally, the first frame of the video is initialized for background modeling or fixed number of video frames [11] is selected that do not have any foreground object. But it does not work with real-time applications where dynamic and complex background exists. Different initialization algorithms (neural-based, statistical, fuzzy, etc) are used depending upon complexity of the background model [17].

Foreground Detection: Each incoming frame of the input video is compared with the background model and this subtraction results into a foreground. This step is a segmentation phase that classifies pixels into either foreground pixels or background pixels. The segmentation can be done by various methods (threshold-based, region-based, clustering-based, edge-based, etc) [19]. Generally, a constant threshold is employed for segmentation. Global thresholding such as Otsu's method is employed for automatic threshold value but it is vulnerable to strong illumination gradient [20] and detects noisy regions as foreground. So there is a shift from global thresholding to adaptive thresholding [21, 31] which smoothly handles strong illumination gradient video frames. This phase outputs a binary video frame representing foreground in white and background in black or vice versa.

Background Maintenance: Background maintenance refers to the process of updation of background model in order to adapt new changes in video scene. The updation of background frame is essential to entail the latest changes into video frames. The selection of maintenance scheme and learning rate are two main challenges in this phase of background subtraction. The learning rate decides the speed of adapting new changes to the background model. The updation of background model is needed to incorporate the motionless objects into the background. Maintenance with IIR filter is commonly used for updating background model [22]. The issue with this maintenance scheme is it employs a single adaptation coefficient (learning rate) and corrupts the background model by considering all the foreground pixels in updation process. Some authors developed algorithms for selective updation by using different learning rates and solve the problem associated with single learning rate. The efficient maintenance scheme obviates erroneous detection due to illumination changes.

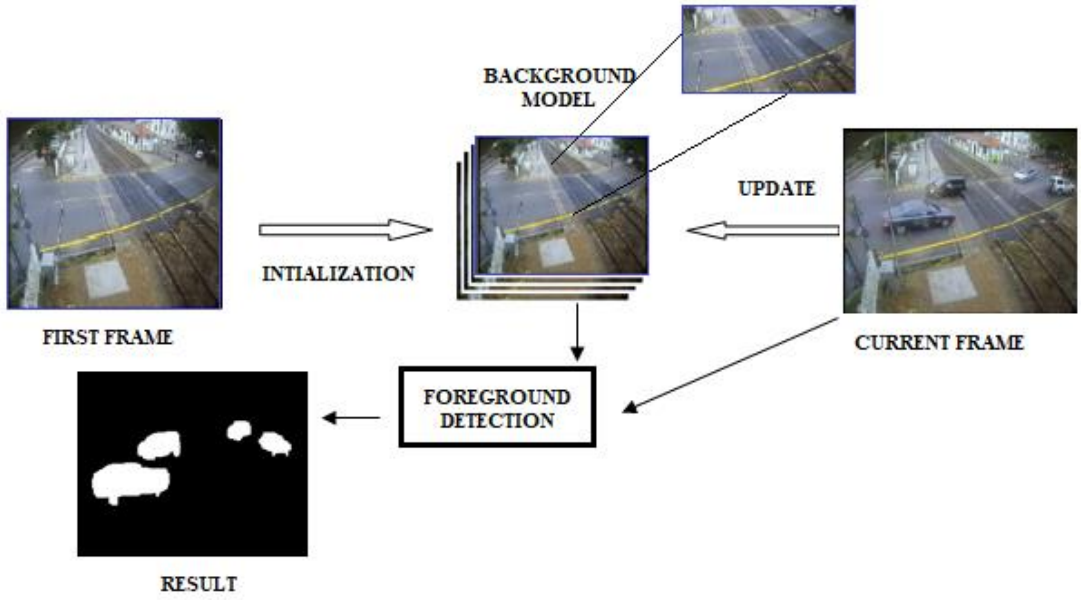


Figure 4. Graphical workflow of background subtraction

Challenges

The major challenges of background subtraction [11, 23] that lead to false detections are listed below:

Dynamic Background: Most background subtraction algorithms assume static background but it is not possible in real-life scenarios. There are some periodical or irregular movements in an outdoor as well as indoor scene. Figure 5 (a) shows video scenes containing dynamic background. The background maintenance component should handle dynamic backgrounds such as floating clouds, raindrops, dangling leaves, swing fountains, swinging of pendulum, moving escalator and swaying curtains.

Illumination Changes: The illumination changes affect the pixels in the video scene and interrupt background model. Video scenes with illumination changes are shown in Figure 5 (b). Switching on/off lights in an indoor scene causes sudden changes in illumination and produces fallacious detection. Gradual illumination changes such as the changeover from sunny days to clouds generates erroneous classification of pixels.

Camouflage: The correspondence between foreground pixels and background pixels create camouflaged regions that result into false detection of foreground objects as background [24].

Shadows: The detection of shadows is itself an active research area. Figure 5 (c) shows video scenes with shadows. The shadow casted by moving object interrupts the process of object detection. The presence of shadow has many consequences [25] such as distorted objects, merging of objects, specious foreground and overlapping shadows.

Partial or Full Occlusion: The occlusion complicates the computation of background model. There are many instances of occlusion in real-life such as moving car is occluded by sign boards, moving person may hide behind tree or pole and some regions of moving object may not be visible due to any fixed infrastructure.

Video Noise: Sensors and compressed videos may add noise to the video signals that degrade the quality of video frames and shows false detections.

Camera Jitter: Videos captured with unstable cameras result into jitter and may disrupt the motion of the moving object.

Intermittent Object Motion: Background subtraction algorithm requires effective background maintenance component to handle irregular movements of objects over time. Video scenes with intermittent object motion are shown in Figure 5 (d). The foreground such as abandoned objects or cars in parking area that become motionless for a short period of time are incorporated into the background but it must be detected again as foreground.

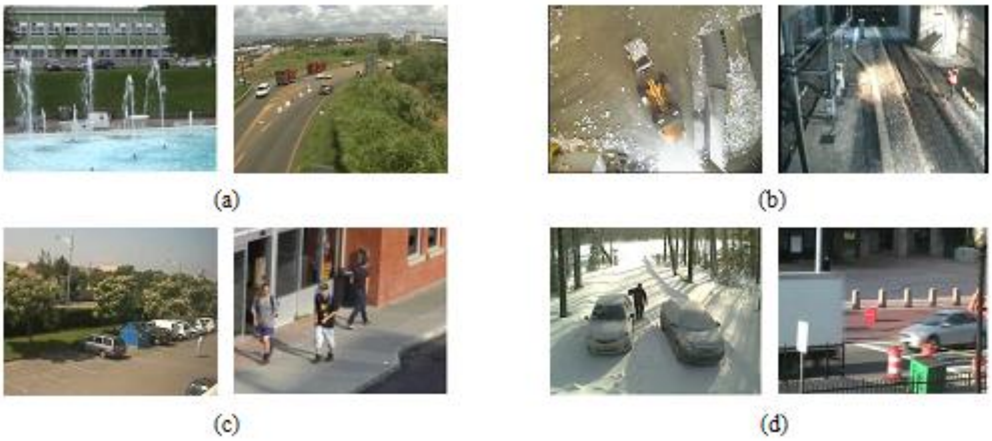


Figure 5. Video scenes: (a) Dynamic Background (b) Illumination changes (c) Shadows and (d) Intermittent object motion. These video scenes are taken from standard datasets [37, 38]

III. BACKGROUND SUBTRACTION METHODS

Background subtraction methods have achieved remarkable success in certain cases. The surveys presented in the literature categorized the background subtraction algorithms into various models [17, 23, and 26]:

Basic methods: These methods employ an average, a weighted mean, an adaptive median, pixel intensity, or a histogram for initialization and maintenance of background model. The classification of pixels as foreground or background is usually done by thresholding [21].

Statistical-based methods: Statistical methods are broadly classified into three categories [26]: gaussian methods (single or multiple), subspace learning methods and support vector methods. The advanced statistical methods use color, edge or texture features and some methods fuse different features such as color and texture [27] for foreground detection in background subtraction process. These methods are robust to dynamic backgrounds and low illumination changes.

Neural-based methods: The weights of the networks are trained to model background and learn to stratify pixels into foreground class or background class. Self organizing neural network, regression neural networks, competitive neural network and multivalued neural networks come under this category. These methods are more efficient because of learning and adaptivity of neural networks [28].

Fuzzy-based methods: As mentioned in [17], these methods are based on fuzzy concepts and introduce them in background modeling, maintenance and foreground detection. Fuzzy-based methods can deal with dynamic backgrounds and illumination variations.

Cluster-based methods: Background modeling is based on clustering where each incoming pixel is matched against clusters and decides whether the pixel belongs to background or not. Codebooks, K-means, genetic K-means methods follow clustering approach. These methods are robust to video noise and dynamic backgrounds.

Deep-learning methods: These methods are broadly classified into two classes [15]: supervised models (e.g., Convolutional Neural Networks (CNNs), Deep Neural Networks (DNNs), and Recurrent Neural Networks (RNNs)) and unsupervised models (e.g., Deep Boltzmann Machines (DBMs), Deep Belief Networks (DBNs), and Auto-encoders).

Other methods: The methods based on tensor models, sparse models, matrices model, neuro-fuzzy models, eigen vectors, low-rank minimization methods, etc are also employed for background subtraction process [11].

IV. RECENT WORKS IN BACKGROUND SUBTRACTION

This section introduced recent achievements of conventional and deep-learning techniques in the field of background subtraction. Table I summarizes the method, achievements and limitations of recent background subtraction algorithms.

Conventional Techniques

Xiang et al. [25] improved the detection of moving objects by combining local intensity ratio model (LIRM) with gaussian mixture model (GMM) that can handle gradual illumination variations and shadows robustly. The morphological operations are employed to handle noise, shadow spots and uneven silhouette. This method does not work with camouflage and sudden illumination variations. Yen et al. [27] introduced a new moving object detection approach for video surveillance. The color and texture based background modeling is combined with hysteresis thresholding and result into an algorithm that restrained the effects of illumination variations, intermittent object motion and shadows. Motion compensation technique used in this method adds noisy regions and has low precision rate in certain cases. Maddalena and Petrosino [28] proposed a neural based background subtraction by implementing self-organizing algorithm. The proposed method is robust to gradual illumination changes, dynamic background and shadows casted by moving object. The performance degrades with sudden light changes and reflections in video scenes. Chen et al. [29] proposed an algorithm (MB-TALBP) for moving object detection. The authors combined background subtraction with edge detection to deal with illumination changes. Background modeling is done by modifying Local binary pattern (LBP) operator. The proposed method is robust to dynamic backgrounds and noisy videos. The performance of this method drops with frequently changed background. Zhou et al. [30] framed the detection of moving object as outlier detection and proposed a unified approach by integrating background learning with object detection. It outperforms other methods in handling dynamic backgrounds by employing low-rank modeling. The foreground is wrongly classified as background for motionless objects and untextured regions in video frames.

A robust scheme named Background motion subtraction (BMS) is introduced by Wu et al. [31] for detecting moving objects from videos taken with moving camera. The adaptive thresholding is applied for foreground segmentation and optimized foreground is extracted by mean-shift segmentation. This method works with different types of video cameras (hand-held cameras, aerial cameras, static cameras, and pan-tilt-zoom cameras) and handles illumination changes effectively. But it can handle detection of moving objects in less video frames. The performance degrades with dynamic backgrounds, fast moving cameras and occlusion.

Deep-Learning Techniques:

Deep-learning techniques revolutionized the field of intelligent video analytics by processing and analyzing large amount of video data [32]. Deep CNN based supervised model has achieved excellent performance in object detection [33].

Christiansen et al. [16] proposed an algorithm by integrating background subtraction with supervised deep convolutional neural network (deep CNN) for detecting anomalies in agricultural fields. The proposed method has low computational time, less memory utilization, high accuracy and also mitigates issues with occlusion and distant objects. The drawback of this approach is it is limited to uniform environments and small occurrence of anomalies. Babaee et al. [23] introduced deep CNN based background subtraction algorithm with spatial-median filtering and global thresholding. It works well with dynamic background, camera jitter, shadows, intermittent object motion, camouflage and thermal videos. But performance drops with bad weather, low frame rate and night videos.

Zhang et al. [34] presented a fast unsupervised deep learning based algorithm that involves two modules for detecting moving objects. First, feature learning is done by deep stacked denoising auto-encoder (SDAE) and then block modeling of binary scenes is done by density analysis. A thresholding based on hash method is used for binarization. It is robust to video noise, bad weather and illumination variations. This method is limited to specific video scenes and requires complex computations.

Braham and Droogenbroeck [35] improved the background subtraction by learning spatial features using deep CNN model and temporal median operator for background modeling. The proposed algorithm deals with hard shadows and night videos. But it requires large number of video frames for training and is also limited to specific scenes. A semi-automatic approach based on cascade CNN model for foreground segmentation form video scenes is proposed by Wang et al. [36]. This algorithm requires little user interventions and handles dynamic background, camera jitter and bad weather. It requires large training frames for complex video scenes especially for night videos.

V. CONCLUSION

This paper clearly manifests the effectiveness and contributions of background subtraction approach for detecting moving objects by reviewing both conventional and deep-learning techniques. The conventional techniques are incapable to handle complex situations. Many statistical methods are reformed by combining different features (color + texture, texture + edge, color + texture + motion) to address complex video scene. Deep-learning techniques for background subtraction have showed remarkable outcomes and provided unified framework to deal with key challenges such as camera jitter, gradual and sudden illumination changes, shadows, camouflage, bad weather, intermittent object motion, and dynamic background. Some deep CNN methods are also robust to night videos and thermal videos. However, deep-learning methods are scenes specific and necessitate large training frames.

In spite of the recent developments in background subtraction, no algorithm can deal with all challenges simultaneously. Effective background subtraction approach is still a great challenge for research community. Future research should consider: recent advancements on deep-learning field, fusion of different techniques to address more complex scenarios, automatic feature selection process and robustness of background model to moving camera.

Table 1. Recent Background Subtraction Algorithm

Reference	Method	Achievements	Limitations
Zhou et al. [30]	Low-rank minimization	Dynamic background	Intermittent object motion & Unsuitable for real-time detection
Xiang et al. [25]	Statistical	Gradual illumination & Shadows	Camouflage & Sudden illumination
Maddalena et al.[28]	Neural	Dynamic background, Shadows & Gradual illumination changes	Sudden illumination changes & Reflections
Zhang et al. [34]	Deep learning	Video noise, Bad weather, & Illumination changes	Complex computations & Specific video scenes
Christiansen et al. [16]	Deep learning	Camera jitter, Shadow, Occlusion, Camouflage & Sudden illumination	Limited to uniform environments
Chen et al. [29]	Advanced Statistical	Illumination changes, Video noise & Dynamic background	Frequently changed background
Braham et al. [35]	Deep learning	Shadows & Night Videos	Specific video scenes & Requires large training frames
Wang et al. [36]	Deep learning	Dynamic background, Camera jitter & Bad weather	Requires large training frames & Night videos
Babaee et al. [23]	Deep learning	Dynamic background, Camera jitter, Camouflage, Shadows, Intermittent object motion & Thermal videos	Bad weather, Low frame-rate & Night videos
Yen et al., [27]	Advanced Statistical	Illumination variations, Shadows & Intermittent object motion	Noisy regions & Low precision rate
Wu et al. [31]	Matrices	Moving camera & Illumination changes	Dynamic background, Fast moving camera & Occlusion

VI. REFERENCES

[1]. Liu, H., Chen, S., & Kubota, N. (2013). Intelligent video systems and analytics: A survey. IEEE Transactions on Industrial Informatics, 9(3), 1222-1233.

[2]. Yilmaz, A., Javed, O., & Shah, M. (2006). Object tracking: A survey. Acm computing surveys (CSUR), 38(4), 13.

[3]. Mishra, P. K., & Saroha, G. P. (2016, March). A study on video surveillance system for object detection and tracking. In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on (pp. 221-226). IEEE.

[4]. Shaikh, S. H., Saeed, K., & Chaki, N. (2014). Moving Object Detection Using Background Subtraction. In Moving Object Detection Using Background Subtraction (pp. 15-23). Springer International Publishing.

[5]. Hu, W., Tan, T., Wang, L., & Maybank, S. (2004). A survey on visual surveillance of object motion and behaviors. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 34(3), 334-352.

[6]. Tiwari, M., & Singhai, R. (2017). A Review of Detection and Tracking of Object from Image and Video Sequences. International Journal of Computational Intelligence Research, 13(5), 745-765.

[7]. Destalem, K., Cho, J., Lee, J., Park, J. H., & Yoo, J. (2015). Dynamic background updating for lightweight moving object detection. International Journal of Computer, Electrical, Automation, Control and Information Engineering, 9(8).

[8]. Zhang, H., & Zhang, H. (2013, April). A moving target detection algorithm based on dynamic scenes. In Computer Science & Education (ICCSE), 2013 8th International Conference on (pp. 995-998). IEEE.

[9]. Gang, L., Shangkun, N., Yugan, Y., Guanglei, W., & Siguo, Z. (2013, July). An improved moving objects detection algorithm. In Wavelet Analysis and Pattern Recognition (ICWAPR), 2013 International Conference on (pp. 96-102). IEEE.

[10]. Cheng, F. C., & Ruan, S. J. (2012). Accurate motion detection using a self-adaptive background matching framework. IEEE Transactions on Intelligent Transportation Systems, 13(2), 671-679.

[11]. Bouwmans, T. (2014). Traditional and recent approaches in background modeling for foreground detection: An overview. Computer Science Review, 11, 31-66.

[12]. Zhang, W., Wu, Q. J., & bing Yin, H. (2010). Moving vehicles detection based on adaptive motion histogram. Digital Signal Processing, 20(3), 793-805.

[13]. Li, Y., Wang, S., Tian, Q., & Ding, X. (2015). Feature representation for statistical-learning-based object detection: A review. Pattern Recognition, 48(11), 3542-3559.

[14]. Yeh, C. H., Lin, C. Y., Muchtar, K., & Kang, L. W. (2014). Real-time background modeling based on a multi-level texture description. Information Sciences, 269, 106-127.

[15]. Sargano, A. B., Angelov, P., & Habib, Z. (2017). A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. Applied Sciences, 7(1), 110.

[16]. Christiansen, P., Nielsen, L. N., Steen, K. A., Jørgensen, R. N., & Karstoft, H. (2016). DeepAnomaly: Combining Background Subtraction and Deep Learning for Detecting Obstacles and Anomalies in an Agricultural Field. Sensors, 16(11), 1904.

[17]. Sobral, A., & Vacavant, A. (2014). A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. Computer Vision and Image Understanding, 122, 4-21.

[18]. Kamate, S., & Yilmazer, N. (2015). Application of object detection and tracking techniques for unmanned aerial vehicles. Procedia Computer Science, 61, 436-441.

[19]. Zaitoun, N. M., & Aqel, M. J. (2015). Survey on image segmentation techniques. Procedia Computer Science, 65, 797-806.

[20]. Huang, Q., Gao, W., & Cai, W. (2005). Thresholding technique with adaptive window selection for uneven lighting image. Pattern recognition letters, 26(6), 801-808.

[21]. SG, A., Karibasappa, K., & Reddy, B. E. (2013).Video segmentation for moving object detection using local change & entropy based adaptive window thresholding. Computer Science & Information Technology, 3(9), 155-166.

[22]. Heikkila, J., & Silvén, O. (2004). A real-time system for monitoring of cyclists and pedestrians. Image and Vision Computing, 22(7), 563-570.

[23]. Babaee, M., Dinh, D. T., & Rigoll, G. (2017). A deep convolutional neural network for background subtraction. arXiv preprint arXiv:1702.01731.

[24]. Zhang, X., Zhu, C., Wang, S., Liu, Y., & Ye, M. (2016). A Bayesian Approach for Camouflaged Moving Object Detection. IEEE Transactions on Circuits and Systems for Video Technology.

[25]. Xiang, J., Fan, H., Liao, H., Xu, J., Sun, W., & Yu, S. (2014). Moving object detection and shadow removing under changing illumination condition. Mathematical Problems in Engineering, 2014.

[26]. Bouwmans, T. (2011). Recent advanced statistical background modeling for foreground detection-a systematic survey. Recent Patents on Computer Science, 4(3), 147-176.

[27]. Yeh, C. H., Lin, C. Y., Muchtar, K., Lai, H. E., & Sun, M. T. (2017). Three-Pronged Compensation and Hysteresis Thresholding for Moving Object Detection in Real-Time Video Surveillance. IEEE Transactions on Industrial Electronics, 64(6), 4945-4955.

[28]. Maddalena, L., & Petrosino, A. (2014). The 3dSOBS+ algorithm for moving object detection. Computer Vision and Image Understanding, 122, 65-73.

[29]. Chen, S., Xu, T., Li, D., Zhang, J., & Jiang, S. (2016). Moving object detection using scanning camera on a high-precision intelligent holder. Sensors, 16(10), 1758.

[30]. Zhou, X., Yang, C., & Yu, W. (2013). Moving object detection by detecting contiguous outliers in the low-rank representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(3), 597-610.

[31]. Wu, Y., He, X., & Nguyen, T. Q. (2017). Moving Object Detection With a Freely Moving Camera via Background Motion Subtraction. IEEE Transactions on Circuits and Systems for Video Technology, 27(2), 236-248.

[32]. Wang, L., & Sng, D. (2015). Deep Learning Algorithms with Applications to Video Analytics for A Smart City: A Survey. arXiv preprint arXiv:1512.03131.

[33]. Ouyang, W., Zeng, X., Wang, X., Qiu, S., Luo, P., Tian, Y., ... & Wang, K. (2017). DeepID-Net: Object Detection with Deformable Part Based Convolutional Neural Networks. IEEE transactions on pattern analysis and machine intelligence, 39(7), 1320-1334.

[34]. Zhang, Y., Li, X., Zhang, Z., Wu, F., & Zhao, L. (2015). Deep learning driven blockwise moving object detection with binary scene modeling. Neurocomputing, 168, 454-463.

[35]. Braham, M., & Van Droogenbroeck, M. (2016, May). Deep background subtraction with scene-specific convolutional neural networks. In Systems, Signals and Image Processing (IWSSIP), 2016 International Conference on(pp. 1-4). IEEE.

[36]. Wang, Y., Luo, Z., & Jodoin, P. M. (2016). Interactive deep learning method for segmenting moving objects. Pattern Recognition Letters.

[37]. Wang, Y., Jodoin, P. M., Porikli, F., Konrad, J., Benezeth, Y., & Ishwar, P. (2014). CDnet 2014: an expanded change detection benchmark dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 387-394).

[38]. Vacavant, A., Chateau, T., Wilhelm, A., & Lequièvre, L. (2012, November). A benchmark dataset for outdoor foreground/background extraction. In Asian Conference on Computer Vision (pp. 291-300). Springer, Berlin, Heidelberg.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Imbalanced Data set in machine Learning : A Comparative Study

Paarth Gupta[1], Pratyush Kumar[2], Manoj Kumar[3]
[1]DoCSE,SMVDU, Katra, Jammu and Kashmir, India
[2]DoCSE,SMVDU, Katra, Jammu and Kashmir, India
[3]AssistantProfessorDoCSE, Katra, Jammu and Kashmir, India

ABSTRACT

A system should be termed as intelligent only when it has the capability of self-learning. Machine learning being one of the most prominent field of computer science can help the system being able to get into the self-learning mode without the need of explicit programming efforts. The major challenge faced by Machine Learning experts in real life scenarios is uneven data distribution leading to imbalanced data set. Thus the proper distribution of elements in the form of sets plays a major role in achieving the self-learning goal. The uneven distribution of elements can be broadly categorized in the majority (negative) class and the minority (positive) class. The distribution of elements in nearly equal proportions is called as balanced data set. A data set is imbalanced when we have a minority class (I.e. the class which is rarer than the other classes namely the majority class). Dealing minority class is becoming more complex as classification rules tend to be fewer and weaker as compared to majority classes. Recent research findings in the area of machine learning along with the data mining have provided deeper insight into the nature of imbalanced learning along with the newer emerging challenges. Thus, this area of research is still popular among research community. In this paper we are focusing on the challenges and its best fit solutions available. Our aim to find the best fit solution by using different machine learning techniques or algorithms. These algorithms may vary in their approaches to solve the given problem. These approaches can be sampling, clustering, Graphical techniques, and statistical techniques or even with the help of classifiers. This paper provides a discussion on the complication of imbalanced data set and solutions concerning lines of future research for each of them.

Keywords: Machine learning, Imbalanced data, Imbalanced clustering, Sampling, Classifiers, Self-Learning.

I. INTRODUCTION

In the recent research done by the researchers in the field of self-learning, they came out with the few building blocks required for an intelligent system. One of the major building blocks is the concept of machine learning which focusses on the development of the computer programs that can access the data and use it for learning purpose. For the effective working, the complete data is divided into various classes based on various parameters. To achieve the objective of partitioning the data in classes we also prefer to have uniform distribution which we also call as balanced data set. But in real life scenario we might come across some cases in which the distribution of classes is not uniform and hence leads to the problem of imbalanced data set. The imbalanced class problem has become a very common problem with the emergence of machine learning. Its importance grow when researchers realized that their dataset is imbalance and this may cause suboptimal classification performance. Among those classes one is major class (having more number of instances) and another is the minor class. Thus, generally the standard classifiers selects the instance from the major class because the ratio of elements in major class to smaller is 1:100,1:1000,1:10000 (and sometimes even more) [1]. This problem is very common in many real life situations and applications like [1] [2] detection of fraudulent telephone calls or credit card transactions, medical problems. Class imbalances have been also observed in many other application problems such as detection of oil spills in satellite images, analyzing financial risk, predicting technical equipment failures, managing network intrusion, text categorization andinformationfiltering. Let us consider an example of class imbalance in which instances in training data belonging to one class heavily outnumber the instances of the other class. E.g. we are visiting the hospital for the test of a rare disease, the chances are we might be having the disease or we might not be having the disease. The possibility of having such a disease is very less so the data set formed is imbalanced. Considering we took the test and the possibilities are the test might be positive (having the disease) or negative. Taking the first case if the result is positive and the test went out correctly which means we have that disease. The test might have gone wrong which will at max lead us to some more tests but if the test comes out negative and the test went out correctly which means we are not having the disease. The last and the most important case is that if the test comes out negative but the test did not go correctly this will mislead the patient as he will be thinking that he is fine but the truth is he was have ng that disease but the test didn't go properly. From this situation we can conclude that it is difficult to apply the concept of machine learning when we are dealing with the situation ofminorityclass[1].

In this real life situation, data describing an infrequent but important event, the learning system may have difficulties to learn the concept related to the minority class. In a data set with the class imbalance problem, the most obvious characteristic is the skewed data distribution between classes [3]. From the recent research this is not only the main problem for the data set including skewed there are various points such as small sample size and the existence of within-class sub concepts. Based on the nature of the problem of the imbalanced data set there are many problems occur such as imbalanced class distribution, small sample size, within class-subclass problem. In further section we are going to discuss the complications related to imbalanced data set and their best known solutions.

II. COMPLICATIONANDITSSOLUTION

This section of our research is based on the thought that though the degree of imbalance is one of the factor that hinders learning but is not the only factor that causes the hindrance. As it turns out, data-set complexity is also the primary determiningfactorofclassificationdeterioration.

2.1 Addressing the imbalanced class problem:

preprocessingandcostsensitivelearning:

This section deals with the problem of two-class imbalance problem both for standard learning algorithms and for ensemble techniques. These approaches are basically divided in three differentgroups.

2.1.1Data–LevelMethods

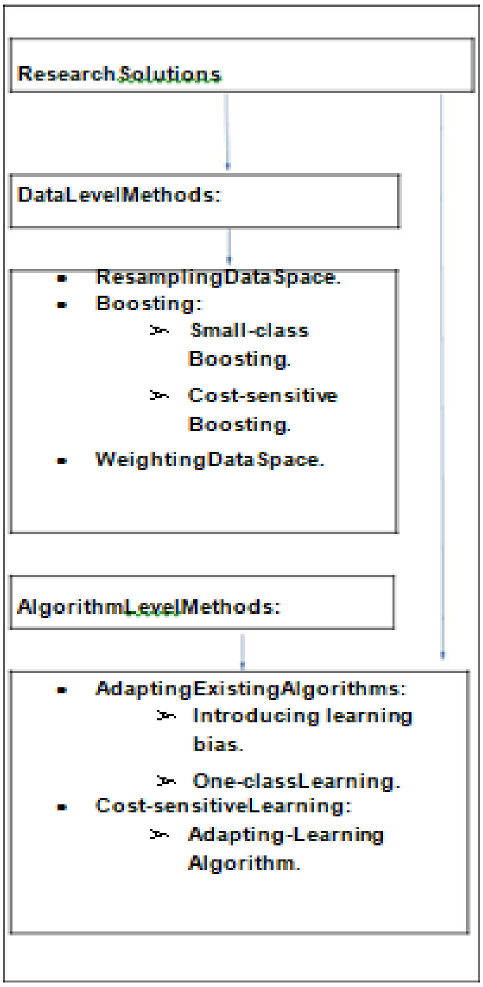
This is one of the easiest methods to handle the above described problem in which we modify the already existing collection of examples just to balance the distribution or in some case we may remove some difficult samples [4][5]. But the only disadvantage of this method is that sometimes all the samples are important and we cannot afford to ignore any one of the existing sample.

2.1.2Algorithm–LevelMethods

This is one of the important methods in which instead of modifying the data we modify the existing algorithms to work efficiently with the minority class. This method gives more accurate result but is less efficient as we might have to change the existing algorithms according to the problem we need to solve and because of this modification the algorithm might take some extra timeaswell[4][6].

2.1.3HybridMethods:

It is basically the combination of the first two methods taking the advantages of both the methods. It is quite a flexible method as we have the flexibility to modify the algorithms according to the need in the problem and we can also remove some of the very rare samples to make the data set balanced and obtain high efficiency [4][7][8].



2.2 Binaryimbalancedclassproblem

This is a problem which was experienced taking various real life applications in consideration like, patient (sick or healthy), access request (valid/authentic or malicious) and so on. In all these scenarios we usually have two options to look forward. These two options are responsible for the construction of two different classes (majority and minority) which are well defined andcanbeeasilydistinguished[9].

The best way to handle such a problem can vary on the type of problem but the best possible solution in most of the case is to try to balance the classes that we already have. To balance the already existing imbalanced classes we usually use the concept of oversampling or undersampling according to the problem we may face. If we use the under sampling method, it creates a subset of the original data-set by eliminating some of the examples of the majority class so that both the class tend to be balance. On the other hand if we use oversampling methods that create a superset of the original data-set by replicating some of the examples of the minority class or creating new ones from the original minority class instances with the same objective [1] [10]. Some of the techniques that uses oversampling or undersampling methods are:

2.2.1 Synthetic Minority Oversampling Technique (SMT)

In this technique the concept of oversampling is used by taking each minority class sample and introducing some synthetic examples along the line segments joining all of the k minority class nearest neighbors. The requirement of amount of oversampling decides the neighbors that is selected from the k-nearest neighbors. While applying the SMT, some majority class examples invade the minority class space and vice versa can also be possible, since the minority class clusters can be expanded by the interpolating of minority class examples. So, there is need of introducing artificial minority class examples deeply into the majority class space. In this situation induction of classifiers can lead to overfitting, in this scenario we use SMT-ENN (edited nearest neighbor), this is the extended version of SMT. As we have already discussed SMT randomly synthesizes minority instances along a line joining a minority instance and it's selected nearest neighbors, while it ignores the nearby majority class instances [1] [10].

2.2.2Random-Oversampling(ROS)

Random-Oversampling is one of the non-heuristic method which has the primary aim to balance the already existing imbalanced classes through the process of replication of the minority class instances as discussed earlier. Recent research proves that random over-sampling can increase the likelihood of occurring overfitting, since it makes exact copies of the minority class examples. This process also increases the computational task if the data set is already fairly large but imbalanced. The major drawback of this technique is that it increases the duplicate samples in the minority class as it follows the simple process of replication[5][15].

2.2.3Random-Undersampling(RUS)

Random-Undersampling is another non-heuristic method which has the primary aim to balance the already existing imbalanced classes through the process of elimination of majority class samples. The logic behind doing such a thing is that it tries to balance out the dataset. The few disadvantages of this technique are that it might discard some of the potentially useful data from the dataset that could be important for the induction process and another problem with this approach is that in the estimation the probability distribution since the distribution is unknown so we take the help of samples available to us [5] [15].

2.3Multi-classImbalanceClassification

In imbalanced data set classification there are some case where we need multiple class distribution and while distributing we may obtain some imbalanced classes as well. Considering a real life scenario of intrusion detection we may have more than two classes of imbalanced class distributions and this hinder the classification performance. Here we take an example of network intrusion detection problem [11] [12], in this distribution while classification of dataset each record represents either an intrusion or a normal connection. Four kinds of attacks are possible in this problem but in the detection of rare classes among fours attacks have very low identification percentage as compared with the other attacks. This increases the complication in the classification performance of the imbalanced data set. The presence of the multiple imbalanced classes in classification of the data set results in complicated situations, so those methods that tackle the class imbalance problem of binary applications are not diretly applicable. For binary-class applications, we have to change the solutions at data level to change the class size ratio of the two classes, either by performing oversampling on the smaller class or down-sampling on the prevalent class, and to get the optimal distribution we run the learning algorithm many times. But due to increase in sample space practically binary class application is not feasible in presence of multiple classes. So, to resolve this problem we extend the binary classifiers. We also use algorithm to boost up the binary-class application to handle the multiple class imbalance problem of imbalanced data set. Here we use the AdaC2.M1 [12] [13] algorithm which has the task to advance the classification of the multi-class imbalanced class. It is a cost sensitive boosting algorithm. It's very effective in biasing the learning from the data set that is directed by the cost setups generated by GA, and eventually creates a significant improvement in the identification performance of those rare instances.

2.3.1Static-SMT

One of the technique to solve the multi-class imbalance classification is the Static-SMT, a preprocessing mechanism. In this technique the resampling procedure is usually applied in ‘n’ steps, where n is the number of classes of the problem. It uses the oversampling technique which has been discussed earlier, with each iteration the resampling procedure selects the minimum size class (minority class) and performs the oversampling by adding the duplicates of the instances of the class in the originaldata-set[1][10].

Comparative analysis with the help of experimental study over the multi class classificationmethodologies: In data mining algorithms, because of the imbalance in the classes formed we have to face a lot of problems. One of the problems is related to the boundaries among the classes i.e. the boundaries among the classes may overlap as there is a very little difference in the samples that lie on the boundaries.
The main problem because of the overlap of boundary samples is that this causes reduce in the performance level. Having such a situation, one of the best possible solution is to reduce the gap between binary class and multiple class imbalanced dataset. In order to implement this technique we have two different strategies[14]:

1. We can try to divide the multiclass problem into simpler binary sub-problems.
2. For each sub-problem that we have we can apply the solutions of two-class imbalanceddata-sets.
3. ignoring the examples that do not belong to the relatedclasses[15].

The OVA (One-versus-all approach) builds a single classifier for each of the classes of the problem, considering the examples of the current class to be positives and the remaining instancesnegatives[16].

2.4 Learning Difficulties with Standard Classifier Modeling Algorithms:

In this section, a subset of well-developed classifier learning algorithms over imbalanced data set is discussed. One of the most popularly used algorithm being the decision tree algorithm, which uses the simple knowledge representation to classify various examples into a finite number of classes. The concept is nearly the same as that in the data structure, where the nodes of the tree represent the content or the attributes. Their edges will be representing the possiblevalues

Table 1

Approach	**Algorithm**	**Remark**	**Reference**	**Fundamental**
OVO + Cost sensitive/ Oversampling.	DecisionTree	Average	Basic algorithm andsimpleOVO	Recursively Splitting thetrainingdata.
	Support Vector Machine	Trulycompetitive	OVO + preprocessing	Binarization and seeking an optimal separating hyperplane to maximize the margin and minimize thetrainingerror.

	K-Nearest Neighbor	Robust performance	Global CS, SL-SMT,SMT	Deciding the class label of test sample by the most abundant class within the K-NearestNeighbors.
	Association Classifiers	Average	OVO and other Algorithm	Deriving classification rules from association pattern.

For the conversion of multiple-class classification problem to a set of binary classification problems we have various techniques, some of them being the OVO (pair wise learning) and OVA approaches.

In OVO technique our main focus is to train the classifier for each possible pair of classes, for corresponding node and finally the leaves which have the responsibility to represent the class labels. The only problem with such technique is that it faces a lot of difficulty in the constructionoftree.

2.4.1DecisionTree

A decision tree can be modelled in two phases named as tree building and tree pruning. The step of tree pruning overcome the overfitting of the training samples and it also improves the generalization capability of a decision tree by trimming the branches of the initial tree. In class imbalance problem decision trees may need to create many tests to distinguish the small classes from the large classes efficiently. In other learning processes, the branches for predicting the small classes may be pruned as being susceptible to overfitting. The basic principle behind pruning is that of predicting errors as there is a high probability that some branches that predict the small classes are removed and the new leaf node is labeled with a dominantclass[17].

2.4.2.Backpropagationneuralnetworks

Another most widely used technique to solve such a problem is by the backpropagation (BP) algorithm, which is most widely used model for such classification problems. In a backpropagation neural network, it usually comprises of one input layer, one output layer, and one or more hidden layers. In each layer we have one or more neurons. The first step of this technique involves initializing the weights to random numbers ranging between -1 to 1. Then our aim is to train the BP network using iterative approach.

While applying this technique we can observe that the error for the samples in the prevalent class reduces whereas the samples for the small classesincreases[18][19].

2.4.3.K-nearestneighbor

Another important and simple solution is K-Nearest Neighbor (KNN) which is an instance-based classifier learning algorithm in which we use specific training instances to make predictions without having to maintain a model derived from data. In this algorithm we compute the distance between the test sample and all of the training samples to determine its k-nearest neighbors. As we have already discussed when we have the imbalanced training data, to identify the samples from the smaller class or the minority class is very difficult. Given a test sample, if we try to calculate k-nearest neighbors it is highly probable to get the result in favor of prevalent class only. Hence, test cases from the small classes are prone to being incorrectlyclassified[20].

III. CONCLUSION

We have discussed the problems because of the imbalanced data set and their challenges along with the appropriate solutions. In data mining community the class imbalance is very pervasive. It is very intrinsic in some applications such as fraud detection, medical diagnosis, and network intrusion detection, modern manufacturing plants, detection of oil spills from radar images of the ocean surface, text classification and direct marketing etc. Some of these applications, such as fraud detection, intrusion detection, medical diagnosis, etc. We come across the experimental studies for the multiple-class imbalanced data-sets with the aim of comparison among the different approaches for the achievement of comparative study. In that section we have highlighted the performance of different algorithm with respect to the binarization techniques with preprocessing of instances such as SMT, SL-SMT, and Global CS algorithm developed for multiple classes. This is going to be the another interesting research issue which is open for learning from imbalanced data set and classification of imbalanced data with multiple class labels. In classic pattern recognition problems there is need of mutually exclusive classes. Classification performance levels decreases when the classes overlap in the feature space. There is a complicated situation occur where the classes are not mutually exclusive. With the applications of these kinds of classes, the class imbalance problem is present. Combined with the multiple class label issue, the class imbalance problem assumes an even more complex situation. Due to the intriguing topics and tremendous potential applications, the classification of imbalanced data will continue to receive more and more attention in both the scientific and the industrial worlds. By this work we tried to provide the basis for the achievement of high quality solutions for imbalanced data-sets with multiple classes, but its significance lies also in the fact that it opens futuretrendsofresearch.

IV. REFERENCES

[1]. Krawczyk, Bartosz. "Learning from imbalanced data: open challenges and future directions." Progress in Artificial Intelligence 5, no.4(2016):221-232.

[2]. Sun, Yanmin, Andrew KC Wong, and Mohamed S. Kamel. "Classification of imbalanced data: A review." International Journal of Pattern Recognition and Artificial Intelligence 23,no.04(2009):687-719.

[3]. Batista, Gustavo EAPA, Ronaldo C. Prati, and Maria Carolina Monard. "A study of the behavior of several methods for balancing machine learning training data." ACM Sigkdd ExplorationsNewsletter 6,no.1(2004):20-29.

[4]. FernaNdez, Alberto, Victoria LoPez, Mikel Galar, MaríA Jose Del Jesus, and Francisco Herrera. "Analysing the classification of imbalanced data-sets with multiple classes: Binarizationtechniquesandad-hoc approaches." Knowledge-based systems 42 (2013):97-110.

[5]. Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16(2002):321-357.

[6]. Barandela, Ricardo, Jose Salvador Sanchez, Vicente Garcıa, and Edgar Rangel. "Strategies for learning in class imbalance problems." Pattern Recognition 36, no. 3 (2003): 849-851.

[7]. Domingos, Pedro. "Metacost: A general method for making classifiers cost-sensitive." In Proceedings of the fifth ACM SIGKDD international conference on Knowledge

discovery and data mining, pp. 155-164. ACM, 1999.

[8]. Wozniak, Michal, Manuel Graña, and Emilio Corchado. "A survey of multiple classifier systems as hybrid systems." Information Fusion 16(2014):3-17.

[9]. Krawczyk, Bartosz. "Learning from imbalanced data: open challenges and future directions." Progress in Artificial Intelligence 5, no.4(2016):221-232.

[10]. Bunkhumpornpat, Chumphol, Krung Sinapiromsaran, and Chidchanok Lursinsap. "Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem." Advances in knowledge discoveryanddatamining (2009):475-482.

[11]. Tax, David MJ, and Robert PW Duin. "Using two-class classifiers for multiclass classification." In Pattern Recognition, 2002. Proceedings. 16th International Conference on, vol. 2, pp. 124-127. IEEE,2002.

[12]. Wang, Shuo, and Xin Yao. "Multiclass imbalance problems: Analysis and potential solutions." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)42, no. 4 (2012):1119-1130.

[13]. Sun, Yanmin, Andrew KC Wong, and Mohamed S. Kamel. "Classification of imbalanced data: A review." International Journal of Pattern Recognition and Artificial Intelligence 23,no.04(2009):687-719.

[14]. Fernandez, Alberto, Mara Jose Del Jesus, and Francisco Herrera. "Multi-class imbalanced data-sets with linguistic fuzzy rule based classification systems based on pairwise learning." In International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, pp. 89-98.Springer,Berlin,Heidelberg,2010.

[15]. FernaNdez, Alberto, Victoria LoPez, Mikel Galar, MaríA Jose Del Jesus, and Francisco Herrera. "Analysing the classification of imbalanced data-sets with multiple classes: Binarizationtechniquesandad-hoc approaches." Knowledge-based systems 42 (2013):97-110.

[16]. Hastie, Trevor, and Robert Tibshirani. "Classification by pairwise coupling." In Advances in neural information processing systems,pp.507-513.1998.

[17]. Cieslak, David A., T. Ryan Hoens, Nitesh V. Chawla, and W. Philip Kegelmeyer. "Hellinger distance decision trees are robust

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

A Comparative Study of Statistical and Machine Learning Techniques of Background Subtraction in Visual Surveillance

ABSTRACT

Video is basically collection of images. Through a single image we can take a screenshot of a scene, which helps in detecting motion with sequence. Now a days, video has popular usage in many applications like identification of exceptional behavior in parking, monitoring of traffic, finding the cause of road accidents, detection of pedestrians, ATMs etc. This is done with the help of many applications that include object tracking, motion segmentation using one of its part background subtractions with the help of various algorithms such as particle filter, mean shift method, kalman filter etc. This paper presents a survey on various algorithms that helps in improving the motion of the object. Research is made on motion detection and tracking in videos along with comparative analysis on various algorithms.

Keywords : Meanshift, Kalman Filter, Particle Filter, Motion Segmentation, Background Subtraction.

I. INTRODUCTION

In todays world with increasing technology and population it becomes very challenging for a human eye to analyze and detect each and everything be it in marketing, bussiness, traffic regulation, health, education and so on.Thus, video analytics comes as a life saver in this situation.a video is basically a collection of images.Video anlysis is the process of automatic analysis of videos in order to detect temporal and spatial activities of an object.There are three important aspects to analyze real world objects through videos :Object detection, Object tracking and analysis.Detection in moving object is the way of recognizing the movement of an object in a given region or area.Tracking is defined as evaluating the path of an object in image plane as it moves around in a scene.In this paper ,a detailed anlysis of various algorithms and techniques of motion detection and object tracking has been made with advantages and drawbacks of each [1].

Image processing is the process of manipulating digital images through digital computer. It particularly focusses on developing such type of computer system that is capable of performing processing on an image. Taking digital image as input for the system than system processes that image using algorithms and provides image as output. An image is basically a two dimensional signal which is defined mathematically as f(x,y) such that x and y are two cordinates horizontally and vertically .The value of f(x,y) at any point denotes pixel value at that point of an image. A signal is a function that provides some information, so a higher dimension can be considered as signal. It can be one dimensional ,two or higher dimensional. One such exampe of two dimension is a digital image. Signal are divided into two types:- analog and digital signal. Analog signals are the one that are defined with respect to time. Also it is a continuous signal and defined over continuous independent variables. Such signal is difficult to examine as they carry huge values and are very much exact due to large part of values. Such signalsare represented by sin waves whereas digital signals are easy to analyse and are not continuous signal. Digital means discrete values and uses proper values to provide any information. As compared to analog signal thay are less accurate because they are distinct samples of analog signal that are taken with respect to time. They are not subjected to noise and denoted by square waves. Digital image is formed by taking an image from camera , sunlight is also used as being a source of energy. Some signal is used for producing the image. When sunlight falls upon the object, the amount of light reflected back by the object is sensed by sensor and a continuous voltage signal is produced by the amount of data which is sensed. Then we replace such data into digital form in order to create digital image[2]. Sampling and quantization techniques are involed which providesa digital image. Frame is basically bit by bit transmission of data.

II. LITERATURE SURVEY

VISUAL SURVEILLANCE :- From vision of computer,visual survelliance is one of the research topic that is used to detect, identify and monitor objects by getting the set of images that helps in understanding and describing the behaviour of object simply by converting the old method of monitoring camers by human operators. Depending upon the human involvement, visual survilliance is defined as manual, semi-automatic and fully automatic. In manual visual survillance, only the human is responsible for detecting and performs all the job while viewing the visual information obtaining from several camers. In semi-automatic visual surveillance both human and system , responsible as object tracking is under computer vision algorithm and human operator is responsible for personal identification, job of classification and recognising activity. Low level of video processing is used in it, and most of the task is performed by human . Whereas in fully automatic, no human involvement and the whole job is performed by the system. On comparing all the three , fully automatic system is more intelligent to track, classify and identify object. Also it helps in reporting and detecting the doubtful behaviour and recognizing the activity of object[3],[4],[5].

Various stages of Visual surveillance

visual surveillance has the following stages such as: background modeling, motion segmentation, classification of foreground moving objects, human recognition , monitoring and examining the behavior of object .

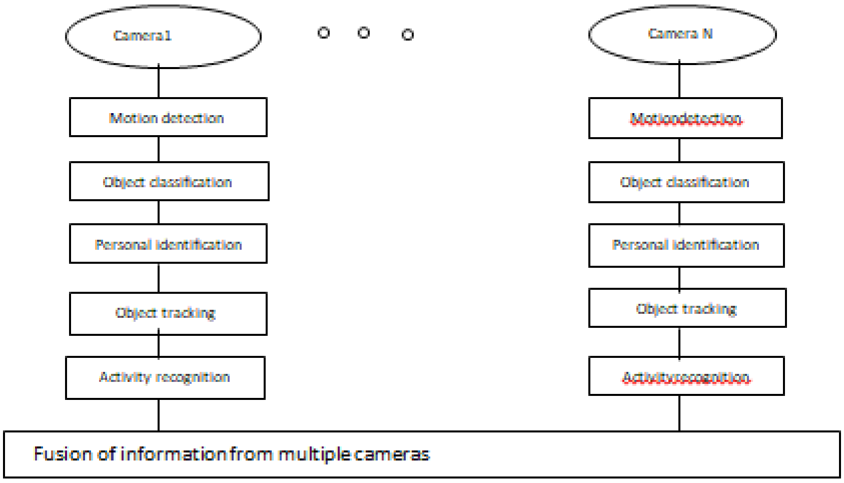


Figure 1. stages of visual surviellance

III. MOTION DETECTION

The first step in visual surveillance system is detecting the motion of an object . Motion detection is the process of segmenting object moving foreground from the remainig image. Motion segmentation is basically done with many techniques such as background subtraction , temporal differencing andso on. Background subtraction is the most popular method among them that helps in detecting the moving regions in an image by taking the exact difference between current image and the background image which is referred and very common technique for detecting the moving object. Very much easy and inexpensive for real time systems, but also sensitive to extraneous events etc. The goal of such technique is to extract current image from reference background image, that is taken during period of time. Otherwise no change occurs and the pixel belongs to the object. The subtracting operations that are applied helps in finding the absolute difference for each pixel, by detecting object that differs from background. If difference is below the threshold then no change in the scene and pixel is observed as it belongs to background. So that is why it is very much dependent on good background maintenance model. There are several algorithms that come under background subtraction[6],[7].

BACKGROUND SUBTRACTION:

Below figure shows the flow chart of background subtraction techniques (BGS).

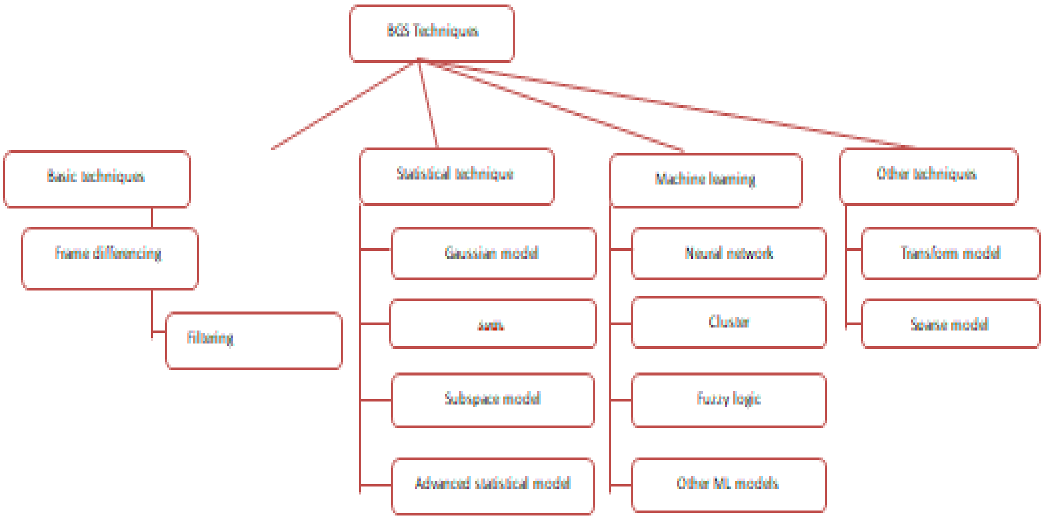


Figure 2. BGS techniques

Statistical techniques

This technique is used to identify the requirements of a process through the method of collection of data from the process and then analyzing them in order to establish control and validate the process capability. It is one such type of background subtraction technique which is more robust towards camera jitter and dynamic background. This technique is further divided into Gaussian model (GM), support vector model (SVM), subspace and advanced statistical model [8]. The figure below shows the flowchart of its various techniques [9], [10], [11].

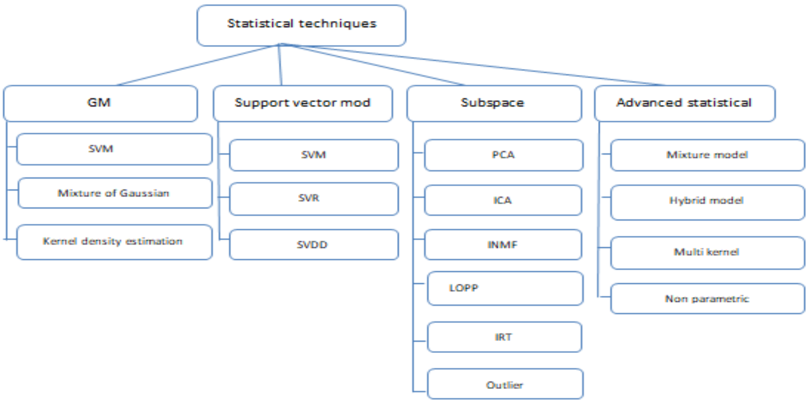


Figure 3

Table 1

Name	Type/category	Precision	Computation time	Performance in bad weather	Performance in low frame rate	Advantages and disadvantages
Single Gaussian method	Gaussian model	0.4036	Less(1.32)	Performs very well	Less robust	Works well for static background only .Not suitable for shadows &camera jitter
Mixture of Gaussian	Gaussian model	0.6336	More(4.91)	Performs well	Most robust	Static background only. Cannot detect shadows or any changes to background
Support vector machine	Support vector model	High precision	Low	Performs well	Less robust	Suitable for applications with camera jitter and dynamic background
ViBe	Non-parametric model	0.7793	More(4.8)	Performs very well	Most robust	Fails to detect shadows, works well for camera jitter and illumination
KDE	Gaussian model	0.3700	Very high(13.80)	Works very well	Less robust	Works well for dynamic background, camera jitter and illumination changes
PCA	Subspace model		Very high	Perform good	robust	Can't be used with dynamic scenes
SVR	Support vector model	High precision	More	Performs well	More robust	Does not work well with more dynamic backgrounds

Median filter	filtering	Good precision	Switching median filter used to improve the speed	Well perform	Less robust	
PBAS	Non-parametric	0.6670	low	Performs good	Most robust	Works well for camera jitter and dynamic background

1. Single Gaussian method:

Ideal conditions like a complete noise-free background cannot be satisfied in real life scenario. Thus, for noise every pixel is used with a Gaussian distribution .The background and foreground represented as

$$|(It - \mu t)|\ \sigma t > k \longrightarrow \text{Foreground}$$

2. Mixture of Gaussian

Single Gaussian method works well for static background only, however, mixture of Gaussian method is a more robust method in this regard [12],[13]. It performs well in bad weather. It supports static background but cannot detect shadows or any changes to background and takes more computation time.

3. Kernel density estimation

It is a non-parametric method for estimating the probability density function of random variables. It is an unstructured approach. It has low robust and works well for dynamic background, camera jitter and illumination changes.

4.Support vector model

It is a more sophisticated model for background subtraction as it uses support vector machine(SVMs) and also uses support vector regression(SVR). In this method the frame is basically partitioned into 4x4 size block and then it is found whether the block is a part of background or not by making use of probabilistic support vector machine(PSVM). SVR also uses the same features as that of SVM for classification with only a few minor differences.

5.Principal component analysis(PCA)

This method makes use of transformation that is orthogonal ,use to convert a series of observations which are possibly correlated into linearly uncorrelated variable which are known as principal components.

6.Visual background extractor(ViBe)

In this method, a series of values taken previously is calculated for each pixel in the same location or in neighborhood. Then a compression is made of these values with the current pixel value in order to determine whether the pixel leads to the background or not. It is a non-parametric statistical approach.

7.Pixel-based adaptive segmenter(PBAS)

In this method , a segmentation decision is taken based on the fact that how close is the value of the pixel to the corresponding pixel model which is created by an array of recently observed background values. If the value of pixel closer than a certain threshold value, then it is said to belong to the background.

Comparative analysis of different statistical techniques :-

Machine learning

Machine learning is one of the types of artificial intelligence which helps a machine to perform activities without explicitly being programmed which at times human can perform better. The basic idea behind machine learning is to build algorithms that can receive output data and use statistical analysis to find out the output of the process. Following are some of the machine learning techniques which can be used for BGS. Now the flowchart below shows different techniques under machine learning.

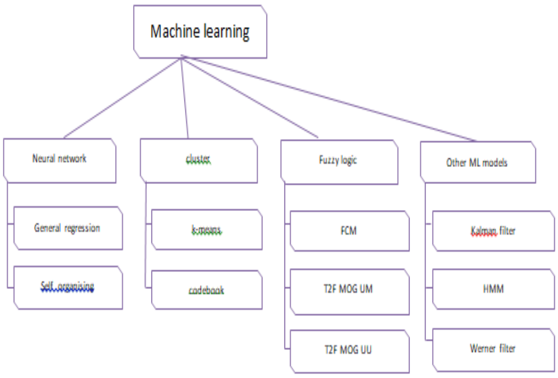


Figure 4

1. Artificial Neural network.

ANN can be used in BGS in a very efficient manner. The background information is stored at each neuron and a neural network which is three layered is used that comprises of aninput layer, hidden layer and output layer. Various variants of neural network have been implemented over years.

2. Codebook

It is an efficient technique to deal with multi model background. A sequence of key values called code words is assigned to each background pixel based on the training sequence which determines the pixel a color which is most likely to take over a certain period.

3. K-means clustering.

Clustering is the process of dividing a set of data points to a small number of clusters. It is is a method in which we find the locations of clusters that reduces the distance of cluster from data points.

4. Kalman filter

This technique uses its recursive nature. It is basically a recursive method in which weighted average of previous estimates and the new information is used and the weights are optimized to reduce the squared error.

5. Hidden Markov model

A method where there is sequence of outputs but the sequence of states the model went through to generate the output is not known. In image processing the conventional block based classification ignored the context information. Therefore in order to improve classification by context hidden markov model was proposed[14],[15].

Comparative analysis of different machine learning techniques :-

Table 2

Name	Approach	Computational time	accuracy	Memory requirement
Artificial neural network.	Non parametric	training is time consuming	Less accurate	Less memory required
Decision tree	Non parametric	Less time consuming	Less accurate	Less memory required
SVM	Non parametric with binary classifier	Training is time consuming	Less accurate	Scalability improves by reducing memory use
Fuzzy logic	Stochastic approach	Less time consuming	Less accurate	Less memory requirement
k-nearest neighbor	Parametric	Less time consuming	Less accurate when large data set is there	More memory needed
Kalman filter	Non parametric	Less time consuming	exact	Limited memory
Hidden markov	Stochastic approach	Time consuming	Less accurate	Less memory

IV. CONCLUSION

This paper presents a comparative study of various background modeling techniques. This paper categorizes the available background modeling approaches into four general types, basic, statistical, machine learning and other approaches. The main two categories i.e statistical and machine learning techniques of background subtraction are focused upon in this paper. Comparison of such methods depends on their cpu or memory requirements as well as their capability of correctly detecting motion in different kind of videos. As far as statistical methods are concerned , methods such as MOG and KDE has proved very good model accuracy. KDE has a high memory requirement. Gaussian methods perform very well but work for static background only while SVM techniques work well with dynamic background and camera jitter and their computation time is also very efficient. In case of machine learning methods , the new approaches like fuzzy logic are less time consuming. Also the kalman filter and k-nearest neighbor approaches are good in this regard. As far as accuracy is concerned kalman filter gives a very accurate result also requires less memory than k-nearest neighbor approach. Thus, overall each method has its own advantages and disadvantages. All we can do is try to incorporate the good features of some techniques into other so that it can result intogood technique.

V. REFERENCES

[1]. Deepak Kumar Panda, and SukadevMeher. Robust Object Tracking Under Background Clutter. In Proceedings of International Conference on Image Information Processing, Nov. 2011 JUIT Shimla, India.

[2]. Deepak Kumar Panda, and SukadevMeher. Robust Object Tracking Under Varying Illumination Conditions. InProceedings of IEEE India Conference INDICON, Dec 2011. BITS Pilani Hyderabad Campus, India.

[3]. L. Li, S. Ranganath, H. Weimin, and K.Sengupta"Framework Framework for Real-Time Behavior Interpretation From Traffic Video "IEEE Tran. On Intelligen Transportation Systems, , Vol. 6, No. 1, pp. 43-53, 2005.

[4]. P. Kumar, H. Weimin, I. U. Gu, and Q. Tian "Statistical Modeling of Complex Backgrounds for Foreground Object Detection" IEEE Trans. On Image Processing, Vol. 13, No. 11, pp. 43-53, November 2004.

[5]. Z Zivkovi "Improving the selection of feature points for tracking" In Pattern Analysis and Applications, vol.7, no. 2, Copyright Springer-Verlag London Limited, 2004.

[6]. Vinay D R, N Lohitesh Kumar, " Object Tracking Using Background Subtraction Algorithm" in International Journal of Engineering Research and General Science Volume 3, Issue 1, January-February, 2015.

[7]. Rohan K. Naik, " A Robust Background Subtraction Technique for Object Detection" in the International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 2, February 2015.

[8]. P. Kumar, H. Weimin, I. U. Gu, and Q. Tian "Statistical Modeling of Complex Backgrounds for Foreground Object Detection" IEEE Trans. On Image Processing, Vol. 13, No. 11, pp. 43-53, November 2004.

[9]. D.-M. Tsai and S.-C. Lai, "Independent component analysis-based background subtraction for indoor surveillance," image Processing, IEEE Transactions on, vol. 18, no. 1, pp. 158–167, 2009.

[10]. T. Bouwmans, "Traditional and recent approaches in background modeling for foreground detection: An overview," Computer Science Review, vol. 11, pp. 31–66, 2014.

[11]. M. Piccardi, "Background subtraction techniques: a review," in Systems, man and cybernetics, 2004 IEEE international conference on, vol. 4. IEEE, 2004, pp. 3099–3104.

[12]. S. Y. Elhabian, K. M. El-Sayed, and S. H. Ahmed, "Moving object detection in spatial domain using background removal techniques-stateof-art," Recent patents on computer science, vol. 1, no. 1, pp. 32–54, 2008.

[13]. T. Bouwmans, F. El Baf, and B. Vachon, "Background modeling using mixture of gaussians for foreground detection-a survey," Recent Patents on Computer Science, vol. 1, no. 3, pp. 219–237, 2008. 14. T. Gao, Z.-g. Liu, W.-c. Gao, and J. Zhang, "A robust technique for background subtraction in traffic video," in Advances in NeuroInformationProcessing. Springer, 2008, pp. 736–744.

[14]. Tao Zhao, RamNevatia, FengjunLv, "Segmentation and Tracking of Multiple Humans in Complex Situations," In Proc. IEEE Computer Society Conf. on Computer Vision and pattern Recognition, 2001,Vol.2, pp. 194 – 201.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Machine Learning Techniques in IoT

Dr. Naveen Kumar Gondhi, Rohini Raina

Department of Computer Science and Engineering, SMVDU, Katra, Jammu and Kashmir, India

ABSTRACT

Internet of Things(IoT) is expanding expeditiously in different fields but has a tremendousimplementation in the branch of savants and commence. Machine learning can also help machines;indulge them collectively to acknowledge them what people want from the data made by Homosapiens. And moreover machine learning has an important role in IoT facet to hold the large extent of data generated by the machines. Machine learning gives internet of things and those machines a mind to think, which is called "encapsulated intelligence" by some researchers. This paper will mainly focus on machinelearningintelligent algorithmslike artificial immune system algorithmBayesian theorem, Geneticalgorithm (GA), Swarm algorithm (SA), algorithm,Bayesian theorem, Reinforcementalgorithm, Ant colonyalgorithm, k-means algorithm and supportive vector machine algorithm and their role in internet of things.

Keywords : IoT, Machine Learning, Encapsulated Intelligence.

I. INTRODUCTION

The Internet of Things (IoT) is a combination of integral computing devices, digital era, or anything which used a network for data transfer. The term thing, in the Internet of Things, can be anything like car with the sensors, home gadgets anything that can be assigned a unique IPaddress .IoT is all about connecting the device through internet.. Internet of things all works on some kind of intelligence and this intelligence is machine learning. Though, there are stills lots of obstacles and challenges to overcome. But this all obstacle can be overcome by machine learning.

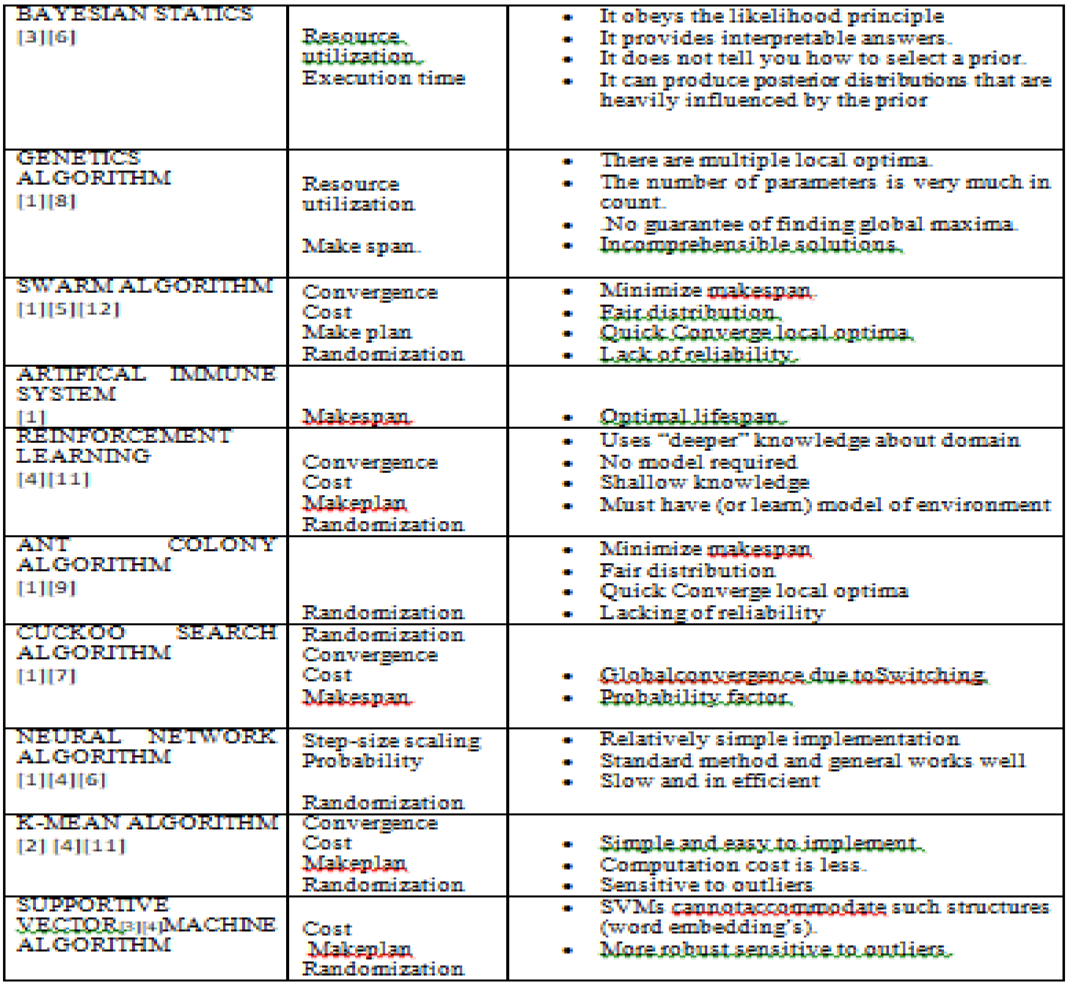
Literature Survey: Machine learning is the discipline of getting computers to act without being explicitly programed. Machine learning has given us everything fromself-drivingcars toWeb search machine learning has made our understanding towards human genome easy. Machine learning made the human being progressive in every aspect. Today machine learning is used in each and every field that we can use it several times without actually knowing it.In this paper, analysis of various intelligent techniques applied to IoT.

1) Artificial Immune System: The Artificial Immune System is [1] a meta-heuristic algorithm based on foundation of the immune system.Artificial immune system is taken from the algorithm of biology which studies the immune system and immunology. In simple words immune system protect the body from the various disease. From this algorithm the idea of creating artificial immune system come to an existence. Artificial immune system also applicable on the internet of things; it is very excellent idea to create a protective layer known as artificial immune system to protect and find to find invader in a smart device where the artificial system is behaving similar to the natural system..The most sought after properties of an artificial immune system are robust, lightweight, error tolerant, distributed.

Robustness is to pass the infected data to the artificial immune system for processing the data might be incomplete or contain the noise. Lightweight property of artificial immune system helps the smart device to not consume a large amount of power to perform its operation. Heterogeneous property helps in protecting the device from transmission of incomplete and malfunction data. Thus the artificial immune system has a vital role in internet of things

Table for advantages and disadvantages of various techniques

Table 1



BAYESIAN STATICS [3][6]	Resource utilization. Execution time	• It obeys the likelihood principle • It provides interpretable answers. • It does not tell you how to select a prior. • It can produce posterior distributions that are heavily influenced by the prior
GENETICS ALGORITHM [1][8]	Resource utilization Make span.	• There are multiple local optima. • The number of parameters is very much in count. • .No guarantee of finding global maxima. • Incomprehensible solutions.
SWARM ALGORITHM [1][5][12]	Convergence Cost Make plan Randomization	• Minimize makespan. • Fair distribution. • Quick Converge local optima. • Lack of reliability.
ARTIFICAL IMMUNE SYSTEM [1]	Makespan.	• Optimal lifespan.
REINFORCEMENT LEARNING [4][11]	Convergence Cost Makeplan Randomization	• Uses "deeper" knowledge about domain • No model required • Shallow knowledge • Must have (or learn) model of environment
ANT COLONY ALGORITHM [1][9]	Randomization	• Minimize makespan • Fair distribution • Quick Converge local optima • Lacking of reliability
CUCKOO SEARCH ALGORITHM [1][7]	Randomization Convergence Cost Makespan.	• Globalconvergence due toSwitching. • Probability factor.
NEURAL NETWORK ALGORITHM [1][4][6]	Step-size scaling Probability Randomization	• Relatively simple implementation • Standard method and general works well • Slow and in efficient
K-MEAN ALGORITHM [2] [4][11]	Convergence Cost Makeplan Randomization	• Simple and easy to implement. • Computation cost is less. • Sensitive to outliers
SUPPORTIVE VECTOR[3][4]MACHINE ALGORITHM	Cost Makeplan Randomization	• SVMs cannotaccommodate such structures (word embedding's). • More robust sensitive to outliers.

2) Genetic algorithm: Genetic Algorithm [1][8] is a searching technique which works randomly based on Darwin theory. It uses current and historical data to analyse the future and this technique is used in VM scheduling. GA is based on the biological concept of increasing the population.. The Genetic Algorithm-Placement of IoT Device (GA-PID) decides the placement point where the IoT device should be allocated to carry out the task. : In this selecting individuals from the parental generation and interchanging their genes, new individuals (descendants) are obtained. Genetic algorithm is used to find the multi object optimization problem. The genetic algorithm is used to minimize path length which is used to give maximum network life. This observation is especially important for the IoT device problem.

3) Swarm algorithm: Swarm algorithm isa highly advanced heuristic intelligent optimization algorithm that follows thebehaviour of animal swarm. It is searching algorithm that gives global best information through collaboration between individuals.Swarm [1][5][12] optimization is used efficiently to enhance physiological multi-sensor data fusion measurement precision in the Internet of Things.Swarm optimization (IPSO) is used to solve the convergence accuracy speed and local optimization of IoT devices.

4) Bayesian theorem: Bayesiantheorem [3] is a statistics theorem which explains that information About the true state is shown in terms of degrees of beliefwhich is also called as Bayesian probabilities. Such kind of interpretation is type of a number of interpretations of probability. It has great implementation in the Internet of things. The algorithm applies on the Internet of things devices to find the occupancy of the room using a PIR Sensor. This algorithm is also estimates the battery running occupying estimation in internet of things.

5)Reinforcement Learning: Reinforcement learning is a method oflearning [4][11] in machine learning to allow machine behave according to the environment or by interacting the environment .it works on the trial and error method. Reinforcement learning works in a cycle of sense-action-goals. Becausereinforcement learning learns from immediate interaction with the environment. Reinforcement learning learns from the immediate interaction with environment. Reinforcement learning is used in IoT.There are sensors,induction refrigerator, a.c, electric glass the mind or the science behind this is reinforcement learning because they adapt the environment and make changes according to it.

6) Ant Colony Algorithm: Ant colony algorithm [1][9] is an approach which is being extracted from the behaviour of the ant's, like the ants which secretes chemical material known as pheromones. By which they implicitly communicate with other ants. When an ant explores and finds some object such as food, it secretes a pheromone along the route back to the colony. This algorithm is also used in IoT in finding the route and communication among these nodes. According to the features of the IoT such as the irregular Network topology, many nodes, the more variable network structure, this algorithm is used to search path, and used to broadcastthe signal which is featured with the random sending. Ant colony algorithm can reduce the broadcast methodefficiently. With the number of nodes in the search in routing was increased, the time of route setup was significantly shortened.

7)Cuckoo search algorithmCuckoo [1] search algorithm is a meta-heuristic algorithm that models natural behaviour of cuckoo species. Cuckoos are the beautiful birds but their aggressive reproduction strategy is more interesting to us. The cuckoos [7]reproduce in such a way that only one egg is laid at a time and laid it in a nest randomly and in next step the nest which has the better quality eggs will be carried further for the next generation. This algorithm has a vitalrole in the Internet of Things (IoT). Error correction is of great significance to achieve Iot precision. Currently, accurately predicting the future dynamic measurement of error is an effective way to improve IoT precision. Aiming to solve the problem of low model accuracy in traditional dynamic measurement error prediction. This study employs to predict the dynamic measurement error of sensors. However, the execution of the SVM depends on setting the appropriate parameters. Hence, the cuckoo search (CS) algorithm is adopted to optimize te key parameters to avoid the local minimum value which can occurs when using the traditional method of parameter optimization.

8) Neural Network Algorithm: Neural network [1][4][6] algorithm is a method used in machine learning to calculate the error contribution of each neuron after a batch of data. The neural network is categorized in two networks: hierarchical network and interconnected, which is categorized according, to the neuron functionality in the different layers. These layers are input , hidden and output layer, which are connected in a sequence. Neural network is widely used in internet of things to accurately classify the input data. The data of sensors has been distinguished on the basis neural network .by neural network .By neural network; the response time of overall network can be reduced. And can increase the performance of the sensors.

9)K- means Algorithm: K-means is a [2]unsupervised learning which is famous for giving the solution of the cluster analysis. This method has some easy rules .This method distinguish a given data into different number of clusters (assume it k-cluster).The main motive is to find k-centroids for every cluster .The centroids are placed in such order that they are not near to each other, they should be far from other one. The next step is to choose each point from a given data set and relate it to the nearest centroid. When there is no point remains, the starting step is ended. On reaching this step re-calculate new k-centroid as barycentre's then step is second is done again but with new k-centroids. As aresult of previous step a loop is generated which signifies that k- centroids move from its place step by step till there are not any further changes. The algorithm main aim is to minimizing the objectivefunction. K-mean algorithm is also used to find the best area in the smart city which is suitable for living and where the air pollution is less and fumes gas from the traffic system is negligible among the whole polluted smart city. K-means clustering algorithm that can use both of the trajectory variables and the associated chemical value to classify source regions of definite chemical category.

10) Support Vector Machine Algorithm In this approach [3] SVM algorithm explains to learn to distinguish data points using labelled practice samples. Basically, the problem is to distinguish those points in two different fragments. These [4] fragments are placed by as far as possible ends and new reading will be distinguish on basis of which side of the ends it is. Supportive vector machine algorithm, which contains optimizing a quadratic function with linear constraints .This type of approach, is highly used in IoT and basically for the automatic traffic accident detectionwhich includes to the interrelation of computing devices and sensors using radiofrequency identification. An intelligent transportation framework based on IoT is the finest application of the supportive vector machine algorithm.

II. CONCLUSION

IoT is modifying our existence. Machine learning changes the scenario of dealing of human with machine and retrieving the data from them. Today machine learning not only connecting the machines but also making the human interaction with machines easy. Some of the application of machine learning comes to existence and more to come in future which is somehow uncertain and magical.

III. REFERENCES

[1]. Dr Naveen Kumar Gondhi, Ayushi Gupta, "Survey on Machine Learning Based Scheduling In Cloud Computing, ISMSI '17, March 25-27, 2017, Hong Kong, Hong Kong 2017 ACM. ISBN 978-1-4503-4798-3/17/0

[2]. Sadegh Bafandeh Imandoust ,Mohammad Bolandraftar "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: TheoreticalBackground, S B Imandoust et al. Int. Journal of Engineering

Research and Applications Vol. 3, Issue 5, Sep-Oct 2013, pp.605-610

[3]. CHIH-CHIA YAO, PAO-TA YU "Effective Training Of Support Vector Machines Using extractive Support Vector Algorithm,1-4244-0973-X/07/$25.00 2007 IEEE

[4]. Yue Xu"Recent Machine Learning Applications to Internet of Things (IoT),http://www.cse.wstl.edu/ jain/cse570-15/ftp/iot –ml/index.html

[5]. Wen-Tsai Sung , Yen-Chun Chiang" Improved Particle Swarm Optimization Algorithm for Android Medical Care IOT using Modified Parameters, Received: 3 February 2012 / Accepted: 19 March 2012 / Published online: 11 April 2012#Springer Science+Business Media, LLC 2012

[6]. DanDanCui,FeiLiu "The Application of BP Neural Network in Internet of Things, Advanced Engineering Forum Vols 6-7 (2012) pp 1098-1102 (2012) Trans Tech Publications, Switzerlanddoi:10.4028/www.scientific.net/AEF.6-7.1098

[7]. AlexanderTeske, RafaelFalcon, AmiyaNayak" Efficient detection of faulty nodes with cuckoo search indiagnosable systems

[8]. Bimlendu shahi,Sujata Dahal,Abhinav Mishra Vinay Kumar.S.B,Prasanna kumar.C "A review over Genetic Algorithm And application of wireless Network systems,International C onference on Information Security &privacy(icisp2015),11-12 december 2015,Nagpur,INDIA

[9]. Chao Cheng, Zhi-hong Qian"An IoT Ant Colony Foraging Routing Algorithm Based on Markov Decision Model, International Conference on on Soft Computing in Information Communication Technology (SCICT 2014)

[10]. Xin Tao, Chunlei Ji" Clustering Massive Small Data for IOT, 2014 2nd International Conference on Systems and Informatics (ICSAI 2014)

[11]. Ale Al-Fuqaha, Mehdi Mohammadi." Semi-supervised Deep Reinforcement Learning Support of IoT and Smart City Services, IEEE INTERNET OF THINGS JOURNAL, VOL. X, NO. X, XXXXX 2017

[12]. A. Carlisle, G. Dozier, An off-the-shelf PSO, in: Proceedings ofthe Workshop on Particle Swarm Optimization, Indianapolis, INPurdue School of Eng. Technol., IUPUI, April 2001.

[13]. B. Scholkopf, A.J. Smola, Learning with kernels: Support vector machines, regularization, optimization,and beyond, Cambridge, Mass: MIT Press, London, 2002.

[14]. P rajesh P Anchalia,Anjan K Koundinya,Srinath N K .MapReduce Design of K-means Clustering Algorithm. IEEE. 2013.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

A Survey on Machine Learning: Concept, Algorithms, and Applications

Sakshini Hangloo[1], Samreen Kour[2], Sudesh Kumar[3]
[1, 2]M.Tech, Department of computer science, Shri Mata Vaishno Devi University, J&K, India
[3]PhD Scholar, Department of computer science, Shri Mata Vaishno Devi University, J&K, India

ABSTRACT

In today's era machine learning concepts and algorithms are heavily used in the digital world. Machine learning algorithms can easily understand how to perform important tasks by generalizing from examples. Machine learning is often feasible and cost-effective approach where manual programming is not. From the past few decagons, Machine learning (ML) made software application more accurate to predict outputs. Also, various algorithms that are designed in machine learning are continuously used for pattern recognition, data clarification, and various other plans and have lead to a distinct research in data mining to determine underground consistencies or inconsistencies in collective data. The main objective of this paper is to discuss various concepts, approaches and procedures of machine learning used in addressing the digital world problems.

Keywords: Machine Learning, Precision, Training data, Procedures

I. INTRODUCTION

Machine learning is an area of computer science that is interrelated with designing the systems in a manner that the system itself learn and upgrade with experience. Learning means identification and interpretation of input data and decision-making based on the provided data.

The programmer sometimes has a certain motive in mind while framing a machine (a software system). Consider the case of Strike Series of Robert Galbraith and Potter Series of J.K. Rowling, two skilled persons were indulged from "The London Sunday Times" and use "Forensic Machine Learning" and they confirm that the Rowling actually wrote the books by the name "Galbraith". They both program a machine learning algorithm and "edify" it with Rowling's and other examples by writers to find out and understand the hidden patterns and "confirm" by Galbraith book. Rowling's and Galbraith's writing matched the most in various features is ended up by this algorithm.

By Machine Learning's use, a researcher trys to obtain an perspective through which the machine, i.e., the algorithm will come up with its own solution based on the example or training data set given to it initially instead of developing an algorithm to mark the problem directly.

A. MACHINE LEARNING: INTERSECTION OF STATISTICS AND COMPUTER SCIENCE

Machine Learning shows a remarkable outgrowth when Computer Science and Statistics both forces are joined together. Computer Science aims at building machines that find some answer to certain problems, and tries to identify if problems are interpretable. The main approaches on which Statistics fundamentally focuses are data inference, modelling hypothesises and measuring quality of the outcomes.

Machine Learning idea is a slightly different but somehow dependent on both. Computer Science focuses on blue-collar programming, while ML marks the issue of getting computers to re-program themselves whenever revealed to new data rooted on some starting learning plans that are given. On the another hand, Statistics aims at probability and data inference, while Machine Learning involves the practicality and efficacy of construction and algorithms to operate those data, constituting certain learning functions and performance measures.

B. HUMAN LEARNING AND MACHINE LEARNING

Machine Learning is the learning of human and animal brain in the technical learning of the nervous system, Psyche and associated domains and a third study area that is closely concerned with it. The researchers suggested that how a machine could understand from experience is not different than how an animal or a human mind understands. However, the researches correlated with statistical - computational approach is far better than the research focusing on exploring machine learning problems using human brain's training methods which did not produce a much optimistic outcome. This might be happen due to the reality that human or animal psyche remains not completely perceivable to date. Association between human study and machine study is growing, nevertheless of these difficulties and for machine learning is used to describe certain understanding techniques present in human or animals. For example, to describe neural signals in humans study, machine study function of time-related difference was presented. It is equitably anticipated that in coming years, this association is to grow greatly.

C. DATA MINING, ARTIFICIAL INTELLIGENCE, AND MACHINE LEARNING

The three approaches Machine learning, Artificial Intelligence, and Data Mining are concerned with each other and together can produce highly effective and responsive outcomes. Data mining places the initial point for both machine learning and artificial intelligence and is primarily about explicating the data. In action, it examines and identifies patterns and correlations that occur in information which is tough to explain manually. Therefore, data mining is not a basic function to show a presumption but function for producing appropriate presumption. Artificial intelligence may be explained as machines that have the capability to resolve a given situation on their own without any human involvement. The important data and the AI illuminating the data produce an answer by itself rather than answer developed straight into the system. The illumination that goes under is nothing but a data mining algorithm. Machine learning promotes the approach to a greater level by giving the data important for a machine to train and modify accordingly when revealed to new data. This is called as "training". It centres on exploring information from large sets of data, and then discovers and determines underneath patterns using various statistical measures to increase its capability to explain new data and gives more efficient outcomes. At last, if a system lacks the ability to learn and improve from its previous exposures then that system cannot examine to be completely intelligent.

II. PRESENT RESEARCH QUESTION & RELATED WORK

The various applications declared above suggest somehow advancement in machine learning and their fundamental theory. Machine learning is a deep learning and various researchers have suggested their views in this field. In this paper, the major research question that is being taken at present are explained and it gives the references to some of recent work.

A. BY THE USE OF UNLABELED DATA IN SUPERVISED LEARNING

For supervised learning, labelled data are necessary. Most of the time, they are often obtainable in less volume, while unlabeled data may be huge [7]. Combining both unlabeled data and labelled data is of great interest, both in a theoretical and a practical sense. In recent time, for joining unlabeled and labelled data various approaches have been recommended. Supervised learning algorithms check the closeness of relation between the features and labels. The problem with this approach is that the predefined information is not consistently provided [8]. Before going for supervised classification, we process, filter and label the information using unsupervised learning, there by adding to the total cost. Supervised learning problems often have the following property: class labels have a high cost while unlabeled examples have little or no cost. This rise in cost can be reduced greatly if unlabelled data is used by supervised learning (e.g., images). In various interesting cases of learning problems with extra presumption, unlabeled data can be truly validated to upgrade the expected perfection of supervised learning [21]. For example, identifying spam emails or classifying web pages. Presently active researchers are seriously taking into consideration the new algorithms or new learning problems for making the use of unlabeled data effectively.

B. TRANSFERRING THE LEARNING EXPERIENCE

To ease the learning process to carry out a new task, transfer learning uses the facts from past similar tasks which are its main goal. The advantage of transfer leaning is mainly regulated by a minimization in the number of training examples. To minimize complexity of samples, training examples need to have a target performance on a sequence of similar learning problems, matched with number required for unfamiliar problems. In various real-life cases, only a few new concepts, training examples or process are often enough for a human learner to grab the new concept and define it. For example, learning for driving a bus becomes very much easier task if we have the knowledge of how to drive a car. For various problems of real life, the supervised algorithm may include learning a group of associated functions than just learning a single one. Even if the scientific determination functions for distinct cities (e.g., Jammu and Canada) are assumed to be comparatively different, some similarities are sure as well.

C. LINKING DIFFERENT ML ALGORITHMS

In various fields or domains, the number of machine learning algorithms have been brought into notice and tested. Among the existing ML algorithms, one experiment of research aims to detect all the possible correlations and suitable cases to use a particular algorithm [23]. Now consider two supervised classification algorithms, Naive Bayes' and Logistic Regression. They both differently tend to various data sets, but when implemented to specific types of training data, their significance can be explained. In general, the ML algorithms conceptual understanding, their combined features, and their respective efficiency and limitations to date will remain an intrinsic research matter.

D. BEST STRATEGICAL APPROACH FOR LEARNERS WHO COLLECT THEIR OWN DATA

A wider research direction centres on learning systems that energetically assembles data for its own processing and learning instead of mechanically using data assembled by some other plans. Most of the research time is given in exploring the powerful scheme to fully pass over the power to the learning algorithm. For example, to check the behaviour of a patient by considering a drug test system to learn the outcome of all possible hidden side effects and trying to reduce them.

E. PRIVACY-PRESERVING DATA MINING

For automatically and intelligently withdrawing information or knowledge from a huge number of data, which can also reveal delicate information about particular understanding the particular's right to privacy, Data mining is a popular technique. Moreover, critical information about business transactions, compromising the free competition in a business setting can be revealed by data mining techniques. Therefore, for this reason, privacy-preserving data mining (PPDM) has become a major field of study. In data mining, PPDM becomes a fresh research area, where data mining algorithms are for possible violation of privacy. PPDM research generally works on three philosophical approaches:(1) data hiding, where delicate raw data like identifiers, name, addresses, etc. are altered, blocked, or trimmed from the original database, in order for the users of the data not to compromise with another person's privacy; (2) rule hiding, where delicate knowledge explored from the data mining process keeps out for use, because private information may be extracted from the disclosed knowledge; (3) secure multiparty computation, in which distributed data is released or shared for computations, but before that it is encrypted; thus, everyone knows about its own inputs and the results but not everything. The PPDM goal is to develop effective algorithms that allow exploring relevant knowledge from a huge number of data, while preventing the delicate data and information from the broadcast.

III. MACHINE LEARNING ALGORITHMS CATEGORIZATION

Over past years a number of ML algorithms have been designed and introduced. These algorithms are broadly grouped into two categories on the basis of learning style and similarity. In this section, we will inculcate some basic idea of various types of ML algorithms.

A. GROUP BY LEARNING STYLE

1. Supervised learning

In supervised learning, the machine is provided with a given set of inputs or training data with their desired outputs or predetermined labels e.g. True/False, Positive/Negative etc. The machine needs to study those given sets of inputs and outputs, and find a general function that could predict the label of test data. The supervised learning can be of regression or classification type.

2. Unsupervised learning

Unlike in supervised learning, here the Input data or training data is not labelled which makes this type of learning harder. One of the approaches is clustering where the training data is grouped on the basis of similarity.

3. Semi-supervised learning

In semi-supervised learning, the training data contains both labelled and unlabeled data. The aim is to develop an algorithm that will predict classes of future test data better than the earlier algorithm that used only the labelled data. The way humans learn is similar to semi-supervised learning.

4. Reinforcement learning

In this type of learning, algorithm maps action to the situation and receives reward or penalty for its actions in trying to solve a problem. After several trial and error runs it learns the best policy i.e. the sequence of actions that maximize the total reward.

B. ALGORITHMS GROUPED BY SIMILARITY

1. Instance-based Algorithms

The Instance-based model simply stores instances of training data instead of developing a definition of target function. Each time when a new problem arises it is compared with the previously stored data in order to predict and determine the value of target function. This is done by assign the value of a target function to the new instance, provided that is a better fit than the former and hence these algorithms are also known as winner-take-all method. Examples of Instance-based Algorithm are K-Nearest Neighbour (KNN), Locally Weighted Learning (LWL), Learning Vector Quantisation (LVQ), Self-Organising Map (SOM), etc.

2. Genetic algorithm

The Genetic algorithm provides a learning method that is similar to biological evolution. Instead of search from general-to-specific hypotheses, GA generates successor hypotheses by repeatedly mutating and crossover of the best currently known hypotheses to generate new genotype in the hope of finding good solutions to a given problem.

3. Decision Tree Algorithms

Decision tree algorithms, one of the most widely used methods for inductive inference. It is a type of supervised learning. A decision tree is a tree-like structure consisting of all possible solutions to a problem based on certain constraints. It begins with a single simple decision or root, which then extends to a various branches until a decision is made, forming a tree and hence named as decision tree. Some of its examples are Classification and Regression Tree (CART), Conditional Decision Trees, Chi-squared Automatic Interaction Detection (CHAID), etc.

4. Bayesian Algorithms

Bayesian algorithms use Bayes' Theorem to solve classification and regression kind of problems. Bayesian offers a possible outlook for logic estimation. It is based on the impression that the quantity of interest is governed by distribution of probability and that optimal decisions can be made by reasoning about these probabilities along with the observed data. Some of the examples of Bayesian algorithm include Naive Bayes, Bayesian Network (BN), Gaussian Naive Bayes, Bayesian Belief Network (BBN), etc.

5. Support Vector Machine (SVM)

In SVM, we plot each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. It uses a separating hyperplane among a set of data points which splits the data into two differently classified groups. SVM is a supervised classification method and can perform both linear and nonlinear classification.

6. Association Rule Learning Algorithms

Association rules aim to discover a relationship between various variables in a huge database. They are widely used in many applications areas like Market Basket analysis, intrusion detection, bioinformatics etc. Common examples are Apriori algorithm, FP Growth algorithm, Éclat algorithm etc.

7. Artificial Neural Network (ANN) Algorithms

ANN is a model based on the structure and operations of actual neural networks of the living being. ANNs are regarded as non-linear models. ANN discovers complex associations between input and output data by selecting sample from data rather than considering the entire data set and thereby reducing cost and time. Examples: Back-Propagation learning, Hop-field Network, Perceptron etc.

8. Deep Learning Algorithms

Deep learning algorithm consists of multiple hidden layers in an artificial neural network. This approach tries to work in the same way as the human brain processes light into vision and sound into hearing. They produce results comparable to human experts and in some cases the results are even better.

Some applications of deep learning are computer vision and speech recognition [1]. Examples of deep learning algorithms are Deep Boltzmann Machine (DBM), Stacked Auto-Encoders Deep Belief Networks (DBN) etc.

9. Dimensionality Reduction Algorithms

Dimensionality simply refers to the number of features or input variables in the dataset. When the number of features is very large, *certain* algorithms struggle to train models effectively, this is called the

Curse of Dimensionality. Dimensionality reduction visualises data with numerous features and helps in implementing supervised classification more efficiently [2]. Examples: Principal Component Analysis (PCA), Multidimensional Scaling (MDS) Principal Component Regression (PCR), Discriminate Analysis (LDA), Partial Least Squares Regression (PLSR), Summon Mapping, etc.

10. Clustering Algorithms

Clustering algorithm divides the population into a number of groups such that the data points in one group are more similar to the other data points in the same group than those which belong to some other groups. In simple words, clustering is concerned with using inherent patterns in the datasets to classify and label the data accordingly [2]. Some of the examples include K-Means, K-Medians, Ward hierarchical clustering, and Mean Shift, Expectation Maximisation (EM) etc.

11. Regression Algorithms

Regression analysis is subset of predictive analytics and visualises the co-relation between dependent and independent variables. Examples of regression models are: Linear Regression, Logistic Regression, and Stepwise Regression, Multivariate Adaptive Regression Spines (MARS) etc.

IV. APPLICATIONS

Machine learning has proved to be the answer to many real-world challenges. In this section, we will discuss some applications of machine learning with some examples. But still, there are a number of problems for which machine learning needs a breakthrough.

1. SPEECH RECOGNITION

In the field of speech recognition, certain methodologies are developed that enable computers to recognize and then translate the spoken language into text. All these systems use machine learning approach for better accuracy. There are various voice-controlled programs such as Apple's Siri, Google Now, Amazon's Alexa, Microsoft's Cortana etc in the market nowadays.

2. COMPUTER-AIDED DIAGNOSES

Computer-aided Diagnosis system assists doctors to interpret medical images. Various medical tests such as X-rays, MRI, ultrasounds etc are the sources of data that describes a patient's condition. Pattern recognition techniques are used to identify suspicious structures in the image to aid Computer-aided diagnosis.

3. COMPUTER VISION

The living beings use their eyes to see the world around them. Computer vision aims to give nearly same capabilities to a machine. It allows a machine to gain high-level understanding from digital images or videos and act accordingly.

Driverless cars are also one of the greatest applications of machine learning where car vision is made possible by advancement in the computer vision technology. To perform these tasks cameras are installed and they get input from these cameras. These tasks lie purely in the pattern recognition domain. A driverless robotic car named STANLEY was first to win the 2005 DARPA Grand Challenge. STANLEY is a Volkswagen Touareg that is equipped with cameras, radar, and laser rangefinders to sense the environment and the onboard software to command the steering, braking, and acceleration [30].

4. GAME PLAYING

IBM's DEEP BLUE became the first computer program to defeat the world champion Garry Kasparov by a score of 3.5 to 2.5. Then the other champions studied Kasparov's loss and were able to draw a few matches in subsequent years. But now highly efficient computer systems have been made and most of the recent human-computer matches have been won by the computer.

5. LOGISTICS PLANNING

In logistics planning, we apply methods for cost reduction, capital reduction, and service improvement. During the Persian Gulf crisis of 1991, the U.S. forces used a Dynamic Analysis and Re-planning Tool (DART), for doing automated logistics planning and scheduling for transportation. This involved approximately 50,000 vehicles, cargo, and people at a time, and had taken into account various parameters like starting points, destinations, routes, and conflict resolution among all these parameters. The AI planning techniques generated a plan in hours that using older methods would have taken weeks.

6. TEXT MINING

It refers to the process of deriving the high-quality information from the text.

There are two different ways of mining the data i.e. Goal-oriented and Method-oriented mining. Any process that generates useful results that are not obvious is called Goal-oriented mining. And any process that involves extracting information from massive amount of data is called Method-oriented mining. Text mining is useful in a number of applications including business intelligence, automated classification of news articles, spam filter, automated placement of advertisement.

V. FUTURE SCOPE

Machine learning is a research area that has attracted the interest of many people and it has the potential to uncover many other problems.

In areas like game playing, logical inference and theorem proving, planning, and medical diagnosis, there are systems that can perform better than human experts. In other areas, such as learning, vision, robotics, and natural language understanding, there is a rapid improvement in performance through the application of better analytical methods. Continued research will give better capabilities in all of these areas.

Some of the most important future problems are discussed here.

A. EXPLAINING HUMAN LEARNING

As mentioned earlier, Machine learning is a field that provides a machine the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning theories try to imitate features of learning in humans and animals. However, the important stimuli in human or animal learning like horror, urgency, excitement, hunger are not yet taken into account in ML algorithms. This is a potential opportunity to discover a more generalised concept of learning.

B. PROGRAMMING LANGUAGES CONTAINING MACHINE LEARNING PRIMITIVES

Majority of applications in ML algorithms are incorporated with manually coded programs as part of the application software. In today's world there is an increasing need for a new programming language that is self-sufficient to support manually written code. This enables the coder to define a set of inputs-outputs for every "to be learned" program and opt for an algorithm. Some of the programming languages like Python are already making use of these concepts but in smaller scope.

C. PERCEPTION

A generalised concept of computer perception that can link ML algorithms is used highly in advanced vision, speech recognition etc. Research in machine perception solves the hard problems of understanding images, sounds, music and video. One of the main problems is the integration of different senses to prepare a system that can induce self-supervised learning to estimate one sensory knowledge using the others.

VI. CONCLUSION

The machine learning field is concerned with the problem of how to construct a computer program that automatically improves with experience. In recent years, many successful machine learning applications have been developed. At the same time, there have been important advances in the theory and algorithms. The foremost target to design more efficient and practical general-purpose learning methods that can perform better over various domains. ML algorithms are completely data-driven and have the ability to examine a large amount of data in smaller intervals of time. Also they are often more accurate and not prone to human bias. ML algorithms have an edge over manual programming as the latter lacks the ability to adapt when exposed to a different environment.

VII. REFERENCES

[1]. Honglak Lee, Roger Grosse, Rajesh Ranganath, Andrew Y. Ng., Convolution Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations, Proceedings of the 26th Annual International Conference on Machine Learning, 2009.

[2]. Kajaree Das, Rabi Narayan Behera., a Survey on Machine Learning: Concept, Algorithms and Application, International Journal of Innovative Research in Computer, and Communication Engineering, Feb 2017.

[3]. N. Cristianini and J. Shawe-Taylor, an Introduction to Support Vector Machines. Cambridge University Press, 2000.

[4]. E. Osuna, R. Freund, and F. Girosi. Support vector machines: training and applications. AI Memo 1602, MIT, May 1997.

[5]. Taiwo Oladipupo Ayodele, Types of Machine Learning Algorithms, New Advances in Machine Learning, Yagang Zhang (Ed.), InTech, 2010

[6]. T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K.Mazaitis, T. Mohamed, N. Nakashole, E. Platanios,A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov,M.Greaves, J. Welling, Never-Ending Learning, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2014

[7]. Wang, J. and Jebara, T. and Chang, S.-F. Semi-supervised learning using greedy max-cut.Journal of Machine Learning Research , Volume 14(1), 771-800 2013

[8]. Chapelle, O. and Sindhwani, V. and Keerthi, S. S. Optimization Techniques for Semi-Supervised Support Vector Machines, Journal of Machine Learning Research , Volume 9, 203–233, 2013

[9]. J. Baxter, A model of inductive bias learning. Journal of Artificial Intelligence Research, 12:149–198, 2000.

[10]. S. Ben-David and R. Schuller, Exploiting task relatedness for multiple task learning, Conference on Learning Theory, 2003.

[11]. W. Dai, G. Xue, Q. Yang, and Y. Yu, Transferring Naive Bayes classifiers for text classification. AAAI Conference on Artificial Intelligence, 2007.

[12]. Z. Marx, M. Rosenstein, L. Kaelbling, and T. Dietterich. Transfer learning with an ensemble of background tasks. In NIPS Workshop on Transfer Learning, 2005.

[13]. R Conway and D Strip, Selective partial access to a database, In Proceedings of ACM Annual Conference, 85 - 89, 1976

[14]. P D Stachour and B M Thuraisingham Design of LDV A multilevel secure relational database

management system, IEEE Trans. Knowledge and Data Eng., Volume 2, Issue 2, 190 - 209, 1990

[15]. R Oppliger, Internet security: Firewalls and beyond, Comm. ACM, Volume 40, Issue 5, 92 - 102, 1997

[16]. Rakesh Agrawal, Ramakrishnan Srikant, Privacy Preserving Data Mining, SIGMOD '00 Proceedings of the 2000 ACM SIGMOD international conference on Management of data, Volume 29 Issue 2,Pages 439-450, 2000

[17]. A. Carlson, J. Betteridge, B.Kisiel, B.Settles,E. R.Hruschka Jr,and T. M. Mitchell, Toward an architecture for never-ending language learning, AAAI, volume 5, 3, 2010

[18]. X. Chen, A. Shrivastava, and A. Gupta, Neil: Extracting visual knowledge from web data, In Proceedings of ICCV, 2013.

[19]. P. Donmezand J. G. Carbonell, Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In proceedings of the 17th ACM conference on information and knowledge management, 619–628. ACM, 2008

[20]. T. M.Mitchell, J. Allen, P. Chalasani, J. Cheng, O. Etzioni, M. N. Ringuetteand J. C. Schlimmer, Theo: A framework for self-improving systems, Arch. for Intelligence 323–356, 1991

[21]. Gregory, P. A. and Gail, A. C. Self-supervised ARTMAP Neural Networks, Volume 23, 265-282, 2010

[22]. Cour, T. and Sapp, B. and Taskar, B. Learning from partial labels, Journal of Machine Learning Research, Volume 12, 1501-1536 2012

[23]. Adankon, M. and Cheriet, M. Genetic algorithm-based training for semi-supervised SVM, Neural Computing and Applications , Volume 19(8), 1197-1206, 2010

[24]. T. M. Mitchell (1997), Machine Learning, McGraw-Hill International.

[25]. Stuart Russell, Peter Norvig (04-Jul-2016), Artificial Intelligence, A Modern Approach, Global Edition, Pearson Education Limited.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Intuitionistic Fuzzy Orienteering Problem and Its Work-Depth Analysis

Madhushi Verma, K. K. Shukla

ABSTRACT

Orienteering is an NP-hard problem that originated from a water sport where a player is required to visit a set of control points connecting the source and the destination, collect the maximum possible rewards or scores associated with the control points and arrive at the destination within the time bound. It finds its application in the tourism industry, telecommunication networks and other computational problems where things like human behaviour and hesitancy of the decision maker must be considered. To tackle the uncertainty involved in the parameters we represent them using trapezoidal intuitionistic fuzzy numbers (TIFN) resulting in intuitionistic fuzzy orienteering problem (IFOP). A technique based on max-min formulation is presented to deal with IFOP using a new method for ranking TIFNs. Also, a work-depth analysis for the parallel version of IFOP is presented to show that IFOP is work-preserving and can be implemented on a multiprocessor model like PRAM to obtain the solution for large instances efficiently.

Keywords: Centroid of Centroids, Fuzzy Optimization, Intuitionistic Fuzzy Orienteering Problem, Orienteering Problem, Trapezoidal Intuitionistic Fuzzy Number.

I. INTRODUCTION

The orienteering problem (OP), which is a mixture of the two well-known problems of combinatorial optimization i.e. the travelling salesman problem (TSP) and the knapsack problem (KP), is NP-Hard. This problem originated from a game where the player has to visit a set of control points connecting the source and the destination within a limited time budget and collect the maximum rewards possible. A lot of real life situations and applications from several fields like logistics, home delivery systems, tourism, building telecommunication networks etc. can be depicted in the form of OP. Several types of OP have been discussed in the literature which includes the team orienteering problem and the simple orienteering problem and a variation of both with time windows [1].

As stated before, the two important parameters associated with OP are time and score, both of which are imprecise in nature and cannot be determined exactly. The best way to tackle the prevailing vagueness is to represent the parameters using fuzzy numbers. Here we prefer intuitionistic fuzzy numbers (IFN) over fuzzy numbers because in OP the endeavor is to obtain a path that helps in achieving the maximum rewards or scores within the specified time bound but in practice, the areas where this problem finds its application, considers the human behavior, knowledge etc. like in tourism industry which leads to uncertainty in determining the values of the two parameters i.e. score and time. The most desirable method of handling these circumstances of insufficient information and lack of precision and certainty is to model the parameters using IFN. In this paper, we model the two parameters (score and time) using trapezoidal intuitionistic fuzzy numbers

(TIFN). TIFNs have been successfully used in several decision-making, applied engineering and scientific problems [2].

A number of heuristics have been proposed to solve the crisp OP. Also, a few approximation algorithms have been stated in the literature that considers this problem. The first heuristic for OP was suggested by Tsiligirides in 1984 [3]. Since then, a lot of heuristics were proposed by Golden et al, Ramesh and Brown, Wang et al and Chao et al [4]-[7]. The other approaches presented to solve OP include the genetic algorithm, tabu search and ant colony optimization suggested by Tasgetiren, Gendreau et al and Liang et al respectively [8]-[10]. Fischetti et al stated a branch and cut heuristic for OP in 1998 and in 2009, two techniques called the pareto ant colony optimization algorithm and the variable neighborhood search algorithm were proposed by Schilde et al for the multi-objective variant of OP [11]-[12]. A Greedy Randomized Adaptive Search Procedure for solving OP was proposed by Campos et al in 2012 [13]. Blum et al suggested a constant factor approximation for the rooted version of OP [14] and Johnson et al proposed an approximation algorithm for the un-rooted version of OP [15].Another approximation algorithm for the time dependent variant of OP was presented by Fomin et al in 2002 [16].

In section II, some necessary definitions are stated and section III, presents a brief description about fuzzy optimization. The mathematical representation of IFOP and the steps of IFOP algorithm is described in section IV and section V respectively. An illustrative example is presented in section VI. In section VII, a work-depth analysis of IFOP is explained. Finally, the paper is concluded in section VIII.

II. PRE-REQUISITES

A. Trapezoidal Intuitionistic fuzzy numbers (TIFN)

In 1986 [17], Atanassov introduced the concept of intuitionistic fuzzy set which is an extension of the Zadeh's fuzzy set where two values are associated with every element of the set, one depicting the degree of belongingness and the other being the degree of non-belongingness. Both these values lie within the real unit interval [0, 1] [2].

An intuitionistic fuzzy set X in U where U is the universe of discourse can be represented as [2]

$$X = \{\langle x, \mu_X(x), \nu_X(x)\rangle : x \in U\}$$

such that

$$0 \le \mu_X(x) + \nu_X(x) \le 1 \ \forall x \in U \qquad (1)$$

Another term that can be linked with every element x in the set X is called the hesitancy degree of x to X and can be defined in the following way:

$$\pi_X(x) = 1 - \mu_X(x) - \nu_X(x)$$

such that

$$0 \le \pi_X(x) \le 1 \, , x \in U$$

In this paper, we use an IFN for which the real line is the universe of discourse i.e. $U = \mathcal{R}$ and can be stated as $X = \{\langle x, \mu_X(x), \nu_X(x)\rangle : x \in \mathcal{R}\}$. An IFN X possess the following properties [18]:

a. The membership function and the non-membership function is fuzzy convex and fuzzy concave respectively (if-convex).
b. There exists at least two points x_1 and x_2 in U such that $\mu_X(x_1) = 1$ and $\nu_X(x_2) = 1$
c. The membership function (μ_X) and the non-membership function (ν_X) is upper semicontinuous and lower semicontinuous respectively.

According to the above definition, TIFN A can be represented using eight numbers $A = \langle (a, b, c, d), (e, f, g, h)\rangle$ where $a, b, c, d, e, f, g, h \ \in \mathcal{R}$ such that $e \le a \le f \le b \le c \le g \le d \le h$ and the functions $L_A, M_A, N_A, K_A : \mathcal{R} \to [0,1]$. The definition of the membership function and a non-membership function of A is as follows [2]:

$$\mu_A(x) = \begin{cases} 0 & \text{if } x < a \\ L_A(x) & \text{if } a \le x < b \\ 1 & \text{if } b \le x \le c \\ M_A(x) & \text{if } c < x \le d \\ 0 & \text{if } d < x \end{cases} \qquad (2.a)$$

$$\nu_A(x) = \begin{cases} 1 & \text{if } x < e \\ N_A(x) & \text{if } e \le x < f \\ 0 & \text{if } f \le x \le g \\ K_A(x) & \text{if } g < x \le h \\ 1 & \text{if } h < x \end{cases} \qquad (2.b)$$

Where $L_A(x) = \frac{x-a}{b-a}$, $M_A(x) = \frac{x-d}{c-d}$, $N_A(x) = \frac{x-f}{e-f}$, $K_A(x) = \frac{x-g}{h-g}$. If $b = c$ and $f = g$ then the trapezoidal intuitionistic fuzzy number is reduced to a triangular intuitionistic fuzzy number.

B. Addition of TIFN

Two TIFN $A_1 = \langle(a_1, b_1, c_1, d_1), (e_1, f_1, g_1, h_1)\rangle$ and $A_2 = \langle(a_2, b_2, c_2, d_2), (e_2, f_2, g_2, h_2)\rangle$ can be added using the following formula [2]:

$$\begin{aligned} A_1 + A_2 &= \langle(a_1, b_1, c_1, d_1), (e_1, f_1, g_1, h_1)\rangle \\ &\quad + \langle(a_2, b_2, c_2, d_2), (e_2, f_2, g_2, h_2)\rangle \\ &= \left\langle \begin{matrix} (a_1 + a_2, b_1 + b_2, c_1 + c_2, d_1 + d_2), \\ (e_1 + e_2, f_1 + f_2, g_1 + g_2, h_1 + h_2) \end{matrix} \right\rangle \qquad (3) \end{aligned}$$

C. Expected Value of TIFN

To determine the degree up to which the constraints of the problem are satisfied by the TIFN (representing either the total score or the total time taken), we need to determine the expected value of TIFN using the below stated formula [2]:

For a given TIFN $A = \langle(a, b, c, d), (e, f, g, h)\rangle$

$$EV(A) = \frac{1}{8}(a + b + c + d + e + f + g + h)$$
(4)

D. Ranking of TIFN

To rank the TIFN, we introduce a technique called Centroid of Centroids (CoC). The centroid of a fuzzy number signifies its geometric centre and is denoted using the formula: $\int_{-\infty}^{\infty} xf(x)dx / \int_{-\infty}^{\infty} f(x)dx$. A trapezoid can be divided into three figures (two triangles and a rectangle) and finding out the centroid of each and joining them forms a triangle. The centroid of this resultant triangle can be considered to be a balancing point and a better point of reference. As the task here is to rank a trapezoidal intuitionistic fuzzy number, we evaluate the centroid for both the trapezoids using the following formula and the figure shown below:

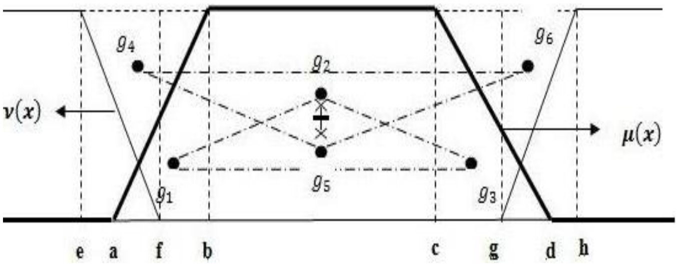


Figure 1. The point of reference used for ranking a TIFN

The centroid of the triangle $g_1g_2g_3(C_1)$

$$C_1 = (x_1, y_1) = \left[\frac{(2a+b+7c+2d)}{18}, \frac{7}{18}\right] \qquad (5)$$

The centroid of the triangle $g_4g_5g_6(C_2)$

$$C_2 = (x_2, y_2) = \left[\frac{(2e+f+2h+7g)}{18}, \frac{11}{18}\right] \qquad (6)$$

Then the rank of the TIFN can be calculated using the following formula:

$$\text{Rank (R)} = \sqrt{\left(\frac{x_1+x_2}{2}\right)^2 + \left(\frac{y_1+y_2}{2}\right)^2} \qquad (7)$$

E. Fuzzy Decision Set ($\boldsymbol{Z}$)

The fuzzy version of the problem under consideration, when formulated as an integer programming problem, can have several goals each of which can be depicted as a membership function and a fuzzy set (F_i) consisting of the elements along with their membership values [19]. The set comprising of the feasible solution elements is called a fuzzy decision set which is as follows:

$Z = F_1 \cap F_2 \cap \dots\dots\dots \cap F_i$

$$\text{i.e. } \mu_Z(x) = \mu_{F_1}(x) * \mu_{F_2}(x) * \dots\dots\dots\dots..* \mu_{F_i}(x) \qquad (8)$$

Here, $*$ is a $t-$norm denoting any operation like algebraic product, minimum etc. and for the stated problem $*$ represents the minimum operation. The element in the fuzzy decision set Z with the highest membership value is the most desirable solution represented by the set Z^* as shown below:

$$\mu_{Z^*}(x^*) = \max[\mu_Z(x)] \qquad (9)$$

III. FUZZY OPTIMIZATION

In most of the problems from the field of engineering design and decision making, it is difficult to conclude with the most optimal solution from a set of feasible solutions. The most appropriate method to deal with this kind of a situation is to tackle the uncertainty in the variables that lead to the optimal solution. The randomness that comes into existence due to natural variations and fluctuations can be handled using the probabilistic concepts but to take care of the uncertainty that is due to the vague nature of the objective, linguistic statements of the decision maker showing his willingness (like acceptable solution or satisfactory solution etc.), qualitative statements etc., we introduce the concept of fuzzy optimization where the optimization problems are solved using fuzzy logic. In the crisp optimization problems, there is an objective function which is to be maximized or minimized and at the same time the stated constraints should be satisfied, if not the solution is unacceptable. However, in fuzzy optimization we induce a certain amount of relaxation to this restriction of satisfying each and every constraint completely. In case of fuzzy optimization, the solution is a matter of degree i.e. we define degree of acceptability or degree of satisfaction which can be expressed using membership functions. So, the objective function and the constraints can be represented as fuzzy goals using membership functions and we intend to come out with a solution which is called the “best compromise solution” that helps in achieving these goals. Along with fuzzy goals, crisp constraints may also be required to state the physical conditions, technological feasibility etc. that should be present in a solution. The technique of fuzzy optimization provides flexibility to the objective function and latitude to the constraints as a result of which we can obtain more than one solution for a particular problem but each one may have a different degree of acceptability and depending upon the willingness and requirement, the decision maker can select the most appropriate solution. Also, fuzzy optimization helps in obtaining a solution which is not a 0-1 type solution by quantifying the preferences of the decision maker and tackling the uncertainty in the decision making problems, that comes into existence due to imprecision ,vagueness etc. through membership functions [20]-[22].

IV. PROBLEM DEFINITION

The OP can be presented in the form of a graph $G(V, E)$ where V and E denote the set of vertices and the set of edges respectively. This graph is a weighted undirected completely connected graph. The weight assigned to every vertex $v_i \in V$ and every edge $e_{ij} \in E$ denotes the parameter score (S_i) and the time taken to traverse each edge (t_{ij}) respectively. The task in OP is to obtain a path P that connects the source vertex (v_1) and the destination vertex (v_n) and also includes any subset ofV such that the total collected score is maximized within the specified time budget T_{max} [1].

Here, we introduce the intuitionistic fuzzy orienteering problem (IFOP) where the parameters (score and time) are represented using TIFN. In IFOP, the strict requirements of the crisp formulation which include the maximization or minimization of the objective function, satisfying each and every constraint and giving equal importance to all the constraints are relaxed to some extent by using fuzzy logic with the aim to provide a more accurate and realistic modeling of the real world. In the fuzzy formulation, we consider the willingness of the decision maker, his aspiration levels and the degree up to which a solution is acceptable or its degree of satisfaction and using intuitionistic fuzzy numbers this modeling can be made more apt as the extra information stating the degree of non-belongingness along with the degree of belongingness is the best way to tackle the vagueness [23].

The fuzzy version of OP provides latitude to the solution by relaxing the constraints to some extent and representing the objective function of maximizing the score and the constraint of satisfying the time bound as fuzzy goals using linear membership functions. The remaining constraints are crisp as shown below:

$$\sum_{i=1}^{N-1} \sum_{j=2}^{N} \widetilde{S_\imath}\, x_{ij} \gtrsim S_{min} \quad (10)$$

$$\sum_{j=2}^{N} x_{1j} = 1 \quad , \quad \sum_{i=1}^{N-1} x_{iN} = 1 \qquad (11)$$

$$\sum_{i=1}^{N-1} x_{ik} \leq 1 \quad \forall\, k = 2, \dots\dots, N-1 \qquad (12)$$

$$\sum_{j=2}^{N} x_{kj} \leq 1 \quad \forall\, k = 2, \dots\dots, N-1 \qquad (13)$$

$$\sum_{i=1}^{N-1} \sum_{j=2}^{N} \widetilde{t_{\imath\jmath}} x_{ij} \lesssim T_{max} \qquad (14)$$

$$2 \leq u_i \leq N \quad \forall\, i = 2, \dots\dots, N \qquad (15)$$

$$u_i - u_j + 1 \leq (N-1)\left(1 - x_{ij}\right) \quad \forall\, i, j = 2, \dots\dots\dots, N \qquad (16)$$

$$x_{ij} \in \{0,1\} \quad \forall\, i, j = 1, \dots\dots, N \qquad (17)$$

The variables with a tilde denote a fuzzy parameter and here we use a TIFN. In the above stated equations, the position of vertex v_i is denoted by the variable u_i and if vertex v_j is explored after v_i then x_{ij} = 1else it is 0.The restriction that for every path the beginning and the end point should be v_1 and v_N, every path remains connected without any vertex being visited more than once and the necessity of removing sub-tours is implemented by the crisp constraints (11), (12)-(13) and (15)-(16) respectively [1]. The two fuzzy goals (10) and (14) represent the necessary condition of maximizing the total reward or score collected and of the total time taken for traversing a path being within the specified upper limit respectively. In the fuzzy formulation, the symbols '≥' and '≤' indicating the 'greater than or equal to' and 'less than or equal to' of the crisp case are replaced by the symbols '≳' signifying the 'fuzzy greater than or equal to' and '≲'signifying the 'fuzzy less than or equal to' relation respectively. These symbols suggest that there is no strict boundaries for the constraints and the violation of the constraint to some extent is also acceptable but with differing degrees [24].

In this paper, we symbolize the two fuzzy goals of minimizing the total time taken by each path and maximizing the total score collected by each path by membership functions as shown in Figure 2 and Figure 1 of [26] respectively. As can be observed from the diagram below, the total score collected for a path should be either equal to S_{min} or greater than S_{min} in the ideal case which gives the most desirable solution but to consider the practical situations, we also accept solutions which have their total score within the range of S_{min} and $S_{min} - P$, each having a different degree of satisfaction. Similarly, for the constraint of satisfying the time budget we specify the limit T_{max} but also accept solutions up to $T_{max} + L$ with different degrees of acceptability. The fuzzy decision set Z and Z^* and the "best compromise solution" derived from the max-min formulation is shown in Figure 3 of [26].

V. IFOP ALGORITHM

Following are the steps to determine the most appropriate path for a given graph $\mathrm{G(V,E)}$ with N nodes. The steps are explained in the next section with the help of an illustrative example:

Step1: Compute all the paths $(\mathrm{P_m})$ that connect the source node $(\mathrm{v_1})$ and the destination node $(\mathrm{v_N})$ and fulfil the condition stated by (11), (12), (13), (15), (16), (17).

Step 2: The following values are calculated for each of the possible paths computed in Step1:

(A) The total collected reward or score and the total time taken (using Definition II(B) and (3)).

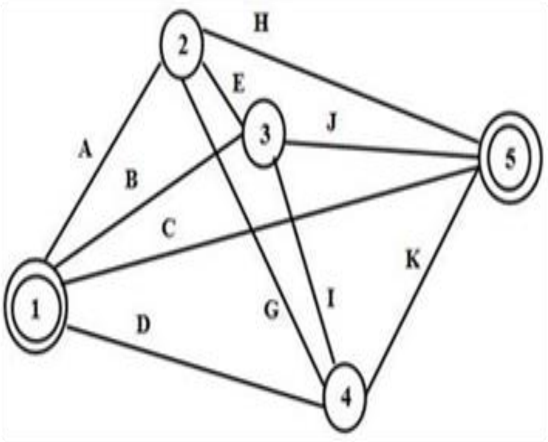
(B) The expected value for the total time taken and the total collected score (using Definition II(C) and (4)).

(C) The membership value for the total time taken and the total collected score represented by the fuzzy set $\mathrm{F_1}$and $\mathrm{F_2}$respectively (using (14), Fig. 2 of [26] and (10), Fig. 1 of [26]).

Step 3: Compute the set of feasible paths depicted by the fuzzy decision set Z (using Definition II (E) and (8)).

Step 4: The final solution representing the most desirable path is denoted by the fuzzy decision set $\mathrm{Z^*}$ (obtained using Definition II (E) and (9)). If the set $\mathrm{Z^*}$ contains more than one path then to conclude with the path that maximizes the total collected score, the paths in $\mathrm{Z^*}$ are ranked according to their total collected score (using (5), (6), (7)).

VI. LUSTRATIVE EXAMPLE



Node	Label	Intuitionistic Fuzzy Score Values $\langle(\mu(x)),(\nu(x))\rangle$
v_1	1	$\langle(1,2,8,9),(0,2,8,10)\rangle$
v_2	2	$\langle(8,9,11,12),(6,8,12,14)\rangle$
v_3	3	$\langle(3,5,9,11),(1,4,10,13)\rangle$
v_4	4	$\langle(17,20,24,27),(14,18,26,30)\rangle$
v_5	5	$\langle(1,2,4,5),(0,1,5,6)\rangle$
	$S_{min} = 25, P = 13$	

Edge	Label	Intuitionistic Fuzzy Time Values $\langle(\mu(x)),(\nu(x))\rangle$
e_{12}	A	$\langle(1,4,6,9),(0,2,8,10)\rangle$
e_{13}	B	$\langle(6,9,13,16),(5,8,14,17)\rangle$
e_{14}	D	$\langle(14,15,21,22),(12,14,22,24)\rangle$
e_{15}	C	$\langle(4,6,8,10),(2,4,10,12)\rangle$
e_{23}	E	$\langle(2,3,3,4),(0,3,3,6)\rangle$
e_{24}	G	$\langle(5,8,8,11),(4,8,8,12)\rangle$
e_{25}	H	$\langle(14,18,22,26),(12,16,24,28)\rangle$
e_{34}	I	$\langle(5,8,12,15),(3,6,14,17)\rangle$
e_{35}	J	$\langle(0,2,10,12),(0,1,11,12)\rangle$
e_{45}	K	$\langle(1,2,2,3),(0,2,2,4)\rangle$
	$T_{max} = 20, L = 15$	

Figure 2. The input graph with $\mathrm{N = 5, v_1 = 1, v_N = 5}$ and the time and score values associated with each edge and vertex respectively

$\mathbf{P_7}$: $1-2-4-5$; $\mathbf{P_8}$: $1-4-2-5$;
$\mathbf{P_9}$: $1-3-4-5$; $\mathbf{P_{10}}$: $1-4-3-5$;
$\mathbf{P_{11}}$: $1-2-3-4-5$; $\mathbf{P_{12}}$: $1-2-4-3-5$;
$\mathbf{P_{13}}$: $1-3-4-2-5$; $\mathbf{P_{14}}$: $1-4-3-2-5$;
$\mathbf{P_{15}}$: $1-4-2-3-5$; $\mathbf{P_{16}}$: $1-3-2-4-5$

Step 2: The actual values for the stated example obtained as a result of step 2 (A), (B), (C) of the IFOP algorithm are shown in Table II.

The fuzzy set $\mathrm{F_1}$ and $\mathrm{F_2}$ denoting the membership value for the total time taken and the total collected score respectively are as follows:

F_1
$= \{P_1/1, P_2/0.67, P_3/1, P_4/1, P_5/1, P_6/0.06, P_7/1, P_8/0,$
$P_9/0.8, P_{10}/0.06, P_{11}/1, P_{12}/0.4, P_{13}/0, P_{14}/0,$
$P_{15}/0, P_{16}/0.73\}$
$F_2 = \{P_1/0, P_2/0.23, P_3/0, P_4/1,$
$P_5/0.77, P_6/0.77, P_7/1,$

$P_8/1, P_9/1, P_{10}/1, P_{11}/1, P_{12}/1, P_{13}/1, P_{14}/1,$
$P_{15}/1, P_{16}/1\}$

Step 3: For the considered input, following is the fuzzy decision Z :

$Z = \{P_1/0, P_2/0.23, P_3/0, P_4/1, P_5/0.77, P_6/0.06,$
$P_7/1, P_8/0, \quad P_9/0.8, P_{10}/0.06, P_{11}/1,$
$P_{12}/0.4, P_{13}/0\}$

Step 4: Following are the paths in Z^* and their corresponding ranks obtained for the given network:

$Z^* = \{P_4/1, P_7/1, P_{11}/1\}$

Table 1. Ranks Of The Desirable Paths

Path	Score	Rank
P_4	$\langle(18,22,32,36),(14,20,34,40)\rangle$	3
P_7	$\langle(26,31,43,48),(20,28,46,54)\rangle$	2
$\mathbf{P_{11}}$	$\langle(\mathbf{29,36,52,59}),(\mathbf{21,32,56,67})\rangle$	**1**

The most desirable path is P_{11} as it has the highest rank. To check the correctness of our result, we set the spreads of the TIFN (for both score and time) to zero in order to convert each input to its crisp equivalent and then perform exhaustive search. This gives the same solution as our algorithm.

VII. WORK-DEPTH ANALYSIS OF IFOP

We present a parallel formulation here, to achieve a better performance and solve the IFOP more efficiently for large instances. Parallel computers can be organized in various ways and several multiprocessor models are known but it is difficult to conclude with one model that is apt for all machines. The method to deal with this situation is to focus on algorithms than on machines. As stated in [25], work-depth model is a technique of presenting the parallelism of an algorithm. For any algorithm, the following terms can be calculated [25]:

Work (W): Total number of operations performed.
Depth (D): Longest chain of dependencies among its operations.
Parallelism ($\mathcal{P}$): The ratio $\frac{W}{D}$

Algorithms with efficient work-depth models can be converted into efficient multiprocessor models and then to actual parallel computers. Work-depth models can be represented in three possible ways:

(a) Circuit Model.
(b) Vector Machine Model.
(c) Language-based Model.

For the parallel formulation of IFOP we use the circuit model which is the most abstract one when compared to the other two models. A circuit has two important components: nodes and directed arcs. The directed arcs and the nodes denote the flow of values and the operations to be performed respectively. Fan-in and fan-out are two terms associated with each node signifying the number of incoming and outgoing arcs respectively. The input to the circuit is provided through input arcs which do not originate from any node. Similarly, the output arcs carry the result out of the circuit and do not have any destination node. The number of nodes denotes the work of the circuit, also called the size of the circuit. A circuit should not contain directed cycles and the count of the nodes on the longest directed path connecting the input and output arc specifies the depth of the circuit. If the parallelism computed for the work-depth model is at least as large as the number of processors then it is said to be work-preserving and can be translated into an efficient multiprocessor model like the PRAM model [25].

For the circuit of IFOP we assume two things:

(a) Number of processors

$= \text{Number of distinct paths between source } (v_1) \text{ and destination } (v_N)$

(b) Number of distinct paths $(p) = \sum_{i=1}^{N-1} \frac{(N-2)!}{[N-(i+1)]!}$ where N is the number of nodes in the given graph G.

The first step of the IFOP of determining all the possible distinct paths is performed sequentially as shown in Figure 3:

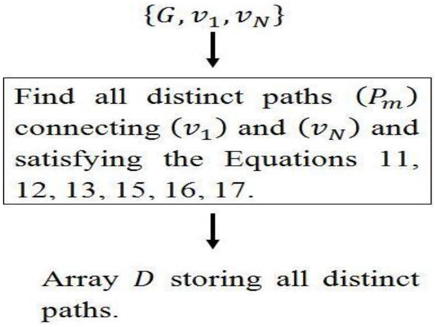

$\{G, v_1, v_N\}$
Find all distinct paths (P_m) connecting (v_1) and (v_N) and satisfying the Equations 11, 12, 13, 15, 16, 17.
Array D storing all distinct paths.

Figure 3. The sequential module executing Step 1 of IFOP that computes all the distinct paths in the given graph G.

The circuit of IFOP is shown in Fig. 4. The parallelism for IFOP is calculated below:

Work (W) = $5p + 2$.

Depth (D) = 5.

Parallelism ($\mathcal{P}$) = $\frac{W}{D} = \frac{5p+2}{5} = \left(p + \frac{2}{5}\right)$.

Therefore, IFOP is work-preserving because the parallelism ($\mathcal{P}$) is$\left(p + \frac{2}{5}\right)$ which is at least as large as the number of processors i.e. p.

VIII. CONCLUSION

In this paper, we considered the orienteering problem which is a NP-Hard problem and formulated the intuitionistic fuzzy version of this problem accounting for uncertainty in the real life areas where this problem finds its application like the tourism industry, logistics etc. We state the problem as a fuzzy integer program with fuzzy goals and crisp constraints and present a fuzzy optimization technique for solving IFOP. The method suggested here considers the hesitancy, aspiration levels, degree of acceptability and satisfaction of the decision maker, thus providing latitude to the solution process. The generated solution is capable of tackling the uncertainty and vagueness involved in the two parameters score and time. To deal with larger instances efficiently, we presented the work-depth analysis of IFOP and showed that the algorithm is work-preserving and thus can be efficiently implemented on a multiprocessor model like the PRAM.

IX. REFERENCES

[1]. P. Vansteenwegen, W. Souffriau, D. V. Oudheusden, "The orienteering problem: A survey", European Journal of Operational Research, vol. 209, pp. 1-10, 2011.

[2]. Jun Ye, "Expected value method for intuitionistic trapezoidal fuzzy multicriteria decision-making problems", Expert Systems with Applications, vol. 38, pp. 11730-11734, 2011.

[3]. T. Tsiligirides, "Heuristic methods applied to orienteering", Journal of the Operational Research Society, vol. 35, pp. 797–809, 1984.

[4]. B. Golden, L. Levy, R. Vohra, "The orienteering problem", Naval Research Logistics, vol. 34, pp. 307–318, 1987.

[5]. R. Ramesh, K. Brown, "An efficient four-phase heuristic for the generalized orienteering problem", Computers and Operations Research, vol. 18, pp. 151–165, 1991.

[6]. Q. Wang, X. Sun, B. Golden, J. Jia, "Using artificial neural networks to solve the orienteering problem", Annals of Operations Research, vol. 61, pp. 111–120, 1995.

[7]. I. Chao, B. Golden, E. Wasil, "Theory and methodology – a fast and effective heuristic for

the orienteering problem", European Journal of Operational Research, vol. 88, pp. 475–489, 1996.

[8]. M. Tasgetiren, "A genetic algorithm with an adaptive penalty function for the orienteering problem", Journal of Economic and Social Research, vol. 4, no. 2, pp. 1–26, 2001.

[9]. M. Gendreau, G. Laporte, F. Semet, " A tabu search heuristic for the undirected selective travelling salesman problem" European Journal of Operational Research, vol. 106, pp. 539–545, 1998a.

[10]. Y. Liang, S. Kulturel-Konak, A. Smith, "Meta heuristics for the orienteering problem", Proceedings of the 2002 Congress on Evolutionary Computation, Hawaii, Honolulu, pp. 384–389, 2002.

[11]. M. Fischetti, J. Salazar, P. Toth, "Solving the orienteering problem through branch-and-cut", INFORMS Journal on Computing, vol. 10, pp. 133–148, 1998.

[12]. M. Schilde, K. F. Doerner, R. F. Hartl, G. Kiechle, "Metaheuristics for the bi-objective orienteering problem", Swarm Intelligence, vol. 3, pp. 179-201, 2009.

[13]. V. Campos, R. Marti, J. Sanchez-Oro, A. Duarte, "GRASP with Path Relinking for the Orienteering Problem", 2013. http://www.uv.es/rmarti/paper/docs/routing7.pdf

[14]. A. Blum, S. Chawla, D. R. Karger, T. Lane, A. Meyerson, M. Minkoff, "Approximation Algorithms for Orienteering and Discounted-Reward TSP", Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS'03), pp. 1-10, 2003.

[15]. D. Johnson, M. Minkoff, S. Phillips, "The prize collecting steiner tree problem: Theory and practice", Proceedings of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms, pp.760–769, 2000.

[16]. F. V. Fomin, A. Lingas, "Approximation algorithms for time-dependent orienteering", Information Processing Letters, vol. 83, pp. 57–62, 2002.

[17]. K. Atanassov, "Intuitionistic fuzzy sets", Fuzzy Sets and Systems, vol. 20, pp. 87-96, 1986.

[18]. P. Grzegorzewski, "Distances and orderings in a family of intuitionistic fuzzy numbers", Proceedings of the 3rd Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT, 2003), Germany, pp. 223-227, September 10-12, 2003.

[19]. M. Jimenez, M. Arenas, A. Bilbao, M. V. Rodriguez, "Linear programming with fuzzy parameters: An interactive method resolution", European Journal of Operational Research, vol. 177, pp. 1599–1609, 2007.

[20]. D. P. Loucks, E. V. Beek, "Water Resources Systems Planning and Management: An Introduction to Methods", Models and Applications, UNESCO Publishing, pp. 135-144, 2005.

[21]. U. Kaymak, J.M. Sousa, "Weighted Constraints in Fuzzy Optimization, Publications in the Report Series Research in Management" , ERIM Research Program: Business Processes, Logistics and Information Systems, pp. 1-21, 2001. http://repub.eur.nl/res/pub/85/erimrs20010403174009.pdf

[22]. J. Ramik, "Soft Computing: Overview and Recent Developments in Fuzzy Optimization", 2001. http://irafm.osu.cz/research_report/118_softco01.pdf

[23]. H. J. Zimmermann, "Fuzzy set theory", WIREs Computational Statistics, vol. 2, pp. 317-332, 2010.

[24]. F. JARRAY, "Discrete Tomography and Fuzzy Integer Programming", Iranian Journal of Fuzzy Systems, vol. 8, pp. 41-48, 2011.

[25]. G. E. Blelloch, "Programming Parallel Algorithms ", Communications of the ACM, vol. 39, no. 3, pp. 85-97, 1996.

[26]. M. Verma, K. K. Shukla, "Application of Fuzzy Optimization to the Orienteering Problem", Advances in Fuzzy Systems, vol. 2015, pp. 1-12, 2015.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Sequence Labeling for Three Word Disambiguation in Telugu Language Sentences

Jinka Sreedhar*[1], Suresh Dara[1], Baijnath Kaushik[2], SK Althaf Hussain Basha[3]

[1]Department of Computer Science and Engineering B.V.Raju Institute of Technology, Narsapur, Telangana, India
[2]Department of Computer Science and Engineering Shri Mata Vaishno Devi University, Katra, Jammu and Kashmir, India
[3]Department of Computer Science and Engineering Gokaraju Rangaraju Institute of Engineering and Technoloy, Hyderabad, India

ABSTRACT

This paper is intended to apply sequence labelings which are introduced to find out the ambiguity in three-words. These words appear to give rise to ambiguity. They seem to be sequence words and this method can be applied only to these types of words. There is another theory of automata which is a mathematical model. By implementing this model to disambiguate the words of sequence it is found there is a kind of mathematical accuracy equal to that of sequence labeling method. The main aim of finding out these methods, is to find out solution to the problem of ambiguity in three-words sequences. Here designing of automata theory for three-words is dealt with the Three-Words disambiguation rules are explained with examples.

Keywords : Natural Language Processing(NLP),Information Retrieval Systems(IRS), Machine Translation(MT), Finite Automata(FA).

I. INTRODUCTION

To explain this theory clearly, five states have been identified. In this process state one may have more than one tag, state two may have more than one tag and state three may have more than one tag. Now one tag has been retained in the word one deleting the remaining tags. In the similar manner, the same procedure is continued in the second word order and third word also[1,2,3,4]. By doing so, the problem of ambiguity has been resolved. When this process comes to state four, it is treated as completed since it gives a complete sense. In the fifth state regarded as a dead state, all the unwanted tags will be appear[5,6,7,8].

With the help of transitional diagrams and transitional tables, the rules are explained. Transitional diagrams contain states, POS tags, start state and final state[5]. These diagrams can be represented with the symbols like Q, ∑, S, F. Here Q stands for states one, two, three and dead states, ∑ contains POS tags, S contains the starting state, that is, state one, and F contains the final state, that is, state four[13,14,15,16,17,18].

Transitional Table is also framed to show how these tags appear in different states and give a picture representation.

W1 :: W2 :: w3 => W1 :: W2 :: W3

Where

W1 ,W2 and W3 are sequence of words in that order.

II. DESIGNING AUTOMATA THEORY FOR THREE WORDS RULES

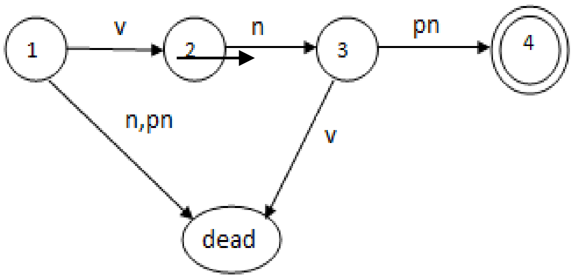


Figure 2.1. Three-word disambiguation for v, n, pn :: n :: pn,v

Where

1, 2, 3 and dead state belong to Q and n, v, pn belongs to ∑.

Here v denotes verb, pn denotes pronoun and n denotes noun.

Q: {1,2,3,4,dead}

∑:{v,n,pn}

S:{1}

F:{4}

Table 2.1. Three-word disambiguation for v, n, pn :: n :: pn,v

∂	v	n	Pn
1	2	dead	Dead
2	-	3	-
3	dead	-	4
4	-	-	-
dead	-	-	-

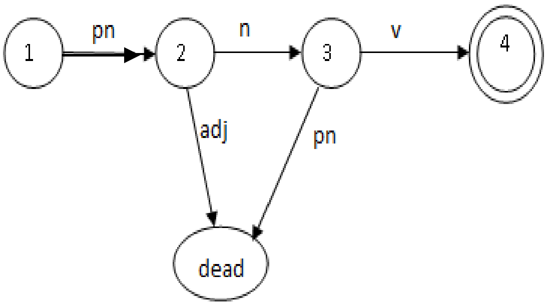


Figure 2.2. Three-word disambiguation for pn :: n, adj :: v, pn

Where

1, 2, 3 and dead state belong to Q and n, v, pn, adj belongs to ∑.

Here v denotes verb, pn denotes pronoun,n denotes noun and adj denotes adjective.

Q:{1,2,3,4,dead}

∑: {pn,n,v,adj}

S: {1}

F: {4}

Table 2.2. Three-word disambiguation for pn :: n, adj :: v, pn

∂	pn	n	V	adj
1	2	-	-	-
2	-	3	-	dead
3	dead	-	4	-
4	-	-	-	-
dead	-	-	-	-

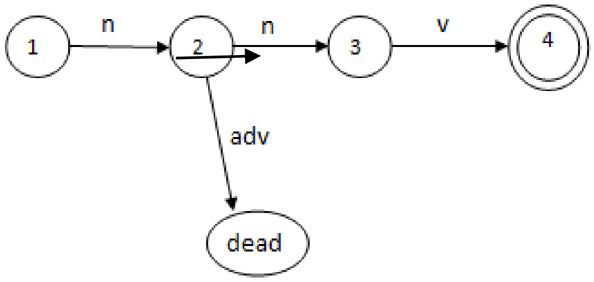


Figure 2.3. Three-word disambiguation for n :: n, adv :: v

Where 1, 2, 3 and dead state belong to Q and n, v, adv belongs to ∑.

Here v denotes verb, n denotes noun and adv denotes adverb.

Q: {1,2,3,4,dead}

∑: {n,v,adv}

S: {1}

F: {4}

Table 2.3. Three-word disambiguation for n :: n, adv :: v

∂	n	v	adv
1	2	-	-
2	3	-	dead
3	-	4	-
4	-	-	-
dead	-	-	-

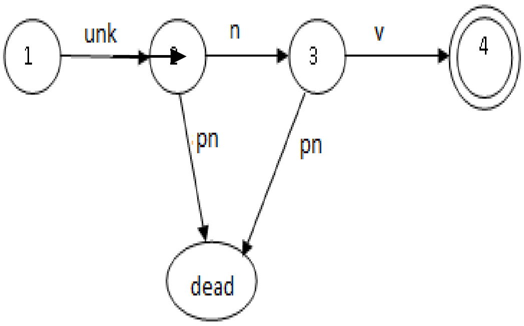


Figure 2.4. Three-word disambiguation for unk :: n, pn :: v, pn

Where

1, 2, 3 and dead state belong to Q and unk, v, pn, nbelongs to ∑.

Here v denotes verb, pn denotes pronoun, n denotes noun and unk denotes unknown.

Q: {1,2,3,4,dead}

∑:{unk,n,v,pn}

S:{1}

F: {4}

Table 2.4. Three-word disambiguation for unk :: n, pn :: v, pn

∂	unk	n	v	pn
1	2	-	-	-
2	-	3	-	dead
3	-	-	4	dead
4	-	-	-	-
dead	-	-	-	-

III. THREE WORD DISAMBIGUATION RULES

W1::W2::W3 ⇒ w1::w2::w3…..(1)
N,v,pn :: n::pn,v ⇒ v::n::pn……….(2)
Pn::n,adj::pn,v ⇒ pn::n::v…..(3)
N::n,adv::v ⇒ n:n:v …. ….(4)
Unk::n,pn::v,pn ⇒ unk::n::v….(5)
n::n,v::v,pn ⇒ n::n::v ….(6)
n::v,pn::n,adv ⇒ n::v::n.….(7)
v,pn::n:pn,v ⇒ v::n::v.(8)
n::v,n::v,pn ⇒ n::n::v……..(9)
n,v:avy::v,pn,adj ⇒ n::avy:v……….(10)
unk::n,adj::v,pn ⇒ unk::n::n…..(11)
pn::v,np::v,pn ⇒ pn::v::v…...(12)
v,p,n,n::v,pn,n,adj::v,pn ⇒ pn::v::v…...(13)
avy::n,adv::v,pn ⇒ avy::n::pn...(14)
n,adj::n::v,pn,n ⇒ n::n::v...…..(15)
n::n,adv::v,pn ⇒ n::n::v..……..(16)
n,adv::adv::v,pn ⇒ n::adv::p…...(17)
v,pn,n::v,pn::avy ⇒ n::pn::avy…….…(18)
adv,n::n,adj::v,pn ⇒ adv::adj::v………..(19)
punc::v,pn,n,adj::v,p ⇒ punc::adj::v…..…...(20)

3.1 Case Study for three word ambiguity

Here is a Telugu sentence which has ambiguous words from Telugu corpus, like

Sentence:

waMdri ceVppina viRayAlu AlociMcevAdu.

Morph Output:

waMdri	waMdri/n
ceVppina	ceVppu/n,v,pn
viRayAlu	viRayaM/n
AlociMcevAdu	AlociMcu/pn,v

Before Applying Disambiguation Rule:

W1 = ceVppu
W2 = viRayaM
W3 = AlociMcu

w1 :: w2 :: w3 => w1 :: w2 :: w3
n,v,pn :: n :: pn,v => v :: n :: pn

In the above sentence, the first word carries tags (n,v,pn) followed by the second word carrying the tag n and followed by a third word carrying the tags (pn,v). Then the tag v is retained from the first word and pn is retained from the third word eliminating the (n,pn) from (n,v,pn), and v from (pn, v).

After Applying Disambiguation Rule:

waMdri ceVppina viRayAlu AlociMcevAdu.
n v n pn punc

3.2 Analysis of Three Word Disambiguation

Here the following figure 3.1 gives an analysis of the Accuracy. While X-axis indicates the number of test sessions, Y-axis indicates the Accuracy. As a result, the proposed method can disambiguate nearly 96% of the ambiguity [9,10,11,12].

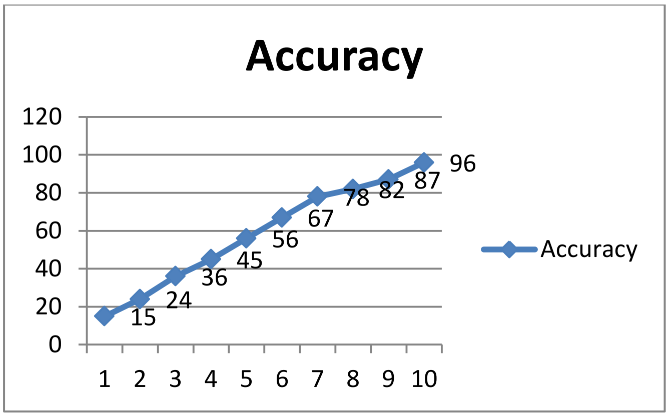


Figure 3.1. Three word disambiguation rules accuracy

IV. CONCLUSION

Here dealing with the designing of three-word rules for Telugu Language Sentence word order. To make things easy to understand some rules have been made which can be applied for the word order of Telugu Language Sentences and it clarifies the ambiguity. All these things are so vividly explained with the help of case studies and theoretical explanations. When these rules are applied whenever needed, they help the user easily to eliminate the ambiguity. These theories help understand the study of disambiguation. By applying disambiguation rules it is found that the proposed method can disambiguate nearly 96% of the ambiguity. The theoretical explanation and disambiguation rules have resulted in the accuracy of evidences.

V. REFERENCES

[1]. Noam Chomsky, "Logical syntax and semantics: their linguistic relevance", vol.31, No.1, pp: 36-45, 1955.

[2]. Noam Chomsky, "On Certain Formal Properties of Grammars", Information and Control, Vol. 9, pp: 137-167, 1959.

[3]. Nou, Chenda and WataruKameyama,Khmer, "POS Tagger: A Transformation-based Approach with Hybrid Unknown Word Handling", Proceedings of the First IEEE International Conference on Semantic Computing (ISCS), Irvine, CA. pp: 482-492, 2007.

[4]. PawanGoyal,LaxmidharBehera,Thomas Martin McGinnity, "Query Representation through Lexical Association for Information Retrieval", IEEE Transactions on Knowledge and Data Engineering, pp: 2260-2273, 2012.

[5]. PengJin,XingyuanChen,"A Word Sense Probabilistic Topic Model", 9th International Conference on Computational Intelligence and Security (CIS), pp: 401-404, 2013.

[6]. PengYuan Liu, "Another View of the Features in Supervised Chinese Word Sense Disambiguation", 9thInternational Conference on Computational Intelligence and Security, ISBN: 978-0-7695-4584-4, pp: 1290-1293, 2011.

[7]. Pengyuan Liu, YongzengXue, Shiqi Li, Shui Liu, "Minimum Normalized Google Distance for Unsupervised Multilingual Chinese-English Word Sense Disambiguation", International Conference on Genetic and Evolutionary Computing, ISBN: 978-0-7695-4281-2, 2010.

[8]. Ping Chen,BowesC,Wei Ding, et..al, "Word Sense Disambiguation with Automatically Acquired Knowledge",IEEE INTELLIGENT SYSTEMS, 2012.

[9]. PrashanthMannem, "Bidirectional Dependency Parser for Hindi, Telugu and Bangla", Proceedings of ICON09 NLP Tools Contest: Indian Language Dependency Parsing, India, 2009.

[10]. Quinlan, J. R, "Induction of decision trees", Mach. Learn. 1, 1, pp: 81–106, 1986.

[11]. Quinlan, J. R, "Programs for Machine Learning", Morgan Kaufmann, San Francisco, CA, 1993.

[12]. R.M.K.Sinha and K. Sivaraman, "Ambiguity Resolution in Anglabharati",TRCS-93-174, Department of Computer Science and Engineering, IIT,Kanpur, India, 1993

[13]. R. Mahesh K. Sinha,"Learning Disambiguation of Hindi Morpheme 'vaalaa' with a Sparse Corpus" 4th International Conference on Machine Learning and Applications, ISBN: 978-0-7695-3926-3, pp: 653-657, 2009.

[14]. J.Sreedhar, S. Viswanadha Raju, A. Vinaya Babu, Amzan Shaik, P.Pavan Kumar "Word Sense Disambiguation : An Empirical Survey" International Journal of Soft Computing and Engineering(IJSCE), Volume-2,Issue-2,May-2012,ISSN:2231-2307.

[15]. J.Sreedhar, S. Viswanadha Raju, A. Vinaya Babu, Amzan Shaik, P.Pavan Kumar"A critical Approaches to Identification of Disambiguation Words in NLP : A Current State of the Art" International Journal of Engineering Trends and Technologies (IJETT), Volume-3,Issue-3,May-2012,ISSN:2231-5381.

[16]. P.Pavan Kumar, J.Sreedhar "Innovative Techniques and Technologies in Translation in a Multilingual Context" 3rd International Conference on Translation Technology and Globalization in Multilingual Context" Delhi, June 23-26, 2012.

[17]. P.Pavan Kumar, J.Sreedhar "Language teaching and MLE in the context of the third revolution" DLA Conference on Dravidian Languages and Translation Technology" HCU, Hyderabad, June 18-20, 2012.

[18]. Hyuk-Chul Kwon, Minho Kim, Youngim Jung, "Hybrid word sense disambiguation using language resources for transliteration of Arabic numerals in Korean", International Conference on Hybrid Information Technology (ICHIT '09), ISBN: 978-1-60558-662-5, pp: 314-321,2009.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Optimization of Worst-Case Execution Time for ASIP using Genetic Algorithm

Mood Venkanna*[1], Rameshwar Rao[2], P. Chandra Sekhar[3]
[1]Department of ECE, UCE, Osmania University, Hyderabad, India
[2]Former VC JNTUH, Department of ECE, UCE, Osmania University, Hyderabad, India
[3]Professor, Department of ECE, UCE, Osmania University, Hyderabad, India

ABSTRACT

The use of Application Specific Processor is available almost in all the areas. Research and developments on ASIP has been progressed since last two decades. However, the minute analysis of these processors is still a great challenge for the engineers and researchers in current scenario. Embedded processor application spread over different areas with a high desire of fast and accurate execution. It requires enhancing the execution time of the processor. The worst-case execution time (WCET) evaluation satisfies the desire of user end along with the hardware and software application of the processor. As a result the reconfiguration of processor architecture can be modified and perfect task scheduling can be performed. For WCET, upper bound on execution time is to be focused. An attempt is made to optimize the WCET to enhance the performance of the processor along with less occupation of space. Genetic Algorithm (GA) as the popular optimization technique is utilized to optimize that can help to the reconfigurable processor performance and also the control flow of the instruction to the processors.

Keywords : Embedded processor, ASIP, Optimization, Genetic Algorithm, WCET.

I. INTRODUCTION

An embedded system relies on Application-specific instruction set processor (ASIP) design to meet its desired performance and cost effectiveness. Additionally, these processors found useful in cellular phones, avionics, automobile control systems etc. in which a slight change in performance or cost may impact drastically the productivity. There are different ASIP designs available such as Co Ware Lisa Tek products, TensilicaXtensa processor, Target Compiler Technologies products and Xilinx Micro Blaze. These ASIPs remain sole solution for the physical constraints and for the desired functions due to programmability and high flexibility. This paper will provide a brief explanation how the ASIP systems can be improved further.

Program Execution time is a crucial component in any real-time system as it may result in catastrophically consequences in case the deadline is missed. The worst execution time measurements using the worst possible program input remains unreliable today. Further, a program's worst execution path may not be not captured during the measurements. A number of researches in WCET (Worst Case Execution Time) analysis and theoretically estimate of WCET models has been conducted during last few years (Asavoae,M.et.al., 2013, Cl´ementBallabriga et al., 2010, Banerjee,A., et.al. 2013, Armin Biere et al., 2013) . However, application of actual real-time operating system code models has been least considered.

In embedded systems, WCET of a program need to be less than a specific threshold particular important in case of synchronous active control loops (Bjørner,N. Dutertre,B. and Moura, L., 2008). For a program, the WCET is computed as a combination of low-level, micro architectural reasoning. This involves pipeline, busses, cache states, cycle-accurate timing as well as higher-level reasoning such as the loop counts, program control flow and the variable pointers. It requires application of abstract interpretation with respect to the micro architecture, deduce elementary block's worst case timings and reassemble to global WCET using the control flow, maximal iteration counts by means of integer linear programming (Cadar, C. et. al, 2008).

In this paper, static method is optimized using genetic algorithm. The control flow and the path analysis is the major focus for optimized WCET. Rest of the paper is organized as the subsections for technique of WCET that follows the optimization method. Next to it the result has been explained. Finally it concludes this piece of wok.

II. TECNIQUES OF WCET AND THE MODEL

WCET has been used in many real-time systems due to safety, reliability, and surfacing of software in automotive systems. It serves as an input to schedule ability analysis in system design. Few of the automated approaches for WCET computation includes:

- ✓ Analytical techniques for test cases that boost the confidence for end to end measurements
- ✓ Static analysis of the software.
- ✓ combined or hybrid approaches that include both measurements and structural analysis
- ✓ Worst-case path determination
  a. Maps control flow graph to an integer linear program
  b. Determines upper bound and associated path

For accurate WCET computation the possible program flow involving function calls and loop iterations including their effects corresponding to hardware features need to be known.

WCET calculation:

For program average-case execution time improvement modern processors contains features such as order execution and cache hierarchies (Chattopadhyay, S., and Roychoudhury, A., 2013, Chu, D., and Jaffar, J., 2011, Wilhelm, R., 2006, Reinhard Wilhelm et al., 2008). Nevertheless, for atight WCET, these features make the system complex. Addition of more complex architectures in model checker increases the number of states which makes the track more prone to state explosion problems. However, availability of sophisticated tools like as Chronos cangues a better running of WCET in single core processors (Chaki, S., and Ivers, J., 2010). On the other hand, Multi core systems posses an additional complexity because of the shared resources or shared memories. Use of shared memory makes problems to obtain tight WCET.

Finding a method for WCET involve approximations thus, the exact WCET can be regarded as unachievable (Wankang Zhao et al., 2006, Kim, S.K. et.al.,1996, White,R., et.al.,1997, Colin, A. et.al., 2000)]. Finding the WCET are based on estimates which may be pessimistic. In such cases, the estimated WCET believes to be higher than the corresponding real or desired WCET. Hence, mostly in WCET analysis an attempt is made to reduce the pessimism with a low enough estimated value that can be of real interest to the system designer.

Static methods takes the task code in hand and do not depend on the executing code involving on a simulators or hardwares. Together with some annotations, the method analyzes the possible control flow through the given task, attempts to combine the control flow with hardware architecture abstract model so as to obtain the desired upper bounds.

Control-Flow Analysis

The control flow analysis is finite and aims to accumulate information on possible execution paths. Any superset can be considered as a safe approximation since the exact paths are not possibly determined. The analysis is difficult on machine code in comparison to the source code than as it is cumbersome to map the machine-code program results because of compilation, change of code optimization and linking in the structure. The basic concept of flow graph has been shown in Figure 1.

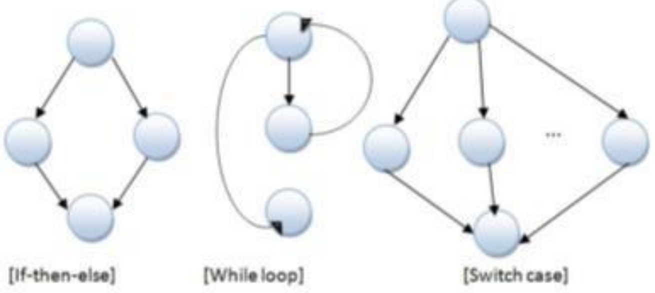


Figure 1. The Basic Concept of Flow Graph

It aims to estimate the WCET in dynamic approaches which may be underestimated since a subset of entire executions has been used for estimation. In static approaches also known as Bound Calculation an upper bound is computed for the entire execution times of task relying on the previous phase flow and timing information. There are three major classes such as the path-based, the structure-based, implicit path enumeration (IPET) based on analytically determination of end-to-end estimate times.

The structure-based approach cannot express every control flow through the syntax tree thus, assumes a straightforward relationship between the source and target program structures. Further, it is not feasible to incorporate additional flow information unlike the IPET. On the other hand, IPET can handle different flow information. It uses constraint programming or integer linear programming techniques and in this flow facts are converted to constraints whose size grows with the number of flow facts.

Let CI be the set of all CIs. We assume a specific configuration j of a CI k ∈CI in hardware to have a constant delay $t_{k,\ j}$ to require area on the reconfigurable fabric $a_{k,\ j}$∈[1, A], and to take a constant reconfiguration delay $r_{k\ j}$ for configuring it on the fabric. For a constant reconfiguration delay, a constant bandwidth for transferring configuration data to the reconfigurable fabric's configuration memory needs to be guaranteed. We assume the CPU to be delayed during reconfiguration in this work, and therefore the system bus could be utilized for reconfiguration at a guaranteed bandwidth. Along with hardware configurations, a CI can be implemented using its original software code j = 0. Since it has been implemented with a software, it does not have a constant delay $t_{k,0,}$ because of specific cache and pipeline analysis. (i.e., $a_{k,0}$= $r_{k,0}$ = 0).

In order to provide flexibility to execute the original software for generated CIs, we introduce CI super blocks. The CI superblocks begin with a conditional branch before every CI (the actual instruction in the binary) which jumps to the functionally equivalent software code when the CI is not implemented in hardware. If a configuration for the CI is available on the reconfigurable fabric, then it is executed instead of jumping to the software. The CI super block ends by joining paths of hardware CI and software. Multiple CI superblocks in the binary can execute the same CI k. Let B be the set of all blocks, that is, basic blocks (not contained in super blocks) as well as super blocks. The function ci(i) determines which CI k is executed by a super block i ∈B, that is, ci: B → CI ∪ {0}, i → k, with ci(i) = 0 ∈CI if i is a basic block

To ensure that exactly one implementation is chosen potentially in software (j= 0)or hardware (j >0) with m_k being the number of hardware configurations of CI $_k$. To only allow solutions that do fit onto the reconfigurable fabric, we introduce the area constraint is the sum of area on the reconfigurable fabric $a_{k,\ j}$ required to implement all CIs using the selected implementation j (for which $y_{k,\ j}$ = 1) needs to be lower than or equal to the total fabric area A. For a program with a count of *N* basic blocks, the objective function is given as

$$max \sum_{i=1}^{N} cix \quad (1)$$
$$i$$

Selecting an instruction set to optimize the WCET bound essentially means. we aim to minimize the WCET over all possible selections, that is, we aim to minimize the maximum execution time.

We extend the ILP formulation of IPET for capturing the implementation alternatives of a CI *k* ∈CI. We introduce new variables $y_{k,\ j}$ ∈ {0, 1} for every implementation *j* with $y_{k,\ j}$= 1 if CI *k* is implemented using alternative *j* and $y_{k,\ j}$= 0 otherwise

$$\sum 0<j<n \ y_{k,}=1 \quad (2)$$

The total cycle contribution of CI *k*'s super block *i* to the WCET bound is given as follows:

$$\sum^{|B|}_{\substack{1\leq i \leq|B| \\ 0<j<me(i)}} e_{i,j} y_{c(i),j'} x_i \quad (3)$$

The WCET for a given selection *y* without accounting for reconfiguration delay can be determined as follows:

$$\text{WCET(y)=max } x \left(\sum^{|B|}_{\substack{i=1 \\ Ci(i)\in Cj}} c_i x_i + \sum^{|B|}_{\substack{1\leq i \leq|B| \\ 0<j<me(i)}} e_{i,j} y_{c(i),j'} x_i \ ) \right) \quad (4)$$

Every CI utilized in a kernel is configured exactly once before entering the kernel (with zero reconfiguration delay for software implementation). Therefore, we obtain the WCET including reconfiguration delay as:

$$\text{WCET}(y) = \text{WCET}'(y) + \sum_{\substack{0\leq i \leq m \\ 0<j<n}} y_{k,j} \ r_{k,j} \quad (5)$$

III. OPTIMIZATION USING GA

Genetic Algorithm is a population-based search method in which the candidate solutions are termed as chromosomes, and the solution is termed as genes in the chromosomes. A search space has been formed using possible chromosomes. These are involved with corresponding fitness function that represents solutions encoded in the corresponding chromosome. The search continues by computing the fitness of a population of chromosomes followed by mutations and recombination with respect to successful chromosomes. The GA execution starts with a set of random initial population which are sampled for a particular task. The process of selection, crossover and mutation are applied on the initial population to get a new and better generation.

The basic Genetic Algorithm:

[Start]: random population of n chromosomes is generated that gives suitable solutions for the task.

[Fitness]: The fitness (*x*)with respect to each chromosome *x* in the population is evaluated.

[New population]: A new population is created using the below steps and repeating them till completion of the new population.

[Selection]: Two parent chromosomes are selected from a population as per their fitness).

[Crossover]: Cross over the parents with across over probability form a new offspring (children). In case of no crossover the offspring is an exact replica of parents.

[Mutation] : Mutate new offspring with a mutation probability at each locus that gives the position in chromosome.

[Accepting]: The new offspring in a new population is placed.

[Replace]: The new generated population is used for a algorithm to be run further.

[Test]: When the end condition is satisfactory, stop, and return to the best solution with respect to the current population.

[Loop]: Go to fitness step.

The three basic steps for Genetic Algorithm, as shown above, are:

1. Selection: In selection (also known as reproduction), the chromosomes from the population to be parents are selected to cross over and produce offspring.

The various methods for parents to cross over are:

I. Roulette-wheel selection
II. Boltzmann selection
III. Tournament selection
IV. Rank selection
V. Steady-state selection

2. Cross over: The off springs are enriched with suitable individuals after the selection phase. Cross over process is continued to the mating pool and expected to create a better string. It also has three steps; firstly, the reproduction stage selects randomly a pair of two individual strings for mating. Secondly, a random cross-site is selected along the string length and at last their position values are swapped between those two strings. Different cross over types are:

I. Single-site cross over
II. Two-point cross over
III. Multi-point cross over
IV. Uniform cross over
V. Matrix cross over

3. Mutation: The strings are mutated once the cross over process is completed. It involves flipping of bits between 0 to 1 and vice versa using a small mutation probability Pm. A number is chosen between 0 to 1 randomly and the bit is changed if the number is less than Pm, otherwise it is unaltered.

Generating optimal test data using GA based on fitness function: On the basis of basis paths, the developed system automatically generates the optimal test data in the CFG. The WCET analysis tool architecture is shown in Figure 1.

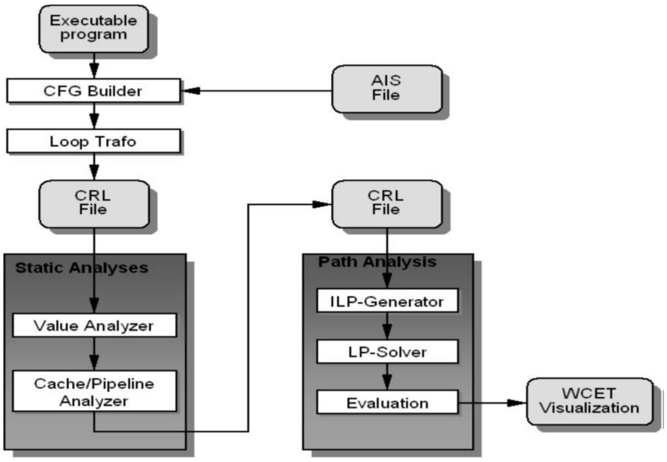


Figure 2. WCET Analysis Tool Architecture

IV. RESULTS AND DISCUSSION

The processor is reconfigured with the optimization. The parameter for genetic algorithm is given in Table-I and related convergence for hardware and software is shown in Figure 3.

Table 1. The parameter for genetic algorithm

Parameter	Value
Generations	350
Population Size	200
Chromosome Length	300
Selection mechanism Tournament	size=2
Crossover	0.85 (fixed point)
Mutation	0.02

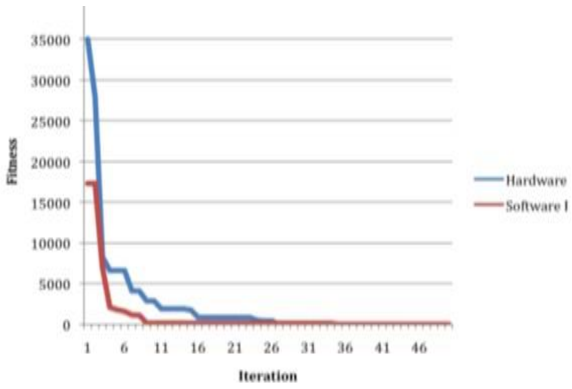


Figure 3. The Fitness for the Specific Function

V. CONCLUSION

It is observed that using satisfy ability modulo theory (SMT), the optimization has been a feasible approach in case of the bounding the WCET of the corresponding loop-free programs i.e., the programs in which the loops may be unrolled. To best of our knowledge, such an approach has been applied successfully for the first time. In all these levels we propose an evolutionary algorithm as the optimization engine, which is helped by other applications, either in a closed loop, either in off-line phases. The development of the computer-aided design or the CAD and compilation tools is one of the major challenges for mapping any application into an effective reconfigurable computing system. This desires the determination of application parts for mapping into the fabric and into the processor. Time of determination and its frequency of mapping into the reconfigurable fabric need to be emphasized in future.

VI. REFERENCES

[1]. Asavoae, M., Maiza,C. Raymond, P., 2013, Program Semantics in Model-Based WCET Analysis: A State of the Art Perspective. WCET Ed. by Claire Maiza. OASICS. 30, 32–41.

[2]. ClementBallabriga et al., 2010, OTAWA: An Open Toolbox for Adaptive WCET Analysis. SEUS. LNCS. Springer, 6399, 35–46.

[3]. Banerjee,A., Chattopadhyay,S. and Roychoudhury,A., 2013, Precise micro architectural modeling for WCET analysis via AI+SAT. IEEE Real-Time and Embedded Technology and Applications Sym- posium (RTAS), IEEE Computer Society, 87–96.

[4]. Armin Biere et al., 2013, The Auspicious Couple: Symbolic Execution and WCET Analysis. WCET, OASIcs. IBFI Schloss Dagstuhl, 30. 53–63. http://drops.dagstuhl.de/opus/volltexte/2013/4122.

[5]. Bjorner,N. Dutertre,B. and Moura, L., 2008, Accelerating lemma learning using joins - DPLL(⊔). Appeared as short paper in LPAR 2008, outside of proceedings.

[6]. Cadar, C. and Sen, K.., 2013, Symbolic Execution for Software Testing: Three Decades Later. Commun. ACM 56.2, 82–90.

[7]. Caspi,P., Raymond P., and Tripakis, S., 2008, Synchronous Programming. Handbook of Real-Time and Embedded Systems. Chapman & Hall / CRC, Chap. 14.

[8]. Chaki, S., and Ivers, J., 2010, Software model checking without source code. English. Innovations in Systems and Software Engineering 6.3, 233–242. ISSN: 1614-5046. doi: 10.1007/s11334-010- 0125-0.

[9]. Chattopadhyay, S., and Roychoudhury, A., 2013 Scalable and precise refinement of cache timing analysis via path-sensitive verification. Real-Time Systems 49.4, 517–562.

[10]. Chu, D., and Jaffar, J., 2011, Symbolic simulation on complicated loops for WCET Path Analysis.EMSOFT. 319–328. ISBN: 978-1-4503-0714-7. doi: 10.1145/2038642.2038692.

[11]. Wilhelm, R., 2006, Determining Bounds on Execution Times. Handbook on Embedded Systems. CRC Press, Chap. 14.

[12]. Reinhard Wilhelm et al., 2008, The worst-case execution-time problem - overview of methods

and survey of tools. ACM Trans. Embedded Comput. Syst.7.3.

[13]. Wankang Zhao et al., 2006, Improving WCET by applying worst-case path optimizations. Real- Time Systems 34.2, 129–152.

[14]. Kim,S.K., Min, S. L. and Ha,R., 1996, Efficient Worst Case Timing Analysis of Data Caching.IEEE Real-Time Technology and Applications Symposium (RTAS'96). 230–240.

[15]. White,R., Muller, F. , Healy,C., Whalley,D., and Harmon, M.,1997, Timing Analysis for Data Caches and Set-Associative Caches. IEEE Real-Time Technology and Applications Symposium (RTAS'97), 192–202.

[16]. Colin, A. and Puaut, I., 2000, Worst Case Execution Time Analysis for a Processor with Branch Prediction.Journal of Real-Time Systems, 18, 2/3, 249–274.

[17]. Mitra, T. and Roychoudhury, A. , 2001, Effects of Branch Prediction on Worst Case Execution Time of Programs. National University of Singapore (NUS),Tech. Rep. 11-01.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Sequence Labeling for Two Word Disambiguation in Telugu Language Sentences

Jinka Sreedhar*[1], A. Jagan[1], SK Althaf Hussain Basha[2], Baijnath Kaushik[3], D. Praveen Kumar[4]

[1]Department of Computer Science and Engineering B.V.Raju Institute of Technology, Narsapur, Telangana, India

[2]Department of Computer Science and Engineering Gokaraju Rangaraju Institute of Engineering and Technoloy,Hyderabad, India

[3]Department of Computer Science and Engineering Shri Mata Vaishno Devi University, Katra, Jammu and Kashmir, India

[4]Department of Computer Science and Engineering Institute of Technology, Dhanbad, Jharkhand, India

ABSTRACT

This paper is intended to apply sequence labelings which are introduced to find out the ambiguity in two-words. These words appear to give rise to ambiguity. They seem to be sequence words and this method can be applied only to these types of words. There is another theory of automata which is a mathematical model. By implementing this model to disambiguate the words of sequence it is found there is a kind of mathematical accuracy equal to that of sequence labeling method. The main aim of finding out these methods, is to find out solution to the problem of ambiguity in two-words sequences. Here designing of automata theory for two-words is dealt with the Two-Words disambiguation rules are explained with examples.

Keywords : Natural Language Processing(NLP),Information Retrieval Systems(IRS), MachineTranslation(MT), Finite Automata(FA), Two-Words disambiguation rules.

I. INTRODUCTION

To explain this theory clearly, four states have been identified. In this process state one may have more than one tag, state two may have more than one tag. Now one tag has been retained in the word one deleting the remaining tags. In the similar manner, the same procedure is continued in the second word order[1,2,3,4]. By doing so, the problem of ambiguity has been resolved. When this process comes to state three, it is treated as completed since it gives a complete sense. In the fourth state regarded as a dead state, all the unwanted tags will be appear[5,6,7,8].

With the help of transitional diagrams and transitional tables, the rules are explained. Transitional diagrams contain states, Parts-of-Speech(POS) tags, start state and final state[5]. These diagrams can be represented with the symbols like Q, ∑, S, F. Here Q stands for states one, two, three and dead states, ∑ contains POS tags, S contains the starting state, that is, state one, and F contains the final state, that is, state three[13,14,15,16,17,18].

Transitional Table is also framed to show how these tags appear in different states and give a picture representation.

W1 :: W2 => W1 :: W2

Where

W1 and W2 are sequence of words in that order.

II. DESIGNING AUTOMATA THEORY FOR TWO WORDS RULES

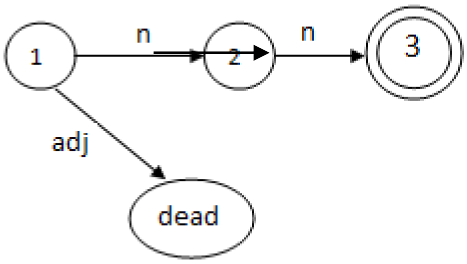


Figure 2.1. Two-word disambiguation for n, adj :: n

Where

1, 2, 3 and dead state belongs to Q and n, adj belongs to ∑.

Here n denotes noun and adj denotes adjective.

Q: {1, 2, 3, dead}

∑: {n,adj}

S: {1}

F: {3}

Table 2.1. Two-word disambiguation for n, adj :: n

∂	n	adj
1	2	dead
2	3	-
3	-	-
dead	-	-

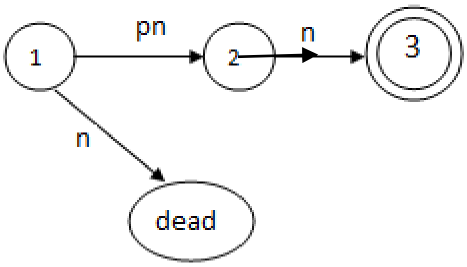


Figure 2.1. Two-word disambiguation for n, pn :: n

Where

1, 2, 3 and dead state belong to Q and n, pn belongs to ∑.

Here n denotes noun and pn denotes pronoun.

Q: {1,2,3,dead}

∑: {pn,n}

S:{1}

F:{3}

Table 2.2. Two-word disambiguation for n, pn :: n

∂	pn	N
1	2	Dead
2	-	3
3	-	-
dead	-	-

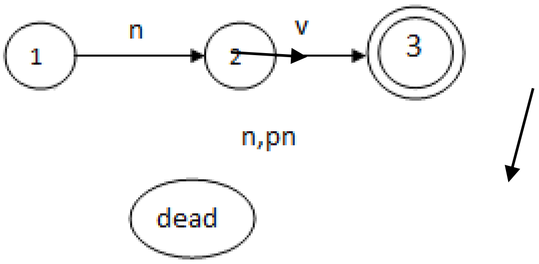


Figure 2.3. Two-word disambiguation for n :: v, n, pn

Where

1, 2, 3 and dead state belong to Q and n, v, pn belongs to ∑.

Here n denotes noun, v denotes verb and pn denotes pronoun.

Q: {1,2,3,dead}

∑: {n,v,pn}

S: {1}

F: {3}

Table 2.3. Two-word disambiguation for n :: v, n, pn

∂	n	v	pn
1	2	-	-
2	dead	3	dead
3	-	-	-
dead	-	-	-

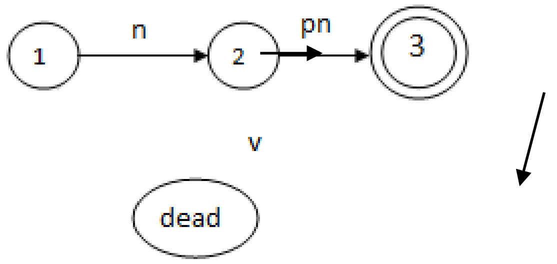


Figure 2.4. Two-word disambiguation for n :: pn, v

Q: {1,2,3,,dead}
∑: {n,v,pn}
S: {1}
F: {3}

Table 2.4. Two-word disambiguation for n :: pn, v

∂	n	pn	v
1	2	-	-
2	-	3	dead
3	-	-	-
dead	-	-	-

Where
1, 2, 3 and dead state belong to Q and n, v, pn belongs to ∑.
Here n denotes noun, v denotes verb and pn denotes pronoun.

Table 2.5. WSD Two-Word Rules with Sentence id's in the Telugu Corpus

S.NO	SENTENCE ID	BEFORE DISAMBIGUATION RULE	AFTER DISAMBIGUATION RULE (RESULT)
1	14784	n,adj :: n => n :: n	n :: n
2	274	n,pn :: n =>pn :: n	pn :: n
3	153	n :: n,pn,v => n :: v	n :: v
4	2291	n :: v,pn => n :: pn	n :: pn
5	10349	avy :: v,pn => avy :: v	avy :: v
6	21560	v ,pn :: avy => v :: avy	v :: avy
7	16646	v,n :: n => n :: n	n :: n
8	24355	n :: n,v => n :: v	n :: v
9	13677	v,pn :: avy =>pn :: avy	pn :: avy
10	442	n :: v,n,pn =>n :: pn	n :: pn
11	531	n :: v,pn => n :: v	n :: v
12	4552	n :: v,pn => n :: pn	n :: pn
13	25974	n :: v,n => n :: n	n :: n
14	12455	pn :: v,pn => pn :: pn	pn :: pn
15	656	avy :: v,pn => avy :: v	avy :: v
16	1893	pn,v :: v => pn :: v	pn :: v
17	590	pn :: adj,n => pn :: n	pn :: n
18	560	n :: v,pn => n :: v	n :: v
19	18714	n,adj :: n => adj :: n	adj :: n

Here n is noun, v is verb, pn is pronoun, adj is adjective and adv is adverb.

From rule 2 when a word carries tags (n,pn) and is followed by another word carrying the tag n, then the tag pn is retained eliminating the n from (n,pn).

From rule 9 a word carrying the tag, such as(n,pn) followed by avy, then most of the times pn will be retained and v will be eliminated. Depending on the context, the linguist will decide which tag will be retained and which one has to be eliminated. These are mostly contextually based syntactic rules. If two-word sequences are unable to resolve unique tags, then three-word, four-word sequence rules may be used for disambiguation.

III. CASE STUDY FOR TWO WORD AMBIGUITY

A Telugu sentence may have ambiguous words from Telugu corpus, like

Sentence: Adaxi aNacivewaku alavAtu padipoyiMxi.

MORPH OUTPUT:

Adaxi	Ada /adj,n
aNacivewaku	aNacivewa/n
alavAtu	alavAtu /n
padipoyiMxi	padu/v,adv,pn,n

Before Applying Disambiguation Rule:

W1 = Ada
W2 = aNacivewa
w1 :: w2 => w1 :: w2
n,adj :: n => n :: n

In the given Telugu sentence the word carries tags (n,adj) and is followed by another word carrying the tag n. Then the tag adj is retained eliminating the n from (n,adj), so from the above sentence adj is eliminated and n is retained.

After Applying Disambiguation Rule:

Adaxi aNacivewaku alavAtupadipoyiMxi .
n n n v
punc

Where punc is punctuation.

IV. ANALYSIS OF TWO WORD DISAMBIGUATION

The following figure 4.1 gives an analysis of the Accuracy. While X-axis indicates the number of test sessions, Y-axis indicates the Accuracy. As a result, the proposed method can disambiguate nearly 98% of ambiguity [59].

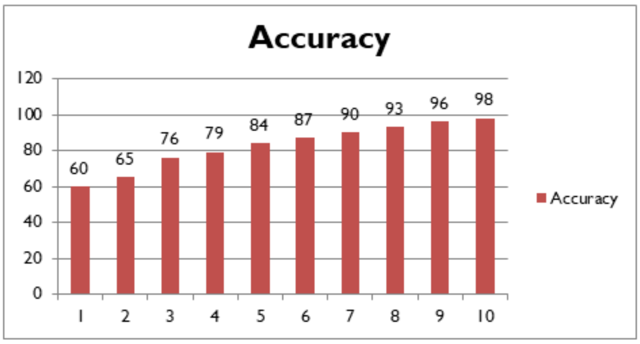


Figure 4.1. Two word disambiguation rules accuracy

V. CONCLUSION

Here dealing with the designing of two-word rules for Telugu Language Sentence word order. To make things easy to understand some rules have been made which can be applied for the word order of Telugu Language Sentences and it clarifies the ambiguity. All these things are so vividly explained with the help of case studies and theoretical explanations. When these rules are applied whenever needed, they help the user easily to eliminate the ambiguity. These theories help understand the study of disambiguation. By applying disambiguation rules it is found that the proposed method can disambiguate nearly 98% of the ambiguity. The theoretical explanation and disambiguation rules have resulted in the accuracy of evidences.

VI. REFERENCES

[1]. Noam Chomsky, "Logical syntax and semantics: their linguistic relevance", vol.31, No.1, pp: 36-45, 1955.

[2]. Noam Chomsky, "On Certain Formal Properties of Grammars", Information and Control, Vol. 9, pp: 137-167, 1959.

[3]. Nou, Chenda and WataruKameyama,Khmer, "POS Tagger: A Transformation-based Approach with Hybrid Unknown Word Handling", Proceedings of the First IEEE International Conference on Semantic Computing (ISCS), Irvine, CA. pp: 482-492, 2007.

[4]. PawanGoyal,LaxmidharBehera,Thomas Martin McGinnity, "Query Representation through Lexical Association for Information Retrieval", IEEE Transactions on Knowledge and Data Engineering, pp: 2260-2273, 2012.

[5]. PengJin,XingyuanChen,"A Word Sense Probabilistic Topic Model", 9th International Conference on Computational Intelligence and Security (CIS), pp: 401-404, 2013.

[6]. PengYuan Liu, "Another View of the Features in Supervised Chinese Word Sense Disambiguation", 9thInternational Conference on Computational Intelligence and Security, ISBN: 978-0-7695-4584-4, pp: 1290-1293, 2011.

[7]. Pengyuan Liu, YongzengXue, Shiqi Li, Shui Liu, "Minimum Normalized Google Distance for Unsupervised Multilingual Chinese-English Word Sense Disambiguation", International Conference on Genetic and Evolutionary Computing, ISBN: 978-0-7695-4281-2, 2010.

[8]. Ping Chen,BowesC,Wei Ding, et..al, "Word Sense Disambiguation with Automatically Acquired Knowledge",IEEE INTELLIGENT SYSTEMS, 2012.

[9]. PrashanthMannem, "Bidirectional Dependency Parser for Hindi, Telugu and Bangla", Proceedings of ICON09 NLP Tools Contest: Indian Language Dependency Parsing, India, 2009.

[10]. Quinlan, J. R, "Induction of decision trees", Mach. Learn. 1, 1, pp: 81–106, 1986.

[11]. Quinlan, J. R, "Programs for Machine Learning", Morgan Kaufmann, San Francisco, CA, 1993.

[12]. R.M.K.Sinha and K. Sivaraman, "Ambiguity Resolution in Anglabharati",TRCS-93-174, Department of Computer Science and Engineering, IIT,Kanpur, India, 1993

[13]. R. Mahesh K. Sinha,"Learning Disambiguation of Hindi Morpheme 'vaalaa' with a Sparse Corpus" 4th International Conference on Machine Learning and Applications, ISBN: 978-0-7695-3926-3, pp: 653-657, 2009.

[14]. J.Sreedhar, S. Viswanadha Raju, A. Vinaya Babu, Amzan Shaik, P.Pavan Kumar "Word Sense Disambiguation : An Empirical Survey" International Journal of Soft Computing and Engineering(IJSCE), Volume-2,Issue-2,May-2012,ISSN:2231-2307.

[15]. J.Sreedhar, S. Viswanadha Raju, A. Vinaya Babu, Amzan Shaik, P.Pavan Kumar"A critical Approaches to Identification of Disambiguation Words in NLP : A Current State of the Art" International Journal of Engineering Trends and Technologies (IJETT), Volume-3,Issue-3,May-2012,ISSN:2231-5381.

[16]. P.Pavan Kumar, J.Sreedhar "Innovative Techniques and Technologies in Translation in a Multilingual Context" 3rd International Conference on Translation Technology and Globalization in Multilingual Context" Delhi, June 23-26, 2012.

[17]. P.Pavan Kumar, J.Sreedhar "Language teaching and MLE in the context of the third revolution" DLA Conference on Dravidian Languages and Translation Technology" HCU, Hyderabad, June 18-20, 2012.

[18]. Hyuk-Chul Kwon, Minho Kim, Youngim Jung, "Hybrid word sense disambiguation using language resources for transliteration of Arabic numerals in Korean", International Conference on Hybrid Information Technology (ICHIT '09), ISBN: 978-1-60558-662-5, pp: 314-321,2009.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Empirical Analysis of Context Sensitive Grammars and Parse Trees for Disambiguiting Telugu Language Sentences

Jinka Sreedhar*[1], SK Althaf Hussain Basha[1], D. Praveen Kumar[2], A. Jagan[3] , Baijnath Kaushik[4]

[1]Department of Computer Science and Engineering Gokaraju Rangaraju Institute of Engineering and Technoloy, Hyderabad, India

[2] Department of Computer Science and Engineering Institute of Technology, Dhanbad, Jharkhand, India

[3]Department of Computer Science and Engineering B.V.Raju Institute of Technology, Narsapur, Telangana, India

[4]Department of Computer Science and Engineering Shri Mata Vaishno Devi University, Katra, Jammu and Kashmir, India

ABSTRACT

This research paper explores the impact of Context Sensitive Grammars (CSG) and Parse Trees for construction of a Telugu Language Sentences. Based on the CSG Rules here we derived the derivations for the respective strings. Later we constructed the Parser Trees for the above said strings. Finally we analysed whether the string is ambiguous or unambiguous. Here for analysis we considered the Large Scale Open Source Telugu carpus. The main aim of finding out these methods, is to find out solution to the problem of ambiguity in Telugu Language Sentences. Here designing of Context Sensitive Grammars Rules and Parse Trees are explained with examples.

Keywords : Natural Language Processing (NLP), Context Sensitive Grammars (CSG) Rules, Parse Trees(PT's), Derivations.

I. INTRODUCTION

The syntax of a language may be specified using a notation called Context Sensitive Grammar (CSG). A context sensitive grammar consists of terminals, non-terminals, a start symbol and production rules. The set of tokens are called the terminal symbols. These are the basic symbols from which strings are formed. Non terminals are the symbols which represent syntactic variables that denote sets of strings. They do not exist in the source program they only help in defining the language generated by the grammar. One of the non-terminals designated as the start symbol. We shall follow the convention of listing the production for the start symbol. The set of strings denoted by the start symbol is the language defined by the grammar. A production rule has a non-terminal symbol on the left hand side followed by an arrow and a sequence of symbols on the right side. This sequence of symbols may contain a combination of terminals and non-terminals[9,11,13].

The organization of this paper is as follows: Section II describes the CSG and its notations, Section III case study IV deals with derivations of CSG Grammar, Section V explores the Parser Trees , Section VI shows the acknowledgements and Section VII deals with conclusion and future enhancement followed by the references.

II. CONTEXT SENSITIVE GRAMMARS

We may have more than one production rule for the same non terminal. In that case, we can group their right hand side by using symbol | to separate the alternate right hand side. The Context Sensitive Grammar explains the sensitive nature of words by its applications. In general a CSG [10,12,17] is a set of recursive rewriting rules called productions that are used to generate patterns of strings and it consists of the following components:

- ✓ A finite set of terminal symbols (Σ).
- ✓ A finite set of non-terminal symbols (NT).
- ✓ A finite set of productions (P).
- ✓ A start symbol (S).

Let G be a Context Sensitive Grammar for which the production rules are:

$$S \Rightarrow NP\ \ VP$$
$$N\ NP \Rightarrow N\ \ NP\ |QC\ \ NP|\ SYM$$
$$V\ VP \Rightarrow VP\ \ NP\ |V\ \ SYM|\ SYM$$
$$NP\ \ VP\ \ \Rightarrow VP$$

Figure 2.1. Context Sensitive Grammar

Where,S is a Sentence, NP is a Noun Phrase, VP is a Verb Phrase, N is a Noun, V is a Verb, QC is a Cardinal, SYM is a Sym.

III. CASE STUDY

In this method of explaining the possibility of avoiding the ambiguity in the structure of Telugu language sentence, one Telugu Sentence has been taken to explain how this arithmetical method has worked out. For example, the sentence is :

1. rEwulu edAxiki mUdu kArla paMtalu paMdiswAru .
   N N QC N N V SYM

Here, in the Telugu sentence 1, each part is segregated and named as N,N,QC,N,N,N,V and SOV and the POS tags explain how this method eliminates the ambiguity of the sense of the sentence when it is applied.

All these views have been taken from the Morphological Analyzer as an example and certain rules have been set in this process of explaining how a sense ambiguity can be avoided in the Telugu language sentence structure.

From sentence 1 the word kAru is taken to explain the meaning of one which is used only in the region of Rayalaseema which has its own dialect and a similar meaning cannot be seen in the two other regions of Telugu speaking states, namely, Andhra and Telangana.

kAru comes under the category of noun. In taking this word as a noun, its meaning has been lost. It is a moving vehicle. In this particular context though it plays the role if an adjective which tells the number, in this context it is considered a Noun.

IV. DERIVATIONS

Here Derivation provides a means for generating the sentences of a language. If one chooses the leftmost non-terminal in a given sentential form then it is called leftmost derivation. If one chooses the rightmost non-terminal in a given sentential form then it is called rightmost derivation. Derivation from S means generation of string w from S. Any language construct can be defined by the CSG [3,15,16]. The above grammar generates different strings by providing many sentential forms as shown below.

$$
\begin{array}{l}
S \Rightarrow NP\ \ VP \\
\Rightarrow NN\ \ NP\ VP \\
\Rightarrow N\ \ NP\ VP \\
\Rightarrow\ N\ \ NN\ \ NP\ VP \\
\Rightarrow N\ \ NNP\ VP \\
\Rightarrow N\ \ N\ \ QC\ \ NP\ VP \\
\Rightarrow NN\ \ QC\ \ NN\ \ NP\ VP \\
\Rightarrow N\ \ N\ \ QC\ \ N\ \ N\ \ NP\ VP \\
\Rightarrow N\ \ N\ \ QC\ \ N\ \ \ N\ \ NP\ VP \\
\Rightarrow N\ \ N\ \ QC\ \ N\ \ N\ VP \\
\Rightarrow N\ \ N\ \ QC\ \ N\ \ N\ \ V\ \ SYM
\end{array}
$$

Figure 4.1. Derivation for the input sentence 1

As an explanation of these derivations, the first one is a sentence deriving Noun Phrase and Verb Phrase, and by taking the Noun Phrase the Noun and Noun Phrase have been derived.

As a second step of derivation, the Noun Phrase has been taken and from it Noun and Noun Phrase have been derived. At the third stage Noun phrase and Verb phrase have been derived. From Noun Phrase, Noun and Noun Phrase have been derived.

Now from the Verb Phrase only the verb has been derived.

From the Figure 4.1, the following are clarified so as to root out the ambiguity in the sentence word order. They are:

N N QC N N V SYM

A different representation of the Parse Tree is given to explain how this ambiguity in the sentence word order can be avoided.

V. PARSE TREES

A parse tree [1,4,5] is an equivalent form of showing a derivation which represents a derivation graphically or pictorially. A parse-tree is an internal structure, created by the compiler or interpreter while parsing some language construction. Parsing is also known as 'syntax analysis'.

A parse tree for a grammar G is a tree where

- ✓ the root is the start symbol for G
- ✓ the interior nodes are the non-terminals of G
- ✓ the leaf nodes are the terminal symbols of G.
- ✓ the children of a node T (from left to right) correspond to the symbols on the right hand side of some production for T in G.

Every terminal string generated by a grammar has a corresponding parse tree; every valid parse tree represents a string generated by the grammar (called the yield of the parse tree).

In this parse tree method of explaining, the first one is S which stands as the root of the parse tree. From this S, the NP and VP have been used to construct the word order. From NP, NN and NP have been taken to derive the Noun, ie. N. Again NP is taken to explain NN and NP, and derive the Noun, ie. N. Again, in the similar manner, NP has been taken directly to derive the Noun, ie. N. Now to derive a verb, Verb Phrase has been taken, all these explained in the Figure 5.1.

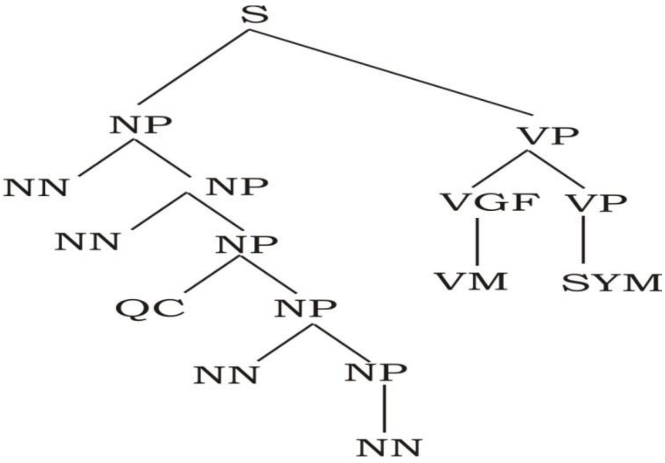


Figure 5.1. Parse tree for Input sentence 1

All these methods explain clearly how a sentence structure can be changed to get the proper sense and to avoid the ambiguity in all kinds of word order.

VI. ACKNOWLEDGMENTS

We are very thankful to all the esteemed authors in a reference list, to make this research article in a better shape and in right direction.

VII. CONCLUSION & FUTURE ENHANCEMENT

Here we described about the impact of noun ambiguity has been analyzed in Telugu Language Sentences empirically. The noun ambiguity in the Telugu Language Sentences is rooted out by applying the Context Sensitive Grammar Rules. There is a scope for further research on verbs, adjectives and adverbs to measure their impact.

VIII. REFERENCES

[1]. Aho, A.V., and Johnson, S.C.1974]. "LR parsing," Computing Surveys 6:2, 99-124.

[2]. Aho, A.V., and Johnson, S.C. , and Ullman, J. D.1975]."Deterministic parsing of ambiguous grammars," Comm. ACM 18:8, 441-452.

[3]. Aho, A.V., and Peterson, T.G,1972]. "A minimum distance error correcting parser for context-free languages," SIAM J. Computing 1:4, 305-312.

[4]. Aho, A.V., and Ullman, J.D. 1972b]. The Theory of Parsing Translation and Compiling, Vol. I:Parsing, Prentice-Hall, Englewood Cliffs, N. J.

[5]. Aho, A.V., and Ullman, J.D. 1972c]. "Optimization of LR(k) parsers," J. Computer and systems Sciences 6:6, 573-602.

[6]. Aho, A.V., and Ullman, J.D.1972a]. The Theory of Parsing, Translation and Compiling, Vol. II: Compiling, Prentice- Hall, Englewood Cliffs, N. J.

[7]. Aho, A.V., and Ullman, J.D.1972b]. " A technique for speeding up LR(k) parsers." SIAM J. Computing 2:2, 106-127.

[8]. Anderson, J. P. 1964]. "A note on some compiling algorithms," comn. ACM 7:3, 149-153.

[9]. Anderson, T., Eve, J., and Horning, J. J.1973]. " Efficient LR(1) paresers," Acta Informatica 2:1, 12-39.

[10]. Backhouse, R.C. 1976]. "An alternative approach to the improvement of LR parsers," Acta Informatica 6:3, 277-296.

[11]. Bar Hillel, Y., Perles, M., and Shamir, E. 1961]. "On formal properties of simple phrase structure grammers," Z. Phonetik, Sprachwissenschaft und Kommunikationsforschung 14, pp. 143-172.

[12]. Barnard, D. T. 1975]. "A survey of syntax error handling techniques," Computer Science Reaserch Group, Univ. of Toronto, Toronto, Ont., Canada.

[13]. Birman, A., and Ullman, J.D. 1973]. "Parsing algorithms with backtrack," Information and Control 23:1, 1-34.

[14]. Bochmann, G. V. 1976]. "Semantic evaluation from left to right," Comm. ACM 19:2, 55-62.

[15]. Brzozowiski, J. A. 1964]. "Derivatives of regular expressions," J. ACM 11:4, 481-488.

[16]. Cheatham, T. E. Jr., and Sattley, K. 1964]. "Syntax directed compiling," Proc. AFIPS 1964 Spring Joint Computer Conf. Spartan Books, Baltimore Md., 31-57.

[17]. Chomsky, N. 1959]. "On Certain formal properties of grammers," Information and Control 2:2, 137-167.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

NLP : Context Free Grammars and Parse Trees for Disambiguiting Telugu Language Sentences

Jinka Sreedhar*[1], SK Althaf Hussain Basha[1] , D. Praveen Kumar[2], A. Jagan[3], Baijnath Kaushik[4]

[1]Department of Computer Science and Engineering Gokaraju Rangaraju Institute of Engineering and Technoloy,Hyderabad, India

[2] Department of Computer Science and Engineering Indian Institute of Technology, Dhanbad, Jharkhand, India

[3]Department of Computer Science and Engineering B.V.Raju Institute of Technology, Narsapur, Telangana, India

[4] Department of Computer Science and Engineering Shri Mata Vaishno Devi University, Katra, Jammu and Kashmir, India

ABSTRACT

Several studies have explained the benefits of using Context Free Grammars(CFGs) of derivations and Parse Trees to reduce the ambiguity in Natural Language Sentences. However, these benefits are dependent on the CFG Rules and Derivation steps. This research paper explains the power of CFGs and Parse Trees for construction of a Telugu Language Sentences. Based on the CFG here we derived the derivations for the respective strings. Later we constructed the Parser Trees for the above said strings. Finally we analysed whether the string is ambiguous or unambiguous. Here we considered the Large Scale Open Source Telugu carpus for analysis.

Keywords : CFG, Parse Trees, Derivations, Telugu Corpus

I. INTRODUCTION

The syntax of a language may be specified using a notation called Context Free Grammar (CFG). A Context Free Grammar consists of terminals, non-terminals, a start symbol and production rules. The set of tokens are called the terminal symbols. These are the basic symbols from which strings are formed. Non terminals are the symbols which represent syntactic variables that denote sets of strings. They do not exist in the source program they only help in defining the language generated by the grammar. One of the non-terminals designated as the start symbol. We shall follow the convention of listing the production for the start symbol. The set of strings denoted by the start symbol is the language defined by the grammar. A production rule has a non-terminal symbol on the left hand side followed by an arrow and a sequence of symbols on the right side. This sequence of symbols may contain a combination of terminals and non-terminals[9,11,13].

The organization of this paper is as follows: Section II describes the CFG and its notations, Section III deals with derivations of CFG Grammar, Section IV explores the Parser Trees , Section V shows the acknowledgements and Section VI deals with conclusion followed by the references.

II. CONTEXT FREE GRAMMARS

Here Context Free Grammar rules and regulations in Telugu language are explained. From this Context Free Grammar Derivations and Parse Trees are for the given sentence are derived[6,7,8]. These methods resolve the problem of ambiguity and help in the understanding of the sense of the sentence without any misunderstanding.

The set of tokens are called the terminals from which strings are composed. Non-terminals represent syntactic variables that denote sets of strings. They only help in defining the language generated by the grammar [9]. The strings denoted by start symbol constitute language as defined by grammar.

We may have more than one production rule for the same non terminal. In that case, we can group their right hand side by using symbol | to separate the alternate right hand side. CFG, sometimes called a phrase structure grammar[2] plays a central role in the description of natural languages. In general a CFG [10,11,12,17] is a set of recursive rewriting rules called productions that are used to generate patterns of strings and it consists of the following components:

- ✓ A finite set of terminal symbols (Σ).
- ✓ A finite set of non-terminal symbols (NT).
- ✓ A finite set of productions (P).
- ✓ A start symbol (S).

Let G be a Context Free Grammar for which the production rules are:

$$\begin{aligned}
&S \Rightarrow NP\ VP\\
&NP \Rightarrow Noun \mid VP\ PP \mid \Pr onoun\\
&VP \Rightarrow Verb \mid NP\ PP \mid VP\ PP \mid \varepsilon\\
&PP \Rightarrow NP\ PP \mid NP\ PP \mid VP\ PP \mid \varepsilon\\
&Noun \Rightarrow n\\
&Verb \Rightarrow v\\
&\Pr onoun \Rightarrow pn
\end{aligned}$$

Figure 2.1. Context Free Grammar

Where, S stands for Sentence, NP stands for Noun Phrase, VP stands for Verb Phrase, PP stands for Prepositional Phrase, Pn stands for pronoun, n stands for noun, v stands for verb.

III. DERIVATIONS

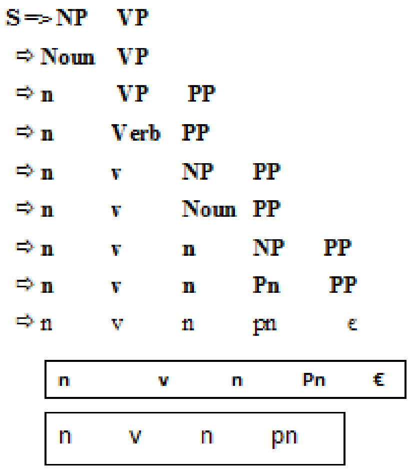
Here Derivation provides a means for generating the sentences of a language. If one chooses the leftmost non-terminal in a given sentential form then it is called leftmost derivation. If one chooses the rightmost non-terminal in a given sentential form then it is called rightmost derivation. Derivation from S means generation of string w from S. Any language construct can be defined by the CFG [3,15,16]. The above grammar generates different strings by providing many sentential forms as shown below.

In the sentence 1, for example, will explain how these POS change for different purposes. Just in order to explain this method, this derivation is explored.

3.1 Methodology for Derivations

1. waMdri ceVppina viRayAlu AlociMcevAdu.
   n v n pn

Here, in the sentence 1, as an example there is a noun phrase and a verb phrase and noun phrase (NP) has been taken to find out noun (n). The ambiguity can be cleared by explaining the sentence 1 in the Figure 3.1.



```
S=>NP    VP
 ⇨ Noun  VP
 ⇨ n     VP    PP
 ⇨ n     Verb  PP
 ⇨ n     v     NP    PP
 ⇨ n     v     Noun  PP
 ⇨ n     v     n     NP    PP
 ⇨ n     v     n     Pn    PP
 ⇨ n     v     n     pn    €

n        v     n     Pn    €

n    v    n    pn
```

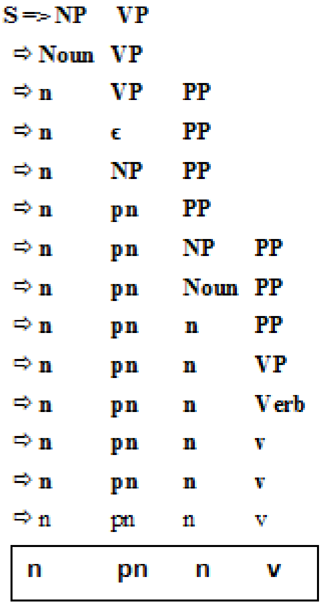
Figure 3.1. Derivation of "n v n pn"

The start symbol of the above grammar is S. Any grammar contains terminals and non-terminals. The non-terminal symbol occurs at the left hand side. These are the symbols which need to be expanded. The non-terminals are replaced by the terminals which it derives.

The above string is derived from S step by step as follows:

- ✓ First the nonterminal NP present at the left side is replaced by its substring noun.
- ✓ Then it is substituted by its substring n.
- ✓ Then VP is substituted by its substring VP PP.
- ✓ Then again VP is substituted by its substring Verb.
- ✓ Then that Verb is substituted by its substring v.
- ✓ Then PP is substituted by its substring NP PP.
- ✓ Then again NP is substituted by its substring Noun, and then Noun is substituted by its substring n.
- ✓ Then again PP is substituted by its substring NP PP.
- ✓ Then again NP is substituted by its substring Pronoun.
- ✓ Finally, PP is substitued by its substring €.
- ✓ So that, finally we obtain the string.

S-> n v n Pn € from S-> NP VP



```
S=>NP    VP
 ⇨ Noun  VP
 ⇨ n     VP    PP
 ⇨ n     €     PP
 ⇨ n     NP    PP
 ⇨ n     pn    PP
 ⇨ n     pn    NP    PP
 ⇨ n     pn    Noun  PP
 ⇨ n     pn    n     PP
 ⇨ n     pn    n     VP    PP
 ⇨ n     pn    n     Verb  PP
 ⇨ n     pn    n     v     PP
 ⇨ n     pn    n     v     €
 ⇨ n     pn    n     v

n     pn    n     v
```

Figure 3.2. Derivation of "n pn n v"

The above string is derived from S step by step as follows:

- ✓ The non-terminal NP present at the left side is replaced by its substring noun.
- ✓ Then it is substituted by its substring n.
- ✓ Then VP is substituted by its substring VP PP.
- ✓ Then again VP is substituted by its substring €.
- ✓ € means null value, so we can just eliminate it.
- ✓ Then PP is substituted by its substring NP PP.
- ✓ Then NP is substituted by its substring pronoun (pn).
- ✓ Then again PP is substituted by its substring NP PP.
- ✓ Then again NP is substituted by its substring Noun, and then Noun is substituted by its substring n.
- ✓ Then again PP is substituted by its substring VP PP.
- ✓ Then again VP is substituted by its substring Verb ,and then Verb is substituted by one of the substring v.
- ✓ Finally, PP is substituted by its substring €.
- ✓ € means null value, so we can just eliminate it.
- ✓ So that, finally we obtain the string

S-> n pn n v from S-> NP PP

IV. PARSE TREES

A parse tree [1,4,5] is an equivalent form of showing a derivation which represents a derivation graphically or pictorially. A parse-tree is an internal structure, created by the compiler or interpreter while parsing some language construction. Parsing is also known as 'syntax analysis'[13,14].

A parse tree for a grammar G is a tree where

- ✓ the root is the start symbol for G
- ✓ the interior nodes are the non-terminals of G
- ✓ the leaf nodes are the terminal symbols of G.
- ✓ the children of a node T (from left to right) correspond to the symbols on the right hand side of some production for T in G.

Every terminal string generated by a grammar has a corresponding parse tree; every valid parse tree represents a string generated by the grammar (called the yield of the parse tree).

Parse Trees for Sentence 1:

Consider the below grammar, implementing the parse tree for the strings generated by this grammar.

S =>NP VP

NP => Noun | Pronoun

VP => Verb | VP PP| ε

PP => NP PP|VP PP| ε

Noun => n

Verb => v

Pronoun =>pn

Figure 4.1. Context Free Grammar

1) This grammar generates the string n v n pn. The parse tree for this string using CFG is as following steps.
2) Create a root labeled with S.
3) For each sentential form αi in the derivation, i ≥ 2, construct a parse tree whose yield is αi ,

We can use induction for constructing the for αi , given the tree for αi-1 as given below:

a. The tree for α1 = S is a single node labeled S.
b. Let αi-1 = X1 X2 …. Xr and αi is derived from αi-1 by replacing Xj by β = Y1 Y2 ….. Yk.

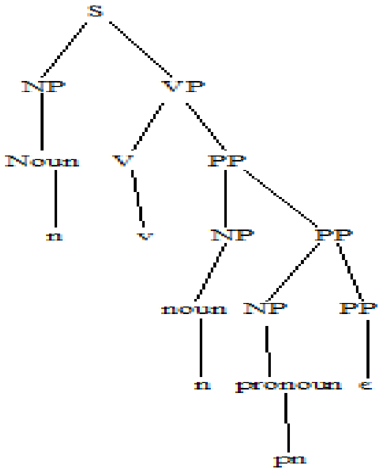


Figure 4.2. Parse Tree for "nvnpn"

S is a start symbol which derives NP VP, NP is a non-terminal which is substituted by noun and it is in turn substituted by the terminal n.

Now VP derives VP PP, PP with NP PP. NP is substituted by noun and with n.

Similarly PP derives NP PP and NP with the terminal pn. Finally, we obtain the string n v n pn.

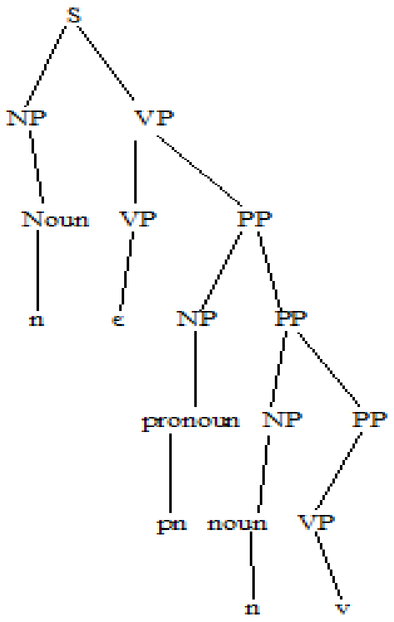


Figure 4.3. Parse Tree for "npnnv"

S is the start symbol for the above grammar which derives NP PP. NP is reduced to noun and inturn by n.

VP derives VP PP and PP to NP PP. Now NP is reduced to pronoun and to the terminal pn. Next PP is substituted by NP PP where NP to noun and PP to VP. Finally, we obtain the string n pn n v.

V. ACKNOWLEDGMENTS

We are very thankful to all the esteemed authors in a reference list, to make this research article in a better shape and in right direction.

6. CONCLUSION and FUTURE ENHANCEMENT

Here we described about the Context Free Grammars, Derivations and Parse Trees. We observed the ambiguity between the Telugu Language Sentences. There is a scope for further research on other Natural Languages of nouns, verbs, adjectives and adverbs to measure their impact

VI. REFERENCES

[1]. Aho, A.V., and Johnson, S. C. 1974]. "LR parsing," Computing Surveys 6:2, 99-124.

[2]. Aho, A.V., and Johnson, S. C. , and Ullman, J. D. 1975]. "Deterministic parsing of ambiguous grammars," Comm. ACM 18:8, 441-452.

[3]. Aho, A.V., and Peterson, T.G, 1972]. "A minimum distance error correcting parser for context-free languages," SIAM J. Computing 1:4, 305-312.

[4]. Aho, A.V., and Ullman, J. D. 1972b]. The Theory of Parsing Translation and Compiling, Vol. I:Parsing, Prentice-Hall, Englewood Cliffs, N. J.

[5]. Aho, A.V., and Ullman, J. D. 1972c]. "Optimization of LR(k) parsers," J. Computer and systems Sciences 6:6, 573-602.

[6]. Aho, A.V., and Ullman, J. D. 1972a]. The Theory of Parsing, Translation and Compiling, Vol. II: Compiling, Prentice- Hall, Englewood Cliffs, N. J.

[7]. Aho, A.V., and Ullman, J. D. 1972b]. " A technique for speeding up LR(k) parsers." SIAM J. Computing 2:2, 106-127.

[8]. Anderson, J. P. 1964]. "A note on some compiling algorithms," comn. ACM 7:3, 149-153.

[9]. Anderson, T., Eve, J., and Horning, J. J.1973]. " Efficient LR(1) paresers," Acta Informatica 2:1, 12-39.

[10]. Backhouse, R.C. 1976]. "An alternative approach to the improvement of LR parsers," Acta Informatica 6:3, 277-296.

[11]. Bar Hillel, Y., Perles, M., and Shamir, E. 1961]. "On formal properties of simple phrase structure grammers," Z. Phonetik, Sprachwissenschaft und Kommunikationsforschung 14, pp. 143-172.

[12]. Barnard, D. T. 1975]. "A survey of syntax error handling techniques," Computer Science

Reaserch Group, Univ. of Toronto, Toronto, Ont., Canada.

[13]. Birman, A., and Ullman, J.D. 1973]. "Parsing algorithms with backtrack," Information and Control 23:1, 1-34.

[14]. Bochmann, G. V. 1976]. "Semantic evaluation from left to right," Comm. ACM 19:2, 55-62.

[15]. Brzozowiski, J. A. 1964]. "Derivatives of regular expressions," J. ACM 11:4, 481-488.

[16]. Cheatham, T. E. Jr., and Sattley, K. 1964]. "Syntax directed compiling," Proc. AFIPS 1964 Spring Joint Computer Conf. Spartan Books, Baltimore Md., 31-57.

[17]. Chomsky, N. 1959]. "On Certain formal properties of grammers," Information and Control 2:2, 137-167.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Computer Aided Analysis of Chest X-Ray Images for Early Detection of Cardiomegaly using Euler Numbers

J Ebenezer[1], A C S Rao[2]

[1]Department of Computer Science, Vignan's Foundation for Science Technology and Research, Guntur, Andhra Pradesh, India

[2]Department of Computer Science, IIT(ISM) Dhanbad, Jharkhand, India

ABSTRACT

Cardiomegaly is an unusual cardiac condition in which the human heart grows larger in size and becomes bigger than it usually is. Cardiomegaly can be detected early by computing Cardiothoracic Ratio(CTR) from chest X-ray images (CXR).As it is difficult for medical experts to examine CXR manually, a Computer-Aided Diagnosis (CAD) system is required to precisely calculate the Cardiothoracic ratio and accurately predict the onset of Cardiomegaly. In this paper, we use euler number based thresholding method for lung region segmentation from CXR images. The resultant binarized image is used for calculating Cardiothoracic Ratio using a computational algorithm. The proposed method is experimented on two datasets: JRST and India. JRST contains 247 chest X-rays and India set contains 100 chest X-rays. An overall accuracy of 96.25% and the overall (lung segmentation time + CTR computation time) average computation of 0.8215 seconds was acheived. The proposed method is compared with existing methods and it gives high accuracy and high performance.

Keywords : Chest X ray images; Computer Aided Analysis;Euler number; Cardiomegaly; Cardiothoracic Ratio Computation.

I. INTRODUCTION

Cardiomegaly or Enlarged heart is a medical condition in which heart size increases which may be due to various factors such as high blood pressure, abnormal heart valve, HIV infection, Kidney disease or genetically inhetited. It is of vital importance to detect Cardiomegaly in early stages as it may give rise to other serious heart diseases like congestive heart failure.

Cardiomegaly is not a disease itself but rather a symptom which marks the onset on various other kinds of diseases like coronary artery disease or congestive heart failure. Therefore early detection of cardiomegaly results in diagnosis of underlying symptom. Heart diseases are life threatening diseases, and it is important to detect their symptoms early. Treatment of the disease at an early stage yeild positive results. Copyright © 201X Inderscience Enterprises Ltd.

2 J Ebenezer et al.

Although Computed Tomograph (CT) and Magnetic Resonance Imaging (MRI) can be more efficient than X-ray, the latter is more generally available and doctors often rely on CXR for making quick decisions in emergency situations. Cuurently medical doctors perform preliminary diagnosis for heart diseases based on chest X-ray images (CXR). The manual process is not only time consuming but also error prone. It is difficult to analyze the chest X-ray images if they are huge in number, which is a common situation in populous countries. To overcome the difficulties, computerised analysis of CXR images can be adopted. CAD improves the diagnostic accuracy and can assist the medical doctors to come to the right conclusion.

Cardiothoracic Ratio (CTR) calculation is a simple, cost-effective yet efficient approach to detect the increase in heart size[16].CTR can be used for predicting cardiomegaly with 95.8% accuracy[1]. The CTR is the ratio between the maximum transverse cardiac diameter (CD) and the maximum thoracic diameter (TH) which is measured between the inner margins of the ribs. It is computed using posteroanterior chest radiography (PA-CXR).

II. RELATEDWORK

The problem of detecting cardiomelagy from chest X-ray using a computer can be further divided into two subproblems. Cardiomegaly can be detected from X-ray images by calculating CTR and following are important steps.

- ✓ Segmentation of the X-ray image
- ✓ Computation of Cardiothoracic Ratio(CTR)

In the first step, the lung region is seperated from the X-ray image. Researchers used different techniques to seperate the lung region. Method of segmentation largely depends on the image which is going to be segmented. Histogram based thresholding methods are most commonly used. But the limitation of this method is that accurate threshold is not guarenteed. Euler number based thresholding was used for real time applications[2, 3]. For chest X-ray segmentation, euler min max function was used[4].

As a second step, CTR is calculated from the segmented lung region. In the past years, several methods were proposed to calculate Cardiothoracic Ratio (CTR) from chest X-ray using a computer [5]. Early work was by Becker, H. C., et al. [6] in 1964, who digitized 70-mm photofluorograms and computer was used to find out the cardiothoracic ratio (CTR). The first derivative of the horizontal spatial signature was used to find out the cardiac boundaries and lung margins.

An automated diagnosis for rheumatic heart disease was developed by HALL et al. [7] and KRUGER et al. [8]. They computed CTR and other cardiac parameters to locate cardiac boundaries using a gray-scale threshold method. Discriminant function was used to classify cardiac silhouettes.

Sezaki and Ukena [9]in 1973 designed a practical instrument for automated mass screening of heart disease. CTR was computed by a scheme that detects the vertical boundary of the rib cage and the heart by analyzing the horizontal profiles. .

Recent work related to chest x-rays is automated approach [10, 11, 17]. However the existing approaches[12] have accuracy range varying from 73% to 86% , which needs to be improved in order to detect the cardiomegaly using computer. Here, in this paper, all the computations are performed by our self-designed image analysis tool: MedIT which is specifically designed to detect and predict the onset of Cardiomelagy. Lung field segmentation using Boundary Map and Snake Segmentation Algorithm 3 In this paper, details about a CAD tool for detection of early symptoms of cardiomegaly are given. The paper is organised as follows. Section 2 describes the preprocessing method for lung segmentation. Once lung objects have been isolated, the prevalence of Cardiomegaly will be identified by measuring the cardiothoracic ratio. Section 3 presents the self-designed partially automated software: Medit for analysing CXR and predicting about the existence of cardiomegaly based on Cardiothoracic ratio computation. Later sectiospresent the results and conclusions.

III. COMPUTER ASSISTED PROCESSING OF CHEST X-RAY IMAGES

The first step towards automated computation of Cardiothoracic Ratio is to create a binarized CXR image having lungs extracted out from the background. Many techniques can be employed to carry out lung segmentation. Euler number based thresholding technique is used to carry out image segmentation. Once the lungs are isolated, then the image can be analysed for detecting cardiomegaly.

3.1 Preprocessing

In order to enhance the quality of the segmented images, the CXR images need to undergo a preprocessing phase. The chest area needs to be cropped out fromCXRimages if theCXRcontains unnecessary background. The contrast of the image is enhanced using histogram equalization. 2-D Gaussian operator is used not only to get smooth image but also to preserve the edge features. Using the Gaussian operator, noise is also removed.

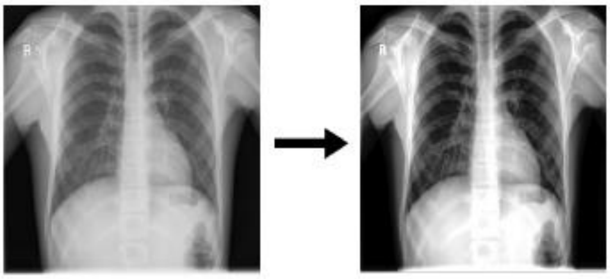


Figure 1. Preprocessing the Lung image. Left image is the image before preprocessing and the right image is the image after preprocessing.

3.2 Lung Segmentation using Euler numbers

Segmentation divides the image into a set of regions R , which consists of homogenous, non-overlapped, connected subregions R$_i$

$$R = \{R_i : i = 1, 2, 3, ..., N\} \qquad (1)$$

The union of all subregions forms the original image i.e.

$$I = R_1 \cup R_2 \cup R_3 \cup ... \cup R_N \qquad (2)$$

4 J Ebenezer et al.

The regions Ri should be connected for all i = 0; 1; 2...N and each region R$_i$ should be homogenous. Different adjacent regions R$_i$ and R$_j$ should be disjoint i.e.

Lung boundary segmentation in Chest X-ray is a problem focused by a number of research groups over the past decade. Different kinds of solutions were proposed. These solutions can be broadly categorized into categories like rule based methods, pixel classification-based methods, deformable model-based methods, and hybrid methods. Here we use Rule-based segmentation method involving thresholding or morphological operations Inspite of existence of various approaches for segmentation, Thresholding is the most common method used. Thresholding converts an input image I to a binary image B as follows:

$$R_i \cap R_j = \phi \qquad (3)$$

where T is the threshold, B (i, j) = 1 for foreground and B (i, j) = 0 for background.

$$B(i,j) = \begin{cases} 1 & \text{if I(i, j)} \geq \text{T} \\ 0 & \text{if I(i, j)} < \text{T} \end{cases} \qquad (4)$$

To find T value histogram based approaches were used. But the main disadvantage is that coherency of image is not guarenteed. Holes and extragenous pixels may appear in the segmented image. To preserve the coherency of image, we propose euler number based technique to find the threshold T. Although euler number based thresholding was applied for real time applications[18], it is not used for image thresholding. The Euler number of an image is an important feature that can be used to describe the topological structure of that image[19]. It is known that this describing feature is invariant up to several image transformations such as translations, rotations, scale changes, affinities, projections and even some non-linear transformations such as deformation of the shapes contained in the image. Mathematically, the Euler number of a binary image can be calculated either by using global computation or by local computations. The following equation is used for computing euler number globally.

$$E = C - H \qquad (5)$$

where C is the number of regions of the image (number of connected components of the object) and H is the number of holes in the image (isolated regions of the imageâĿ™s background). Euler number E of an binary image can be calculated using local computations.

$$E(t) = \frac{1}{4}[q_1(t) - q_3(t) - 2q_d(t)] \qquad (6)$$

where t is the threshold value which is used to obtain the binary image from a gray level image, q_1 denotes the number of 2x2 matrices in the image with one 1 and remaining 0's. There are four different possible matices which count as q_1 ·q_3 denotes the number of 2x2 matrices in the image with three 1's and remaining one 0. There are four such different possible matices which could be counted as q_3· q_d denotes the number of diagonal 2x2 matices. There are two different possible matrices of qd type. Equations 3.5 and 3.6 are expected to give same Euler number E for a given binary image. For chest X-ray, it is expected to seperate two lung regions from the given image using segmentation technique. Therefore, Lung field segmentation using Boundary Map and Snake Segmentation Algorithm 5 the expected euler number is 2, since the expected number of connected components are 2 and the expected number of holes are 0. Using equation 3.5, Euler number E can be calculated in the following manner.

$$E = C - H = 2 - 0 = 2 \qquad (7)$$

Since equations 3.5 and 3.6 are expected to give the same Euler number E for a given binary image, Equation 3.6 can be made equal to 2.

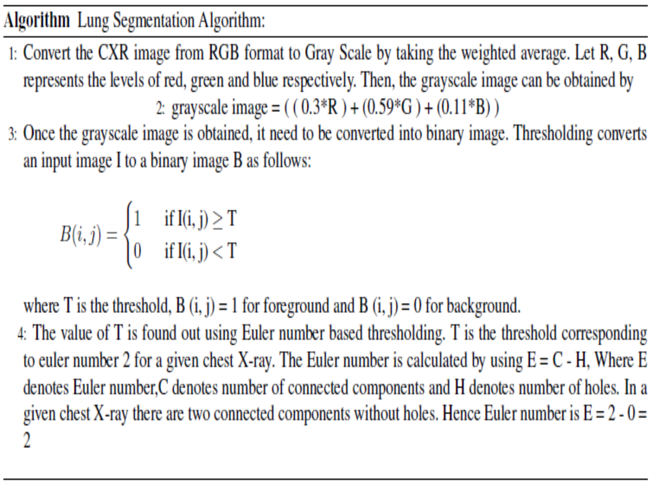
$$1/4[q_1(t) - q_3(t) - 2q_d(t)] = 2 \qquad (8)$$

From equation 3.8, required threshold is the threshold th at which the Euler number becomes 2 i.e.

$$E(t_h) = 2 \qquad (9)$$

Let us assume that T_h is a set of all thresholds t_h for which equation 3.9 is true. There are two possible cases: T_h may be a singleton set or T_h may contain multiple values. It has been proved that the graph containing different threshold values on X-axis and corresponding euler numbers on Y-axis for a given image is decaying exponential. Hence, for a given euler number, a corresponding threshold value can be found. With this observation, second case is ruled out and T_h is a singleton set and contains single value which is the required threshold value.

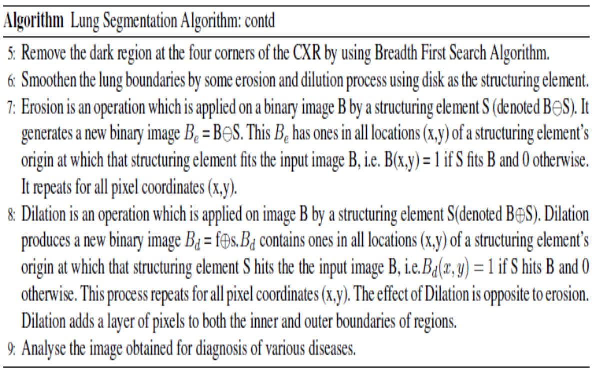
Algorithm Lung Segmentation Algorithm:



Algorithm Lung Segmentation Algorithm:

```
1: Convert the CXR image from RGB format to Gray Scale by taking the weighted average. Let R, G, B
   represents the levels of red, green and blue respectively. Then, the grayscale image can be obtained by
2: grayscale image = ( ( 0.3*R ) + (0.59*G ) + (0.11*B) )
3: Once the grayscale image is obtained, it need to be converted into binary image. Thresholding converts
   an input image I to a binary image B as follows:
   $B(i,j) = \begin{cases} 1 & \text{if } I(i,j) \geq T \\ 0 & \text{if } I(i,j) < T \end{cases}$
   where T is the threshold, B (i, j) = 1 for foreground and B (i, j) = 0 for background.
4: The value of T is found out using Euler number based thresholding. T is the threshold corresponding
   to euler number 2 for a given chest X-ray. The Euler number is calculated by using E = C - H, Where E
   denotes Euler number,C denotes number of connected components and H denotes number of holes. In a
   given chest X-ray there are two connected components without holes. Hence Euler number is E = 2 - 0 =
   2
```

6 J Ebenezer st al.



```
Algorithm Lung Segmentation Algorithm: contd
Remove the dark region at the four corners of the CXR by using Breadth First Search Algorithm.
Smoothen the lung boundaries by some erosion and dilution process using disk as the structuring element.
Erosion is an operation which is applied on a binary image B by a structuring element S (denoted B⊖S). It
    generates a new binary image $B_e$ = B⊖S. This $B_e$ has ones in all locations (x,y) of a structuring element's
    origin at which that structuring element fits the input image B, i.e. B(x,y) = 1 if S fits B and 0 otherwise.
    It repeats for all pixel coordinates (x,y).
Dilation is an operation which is applied on image B by a structuring element S(denoted B⊕S). Dilation
    produces a new binary image $B_d$ = f⊕s.$B_d$ contains ones in all locations (x,y) of a structuring element's
    origin at which that structuring element S hits the the input image B, i.e.$B_d(x,y) = 1$ if S hits B and 0
    otherwise. This process repeats for all pixel coordinates (x,y). The effect of Dilation is opposite to erosion.
    Dilation adds a layer of pixels to both the inner and outer boundaries of regions.
Analyse the image obtained for diagnosis of various diseases.
```

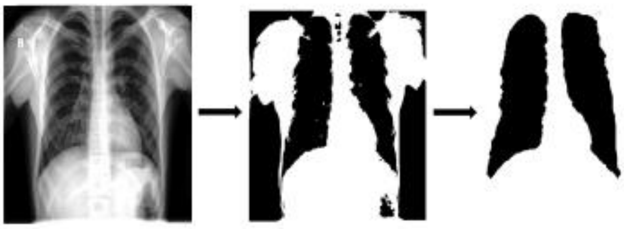


Figure 2. Lungs are segmented from the preprocessed image. First image is the preprocessed image. Second image is obtained using Euler numbers. Third one is the final image obtained by further filtration

3.3 Measuring Cardiothoracic Ratio (CTR)

For calculating the cardiothoracic ratio, first cardiac diameter and thoracic diameter needs to be computed. The maximum transverse cardiac diameter (CD) can be represented as a sum of MRD (greatest perpendicular diameter from midline to right heart border) and MLD (greatest perpendicular diameter from midline to left heart border).

$$CD = MRD + MLD \quad (10)$$

Thoracic diameter (TD) is the widest distance between the internal surfaces of the ribs on the left side and the right side. The Cardiothoracic Ratio (CTR) can then be calculated as

$$CTR = \frac{CD}{TD} \quad (11)$$

Lung field segmentation using Boundary Map and Snake Segmentation Algorithm 7

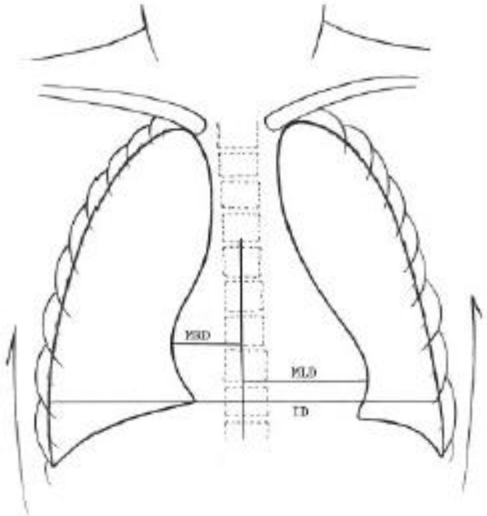
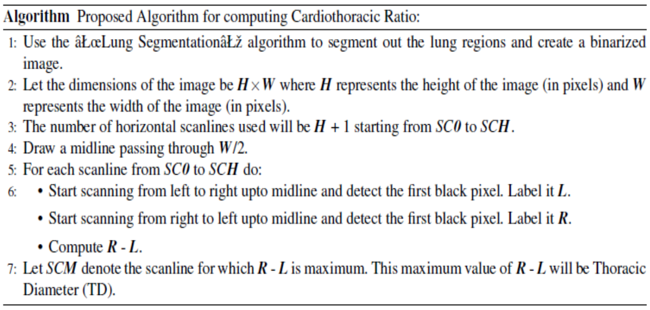
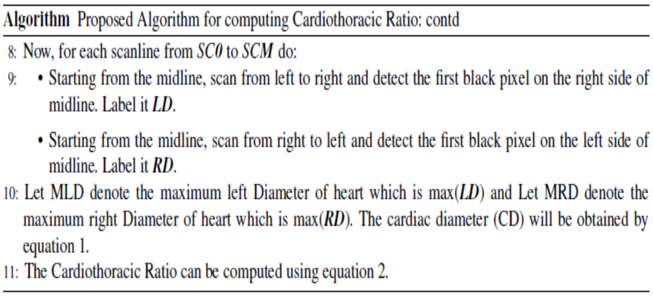


Figure 3. Figure depicting the MRD (maximum right diameter), MLD (maximum left diameter) and TD (thoracic diameter).



```
Algorithm Proposed Algorithm for computing Cardiothoracic Ratio:
Use the âŁœLung SegmentationâŁž algorithm to segment out the lung regions and create a binarized
    image.
Let the dimensions of the image be $H \times W$ where $H$ represents the height of the image (in pixels) and $W$
    represents the width of the image (in pixels).
The number of horizontal scanlines used will be $H + 1$ starting from $SC0$ to $SCH$.
Draw a midline passing through $W/2$.
For each scanline from $SC0$ to $SCH$ do:
    • Start scanning from left to right upto midline and detect the first black pixel. Label it $L$.
    • Start scanning from right to left upto midline and detect the first black pixel. Label it $R$.
    • Compute $R$ - $L$.
Let $SCM$ denote the scanline for which $R$ - $L$ is maximum. This maximum value of $R$ - $L$ will be Thoracic
    Diameter (TD).
```



```
Algorithm Proposed Algorithm for computing Cardiothoracic Ratio: contd
Now, for each scanline from $SC0$ to $SCM$ do:
    • Starting from the midline, scan from left to right and detect the first black pixel on the right side of
      midline. Label it $LD$.
    • Starting from the midline, scan from right to left and detect the first black pixel on the left side of
      midline. Label it $RD$.
Let MLD denote the maximum left Diameter of heart which is max($LD$) and Let MRD denote the
    maximum right Diameter of heart which is max($RD$). The cardiac diameter (CD) will be obtained by
    equation 1.
The Cardiothoracic Ratio can be computed using equation 2.
```

The prevalence of Cardiomegaly can be detected if the Cardiothoracic ratio is ≥ 0.5. Nakamori et al. [1] have emphasised on the usefulness of Cardiothoracic Ratio in detection of Cardiomegaly. They have shown in their work that accuracy of CTR in detection of Cardiomegaly is 95.8 %. Studies [13, 14] have shown that CTR can be influenced by many cardiac and extracardiac factors. Some of the factors that influence CTR are the examination technique, the patient's biotype, the patient's physiological status, thoracic alterations, the size of the lungs, the breathing phase, the cardiac cycle phase, and heart rate at the time of examination. But the classic criterion of 0.5 has been adopted as the most appropriate value for Cardiomegaly predictions.

Figure 4 and Figure 5 illustrate the difference between CXR of a normal heart and enlarged heart.

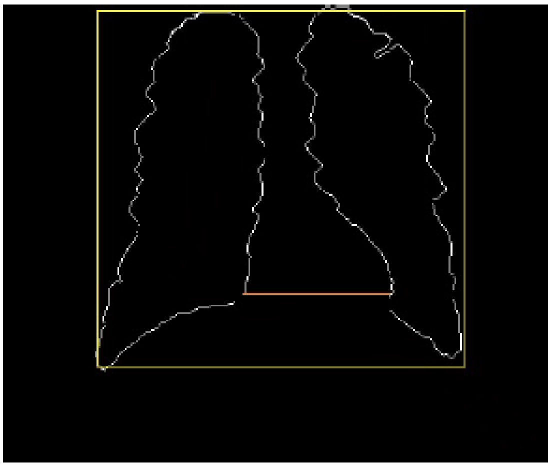


Figure 4. CXR of a Normal Heart.

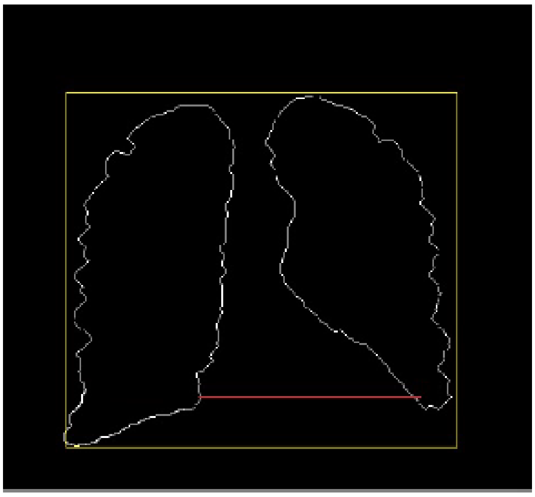


Figure 5. CXR of a Enlarged Heart.

Lung field segmentation using Boundary Map and Snake Segmentation Algorithm 9

IV. EXPERIMENTAL RESULTS

4.1 MedIT : A Tool to detect Cardiomegaly

Based on the above mentioned algorithms, we implemented it in matlab and designed a software named MedIT. Snapshots of the software are shown below.

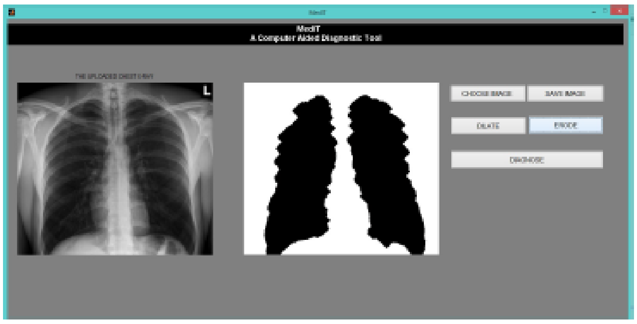
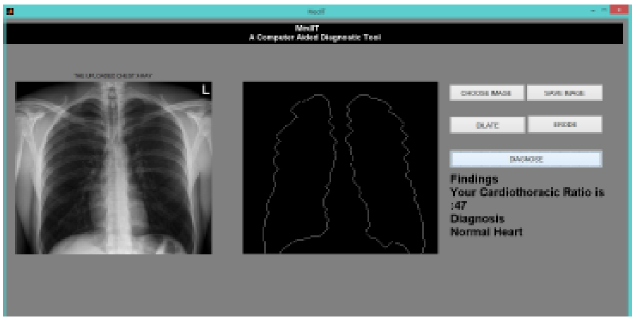


Figure 6. Lung segmentation from the CXR image.

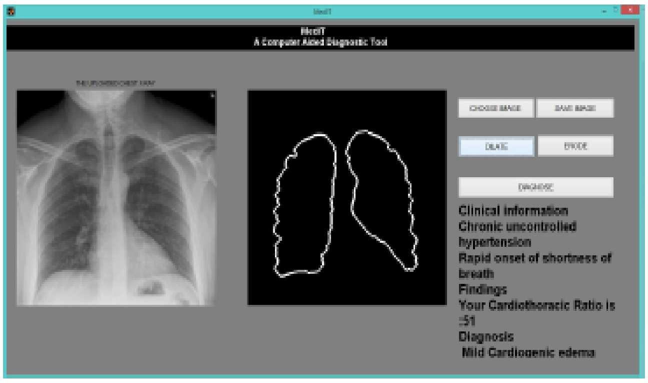
Further fine tuning of the morphed image can be done using Dilate and Erode option buttons provided in the software. Dilation and Erosion are the most basic morphological operators. Pixels are added to the boundaries of objects in an image using Dilation whereas pixels are removed from object boundaries using Erosion.



Findings
Your Cardiothoracic Ratio is :47
Diagnosis
Normal Heart

Figure 7. Analysing the CXR image of a Normal heart sample.

10 J Ebenezer et al.



Clinical information
Chronic uncontrolled hypertension
Rapid onset of shortness of breath
Findings
Your Cardiothoracic Ratio is :51
Diagnosis
Mild Cardiogenic edema

Figure 8. Analysing the CXR image of a Diseased heart sample.

4.2 Chest X-Ray Image Dataset

The proposed Cardiothoracic Ratio computation algorithm is evaluated using two different CXR datasets.

1. JSRT DataSet: This dataset was compiled by the Japanese Society of Radiological Technology (JSRT) [15]. The set contains 247 chest X-rays. Among 247 images, 154 have normal lungs and 93 have abnormal lungs. All X-ray images have a size of 20482048 pixels and a gray-scale color depth of 12 bit.

2. India Set: This dataset contains 100 chest X-rays which are collected from a private clinic in India with resolutions of 19201080 . The gray-scale color depth is 12 bit. The dataset contains 50 CXR images of normal patients and rest 50 of those who are medically diagnosed by doctors as patients of cardiomegaly. The dataset contains CXR images of both male and female patients of age ranging from 20 to 65 years.

4.3 Cardiothoracic ratio(CTR) computation

We examined the CXR images from the above mentioned datasets using MedIT. Each image is individually selected for testing, morphological operations namely Dilation and Erosion are applied and then finally assessed for Cardiomegaly. Table1 shows the results of computation of Cardiothoracic Ratio of 10 patients and Figure 7 depicts the input and output chest X-ray images of these patients.

In Table1, the CXR X5, X7 and X8 are having Cardiothoracic ratio greater than 50% .Our system has identified the prevalence of Cardiomegaly in these patients. The average computation time for lung segmentation is 0.82 seconds and the average computation time for calculating Cardiothoracic Ratio as per our proposed algorithm is 0.0015 seconds. In literature, many have presented good alorithms for computing Cardiothoracic Ratio but our proposed algorithm has achieved the best results so far.

Table 1. Computation time to measure CTR.

CXR	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Cardiothoracic Ratio(CTR)	42.1	43.2	40.3	42.8	53.9	40.1	51.5	53.4	42.8	38.8
Computation time for Lung Segmentation (in sec)	0.86	0.84	0.81	0.84	0.91	0.79	0.82	0.77	0.76	0.72
Computation time for CTR calculation (in sec)	0.0014	0.0020	0.0011	0.0010	0.0019	0.0011	0.0013	0.0017	0.0018	0.0022

Lung field segmentation using Boundary Map and Snake Segmentation Algorithm 11

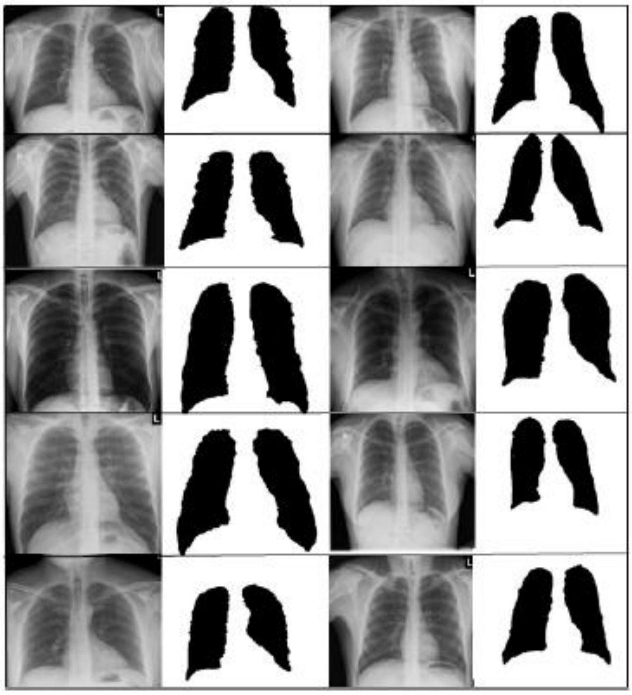


Figure 9. Depicting the input CXR images and their corresponding morphed images.

4.4 Test for Accuracy

If a system can differentiate the patient and healthy cases correctly, then the system is said to have high efficiency. In order to approximate the accuracy of a system, the proportion of true positive and true negative should be calculated in all evaluated cases. This can be stated mathematically as :

$$Accuracy = \frac{Tp + Tn}{Tp + Tn + Fn + Fp} \qquad (12)$$

where

True positive (T_P) denotes the number of cases which are correctly identified as patient

False positive (F_P) denotes the number of cases which are incorrectly identified as patient

True negative(T_n) denotes the number of cases which are correctly identified as healthy

False negative (F_n) denotes the number of cases which are incorrectly identified as healthy

We have tested total 347 Chest X-ray images out of which 143 were of patients suffering from cardiomegaly and rest were of persons with normal cardiac size. Our system has correctly identified 138 out of 143 as patients of cardiomegaly and rest 5 have been incorrectly identified as healthy. This means TP = 138 and FN = 5 . Out of 204 normal X-rays, we have detected 196 X-rays as normal and rest 8 have been incorrectly identified as unhealthy. This means TN = 196 and FP = 8. So, our system was able to achieve an accuracy of 96.25% with a positive error of 3.92% and a negative error of 3.49%. Positive error occurs when the system incorrectly predicts normal as abnormal due to overestimation in calculation and negative error occurs when system incorrectly predicts abnormal cardiac conditions as normal.

V. CONCLUSION

This paper presents a new method for computer assisted CXR analysis for diagnosing Cardiomegaly. Segmentation of the image is done using euler number based thresholding. An algorithm for Cardiothoracic Ratio(CTR) computation is developed and implemented. The experimental results obtained with the proposed algorithm are very encouraging. The proposed method outperforms the existing CTR computation algoritms on the chest x-ray images of different categories. We have achieved an overall accuracy of 96.25% and the overall(lung segmentation time + CTR computation time) average computation of 0.8215 seconds.

VI. REFERENCES

[1]. N. Nakamori, K. Doi, H. MacMAHON, Y. Sasaki, and S. Montner, "Effect of heart-size parameters computed from digital chest radiographs on detection of cardiomegaly: Potential usefulness for computer-aided diagnosis.," Investigative radiology, vol. 26, no. 6, pp. 546–550, 1991.

[2]. M.-H. Chen and P.-F. Yan, "A fast algorithm to calculate the euler number for binary images," Pattern Recognition Letters, vol. 8, no. 5, pp. 295–297, 1988.

[3]. S. B. Gray, "Local properties of binary images in two dimensions," IEEE Transactions on Computers, vol. 100, no. 5, pp. 551–561, 1971.

[4]. L. Wong and H. Ewe, "A study of lung cancer detection using chest x-ray images," in Proc. 3rd APT Telemedicine Workshop, Kuala Lumpur, vol. 3, pp. 210–214, 2005.

[5]. B. Van Ginneken, B. T. H. Romeny, and M. A. Viergever, "Computer-aided diagnosis in chest radiography: a survey," IEEE Transactions on medical imaging, vol. 20, no. 12, pp. 1228–1241, 2001.

[6]. H. Becker, W. Nettleton, P. Meyers, J. Sweeney, and C. Nice, "Digital computer determination of a medical diagnostic index directly from chest x-ray images," IEEE Transactions on Biomedical Engineering, no. 3, pp. 67–72, 1964.

[7]. D. Hall, G. Lodwick, R. Kruger, S. Dwyer, and J. Townes, "Direct computer diagnosis of rheumatic heart disease 1," Radiology, vol. 101, no. 3, pp. 497–509, 1971.

[8]. R. P. Kruger, J. R.Townes, D. L. Hall, S. J. Dwyer, and G. S. Lodwick, "Automated radiographic diagnosis via feature extraction and classification of cardiac size and shape descriptors," IEEE Transactions on Biomedical Engineering, no. 3, pp. 174–186, 1972.

[9]. N. Sezaki and K. Ukena, "Automatic computation of the cardiothoracic ratio with application to mass screening," IEEE Transactions on Biomedical Engineering, no. 4, pp. 248–253, 1973.

[10]. A. H. Dallal, C. Agarwal, M. R. Arbabshirani, A. Patel, and G. Moore, "Automatic estimation of heart boundaries and cardiothoracic ratio from chest x-ray images," in SPIE Medical Imaging, pp. 101340K– 101340K, International Society for Optics and Photonics, 2017.

[11]. L. Cong, L. Jiang, G. Chen, and Q. Li, "Fully automated calculation of cardiothoracic ratio in digital chest radiographs," in SPIE Medical Imaging, pp. 1013432–1013432, International Society for Optics and Photonics, 2017. Lung field segmentation using Boundary Map and Snake Segmentation Algorithm 13

[12]. H. MacMahon, K. Doi, H.-P. Chan, M. L. Giger, S. Katsuragawa, and N. Nakamori, "Computer-aided diagnosis in chest radiology," Journal of thoracic imaging, vol. 5, no. 1, pp. 67–76, 1990.

[13]. K. Nickol and A.Wade, "Radiographic heart size and cardiothoracic ratio in three ethnic groups: a basis for a simple screening test for cardiac enlargement in men," The British journal of radiology, vol. 55, no. 654, pp. 399–403, 1982.

[14]. Y. Mensah, K. Mensah, S. Asiamah, H. Gbadamosi, E. Idun, W. Brakohiapa, and A. Oddoye, "Establishing the cardiothoracic ratio using chest radiographs in an indigenous ghanaian population: a simple tool for cardiomegaly screening," Ghana medical journal, vol. 49, no. 3, pp. 159–164, 2015.

[15]. J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T.Kobayashi, K.-i.Komatsu, M. Matsui, H. Fujita, Y.Kodera, and K. Doi, "Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," American Journal of Roentgenology, vol. 174, no. 1, pp. 71–74, 2000.

[16]. Esmail, Hanif and Oni, Tolu and Thienemann, Friedrich and Omar-Davies, Nashreen and Wilkinson, Robert J and Ntsekhe, Mpiko,"Cardio-thoracic ratio is stable, reproducible and has potential as a screening tool for HIV-1 related cardiac disorders in resource poor settings," Public Library of Science, vol. 11,no. 10,pp. 63-49,2016.

[17]. Candemir, Sema and Jaeger, Stefan and Lin, Wilson and Xue, Zhiyun and Antani, Sameer and Thoma, George, " Automatic heart localization and radiographic index computation in chest x-rays," Proc. of SPIE Vol, vol. 9785, pp. 1-17, 2016.

[18]. Lakhani, Paras and Sundaram, Baskaran, "Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks," Radiological Society of North America,pp. 16-26,2017.

[19]. He, Li-Feng and Chao, Yu-Yan and Suzuki, Kenji,"An algorithm for connected-component labeling, hole labeling and Euler number computing", Journal of Computer Science and Technology Springer,vol. 28,no. 3,pp. 468-478, 2013.

International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Synchronous and Asynchronous Replication

Katembo Kituta Ezechiel[1], Dr. Ruchi Agarwal[2], Dr. Baijnath Kaushik[3]
[1]CSE-SET, Sharda University, Greater Noida,Uttar Pradesh, India
[2]JIMS Engineering Management Technical Campus, Greater Noida, Uttar Pradesh, India
[3]Shri Mata Vaishno Devi University, Jammu and Kashmir, India

ABSTRACT

The objective of this study is to analyse the various technical aspects of synchronization between databases while replication is occurred. To achieve this feat, we embarked in a systematic review of the literature around this theme, an examination that revealed that synchronization is not done with more than one Peer at the time. Therefore, we have been motivated to develop an algorithm which supports the synchronization with more than one Peer.

Background and Objective: This paper presents a systematic review of the two replication approaches, namely: synchronous replication and asynchronous replication. The purpose of this study is to analyse the different technical aspects of synchronization between databases during replication. Methodology: To achieve this, we used the documentary method that allowed us to collect the necessary documents containing the literature that fits with this research and that allowed us to conduct this review. Apart from this one, algorithmic helped us to develop a model that supports synchronization with more than one peer. Result: Our long-awaited result being an algorithm, we have similarly taken care to present the fruits of the efforts of our predecessors who would have proceeded in the same way as us. Conclusion: Finally, we found that it is necessary to have a synchronization algorithm independently of the DBMSs that can be used by the designers of distributed databases.

Keywords : Algorithm, Synchronization, Replication, Peer, Database.

I. INTRODUCTION

In computer science, a distributed database is a database in which storage devices are not all attached to a common Processing Unit Such (CPU). It may be stored in multiple computers, located in the same physical location; or may be dispersed over a network of interconnected computers [1].

The design of a distributed database requires that it be entirely resident on various sites in a computer network or its portion. In this logic, there must be at least two sites hosting the database and not necessarily each site in the network. The strategies can be broadly divided into replication and fragmentation [2]. However, in most cases, a combination of the two is used. But, as far as this work is concerned, our interest is more fixed on replication of data.

Replication is a set of technologies for copying and distributing data and database objects from one database to another and then synchronizing between databases to maintain consistency [3]. It stores separate copies of the database at two or more sites. It is a popular fault tolerance technique of distributed databases [4].

From this point of view, we retain two conceptions which we will first of all clear up: synchronization and consistency.

Data synchronization is the process of establishing consistency among data from a source to target data storage and vice versa, and the continuous harmonization of the data over time. Data synchronization technologies are designed to synchronize a single set of data between two or more devices, automatically copying changes back and forth [5].

In turn, the consistency of Replication models is essential to abstract away execution particulars, and to classify the functionality of a given system. Also a consistency model is a method for come to a joint considerate of each other’s rights and responsibilities [6].

II. TYPES OF REPLICATION

There are two types of replication which are as follows [7]:

- Synchronous Replication: All copies of a modified relation (fragment) must be updated before commit. Here, the most up to date value of an item is guaranteed to the end user. There are two different methods of synchronous replication:
  1) Read-Any, Write-All: This method is beneficial in case well when reads are much more frequent than writes.
     1. Read-Any: when reading an item, access any of the replicas.
     2. Write-All: when writing an item, must update all of the replicas.
  2) Voting:
     1. When writing, update some fraction of the replicas.
     2. When reading, read enough copies to ensure you get at least one copy of the most recent value.
        Use a version number to determine which value is most recent the copies "vote" on the value of the item
- Asynchronous Replication: Asynchronous replication allows different copies of the same object to have different values for short periods of time. Data is updated after a predefined interval of time.
  1) Primary Site: In primary-site replication, one copy of data is assigned as the master copy. Updation of data is possible only with in the master copy. The secondary copies of data can only be read. Changes to the master are periodically propagated to the secondary copies.
  2) Peer-to-Peer: In peer-to-peer replication, more than one replica is updatable. In addition, a conflict resolution strategy must be used to deal with conflicting changes made at different sites.

III. LITERATURE REVIEW

We are not the first to direct our thoughts on the problem of data replication and its techniques. It is therefore essential for us to review the literature that fits in with this theme in order to justify our research.

Chaturvedi & Prof. Jain [8] focuses on replication in distributed file systems. They present replication as a strategic key in distributed systems for enhancing accuracy, availability, and performance. They gave also a brief introduction to replication and various algorithms have been discussed and a detailed study has been performed. Although they concluded that optimistic replication systems have been in use for some time. Client - Server model is generally used but when dealing with mobile, peer-to-peer model is used; this allows direct communication and synchronization between all the peers but has scalability problems.

Gudakesa et al. tackled the data synchronization method usable in the two-ways of data synchronization (Master-Slave / Slave-Master) [9]. They indicate that synchronization of data is an integral part of replication in that it ensures the reliability and similarity between copies of the database and thus the objects and data. They added that to synchronize the data several methods are applicable; so they presented the one that uses the audit log that records all the activities that occur to the database. In this context, they have shown that this technique can be applied to almost all database management systems (DBMS). In the end, they also targeted possible problems that can occur when synchronization is running and how to solve this problem.

Kaushik B. et al [23-25] developed a neural network based model to evaluate the performance of reliability evaluation in complex and high performance network model. Agrawal et al. proposed taxonomy for partitioned replicated database systems that reside in the cloud [10]. Taxonomy is a practice derived from the need for scientific classification of species. Indeed, the authors of this article have focused on databases that support both transactions and replication. They used several advanced systems to illustrate specific instances of taxonomy. Thus, they quickly realized that taxonomy is distorted, in the sense that there are several subcategories of replicated transaction systems, but only one category of replicated object systems. Finally, they concluded that taxonomy represents a first attempt at principle to understand and classify the transaction mechanisms of many proposed systems of the state of the art.

Krishna et al. advocated producing a cloud Operating System in which the user can upload file from mobile or Personal Computer to the cloud storage [11]. They have implemented a Windows platform on which a user's files can be automatically synchronized with his devices. Thus the user can view his files anywhere; in analogy with the existing systems, they were able to demonstrate that here the files were to be downloaded manually. On finish, automatic data synchronization between devices of the user was studied.

Ranjan & Agarwal [22] analysed the customer transactional database and try to understand the true value of the customer.

Agarwal.et al [19] reviewed parallel apriori algorithm on mad reduce framework for performance enhancement by dividing the transactional database into data chunks and distributes them among different machines. Ranjan and Agarwal [21] also analysed the database of a retail firm and found the behaviour of purchased products using classification and Regression Modelling.

Kelemu & Prof. Patil [12] defined a strategy which combines together horizontal fragmentation and vertical fragmentation of the relational database table in distributed system. This strategy is hybrid fragmentation. The subspace grouping algorithm on which hybrid based fragmentation is an incomparable advance in the application of partitioning data with the tuples (rows) and attributes (columns) of a database relation (table). They referred to project clustering as an advance that uses the sub-space grouping algorithm. Thus, it has been developed a new algorithm and it has been implement and also analysis with the random algorithm. This approach has improved the distribution of relations of databases containing large tuples and attributes.

Agarwal & Ranjan [20] developed a model to measure and manage the performance of a retail sector using multiple databases. Salunke & Potdar presented some static and dynamic database partitioning techniques to support mission critical databases [13].

Following techniques have been discussed:

- Partitioning based on selectivity;
- A greedy algorithm to select the "best" dimension tables of a star schema;
- Reloading in a main memory database system supported by database partitioning techniques;
- The distribution of a database based on a grouping approach and a genetic algorithm;
- A dynamic vertical partitioning approach for the distributed database system.

Hiremath & Dr. Kishor [14] studied the various problems areas and advantages and disadvantages of Distributed Database . They showed that Distributed database would allow end user to create and store data anywhere in the network where database is situated. In this approach, while storing and accessing data from distributed database through computer network, there are various problems that occurs such deadlock, concurrency and data allocation using fragmentation and replication. To manage these problems, it is necessary to design the distributed database carefully manner. Apart that they have presented some distributed database architecture strategies:

- Top-Down Approach: mostly used when we have to implement distributed database from beginning;
- Bottom-up Approach: This type of Approach is used when distributed database already exists and we have to add another database in existing environment.

Akshay &Yogesh [15] reviewed the security issues and concurrency control in distributed database system and data security aspects of client/server architecture. Here it was pointed out that the most important problem is the security that could occur and potentially compromises access control and system integrity. For example, a solution has been proposed for certain security aspects such as multilevel access control, privacy, reliability, integrity, and recovery for a distributed database system. In addition, some competitive control algorithms were examined, for example: the basic timestamp algorithm, the distributed two-phase lock protocol (2PL), the distributed optimized protocol and the wound-wait algorithm.

Imam et al. proposed systematic literature review on Data synchronization between mobile devices and server-side databases [16]. The objective of this study was to examine the state of the art in various aspects of mobile devices with regard to data sharing and synchronization between mobile device databases and server-side databases, opposing to server-server synchronization. Relevant literature was chosen for examinations and several aspects were identified as being directly or indirectly involved in synchronization. The issues that have been discussed are the synchronization of data between mobile device and server databases, cloud-based solutions, data inconsistency, conflict resolution strategies, data processing, dependency Suppliers, summary of messages and algorithms used and common tools adopted.

Shabani et al. have designed an algorithm for data synchronization based on Web Services (WS), which allows software applications to work well on both configurations "Online" and “Offline”, in the absence of the network [17]. In this study, it has been shown that the use of synchronization is a major challenge for the institutions in general because the agents or personnel (users) are less aware of the network failures and the system works without any involuntary interruption. In particular, the confidence of the administrative and academic staff of the University of Prishtina is increasing as there is no waiting for documents because the network fails. The results of this synchronization set up are judged to be positive because there is a reliability which is observed in the user software applications.

Fadoua & Amel established a standard synchronization process between different sites of a distributed database architecture including database heterogeneity, variable synchronization delays, network capability restrictions and fault management ability [18]. To go straight to the goal, apart from the introduction and the conclusion, it has been:

- Described the different approaches and their limitations;
- Presented the suggested synchronization protocol;
- Detailed the data serialization and the deserialization mechanism regarding network bandwidth consumption and packet preparation time impact;
- Studied the protocol performance on the most optimistic and pessimistic scenarios;
- Reported the experimental tests of the implemented example.

IV. METHODOLOGY

After the course of the literature above offered by our predecessors in this field, we realized however that all of these possibilities for synchronization are being making over a centralized Peer-to-Peer Architecture because they require a central server. Thus, we notice that existing synchronizers are limited because either they require a synchronization of not more than two copies at a time.

It would be interesting to set up another collaborative environment to remove such limitations in such a way that synchronization is now done with several copies. We approach in this direction because visibly centralized Peer-to-Peer architecture also presents a major problem that is such that it only offers a single gateway, its central server, which is the Achilles heel of everything over the network. It would be enough that this server knew a breakdown to block or to disconnect all the users and to stop the operation of the whole system because no peer will have the data no longer updated.

Here we present the methodology of the work we used to achieve our goals. So, for our case, our gaze remains focused on:

- The documentary method, a set of steps to search, identify and find documents related to a subject by developing a research strategy. It helped us to collect and analyse the documents that we used to identify the shortcomings of existing synchronizers.
- The Algorithmic as method or technique helped us to design the algorithm of instructions and steps of a Pure Peer-to-Peer Synchronizer.

V. RESULT

This section presents the models collected from the documents of our predecessors in relation to the objectives that this work has set itself. Then it analyses the defined methodology and then presents the discussion of the results obtained.

Review of the literature around this theme revealed that synchronization is not done with more than one peer at a time. Some synchronization models, results of some works, are presented in Figure 1. and Figure 2.

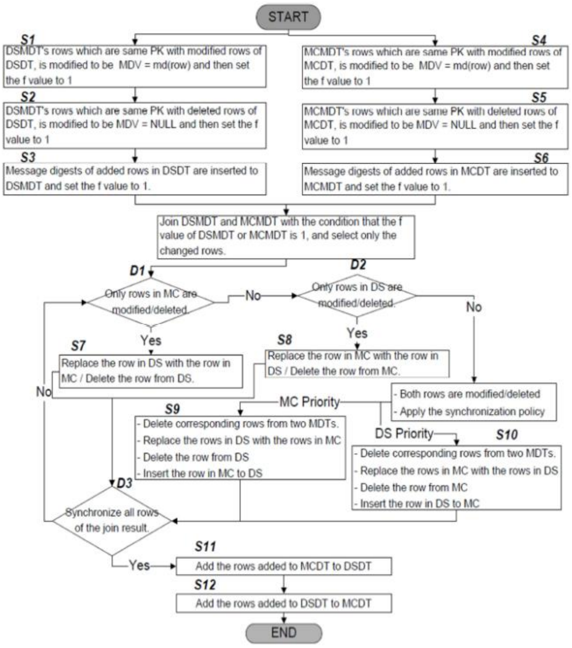

START
S1
DSMDT's rows which are same PK with modified rows of DSDT, is modified to be MDV = md(row) and then set the f value to 1
S2
DSMDT's rows which are same PK with deleted rows of DSDT, is modified to be MDV = NULL and then set the f value to 1
S3
Message digests of added rows in DSDT are inserted to DSMDT and set the f value to 1.
S4
MCMDT's rows which are same PK with modified rows of MCDT, is modified to be MDV = md(row) and then set the f value to 1
S5
MCMDT's rows which are same PK with deleted rows of MCDT, is modified to be MDV = NULL and then set the f value to 1
S6
Message digests of added rows in MCDT are inserted to MCMDT and set the f value to 1.
Join DSMDT and MCMDT with the condition that the f value of DSMDT or MCMDT is 1, and select only the changed rows.
D1
Only rows in MC are modified/deleted.
No
D2
Only rows in DS are modified/deleted.
No
Yes
S7
Replace the row in DS with the row in MC / Delete the row from DS.
Yes
S8
Replace the row in MC with the row in DS / Delete the row from MC.
No
- Both rows are modified/deleted
- Apply the synchronization policy
MC Priority
S9
- Delete corresponding rows from two MDTs.
- Replace the rows in DS with the rows in MC
- Delete the row from DS
- Insert the row in MC to DS
DS Priority
S10
- Delete corresponding rows from two MDTs.
- Replace the rows in MC with the rows in DS
- Delete the row from MC
- Insert the row in DS to MC
D3
Synchronize all rows of the join result.
Yes
S11
Add the rows added to MCDT to DSDT
S12
Add the rows added to DSDT to MCDT
END

Figure 1. Synchronization Algorithm [16]

DS: Database Server

DSMDT: Database Server Mobile Digest Table

MDV: Message Digest Value

PK: Primary Key

S1 to S12: Step 1 to Step12

D1 to D3: Decision 1 to Decision 3

DSDT: Database Server Data Table

MC: Mobile Client

MCDT: Mobile Client Data Table

f: Flag

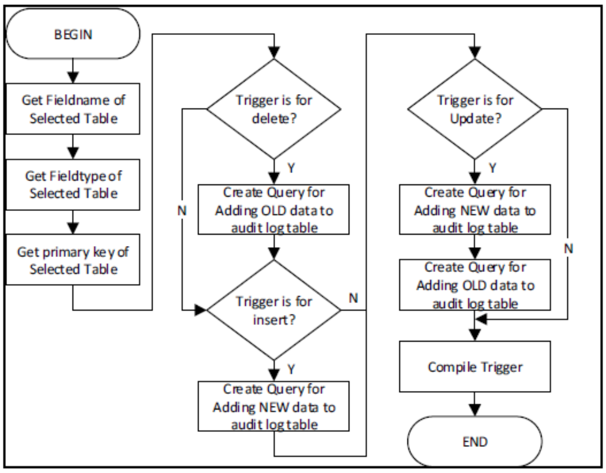

BEGIN
Get Fieldname of Selected Table
Get Fieldtype of Selected Table
Get primary key of Selected Table
Trigger is for delete?
Y
N
Create Query for Adding OLD data to audit log table
Trigger is for insert?
N
Y
Create Query for Adding NEW data to audit log table
Trigger is for Update?
Y
Create Query for Adding NEW data to audit log table
N
Create Query for Adding OLD data to audit log table
Compile Trigger
END

Figure 2. How Audit Log Trigger Created [9]

This is used when it's mandatory to keep the same copy of the data in two data storage node on the network. It is Two-ways synchronization used Audit log mechanisms [9].

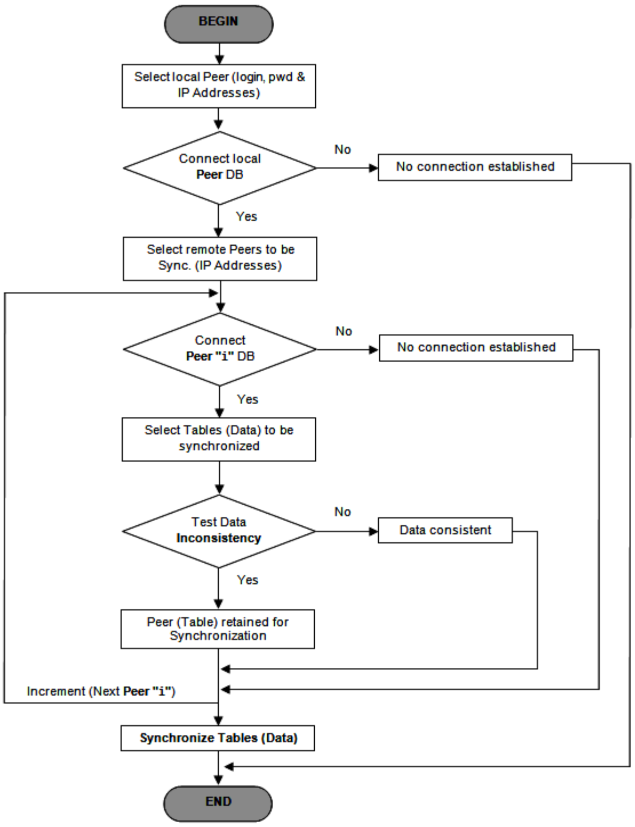

BEGIN
Select local Peer (login, pwd & IP Addresses)
Connect local Peer DB
No
No connection established
Yes
Select remote Peers to be Sync. (IP Addresses)
Connect Peer "i" DB
No
No connection established
Yes
Select Tables (Data) to be synchronized
Test Data Inconsistency
No
Data consistent
Yes
Peer (Table) retained for Synchronization
Increment (Next Peer "i")
Synchronize Tables (Data)
END

Figure 3. Algorithm for a Pure Peer-to-Peer synchronization

Therefore, it has been necessary to design a synchronization algorithm that can update several copies of the databases at the same time. Each machine in its roles is identical to another, and then we will call this type of system Pure Peer-to-Peer Synchronizer.

As shown in Figure 3 the steps of the algorithm for a Pure Peer-to-Peer synchronization are given below:

Step1. Select the local Peer and connect on it by providing the login, password and the IP address (facultative): if these provided parameters are incorrect then no connection established else next step;

Step2. Select remote Peers, by indicating theirs IP addresses, to be Sync and test connection with them one to one: if Peer non-jointed then no connection established, next Peer else next step;
Step3. Select Tables (Data) to be synchronised and test Data inconsistency: if Data consistent then next Peer else Peer (Table) retained for Synchronization, next step;
Step4. Synchronize Tables (Data) of all retained Peers at the same time.

VI. CONCLUSION

In this paper we have managed to conduct literature study on synchronization while replication is occurred. Although in relevant literature, we have realized that synchronization is not done with more than one Peer at the time; this has been our motivation to develop an algorithm which support the synchronization with more than one Peer.

This algorithm may be used by Distributed Databases and applications designers since they will need interaction between different DBMSs.

As future work we will write a complete algorithm in which it will be presented step by step scenarios to Synchronize Tables (Data) i.e. the internal reality of a Pure Peer-to-Peer synchronization algorithm.

VII. REFERENCES

[1]. Kaur, K., & Singh, H. (2016). Distributed database system on web server: A Review. International Journal of Computer Techniques, 3(6), 12-16.

[2]. Ozsu, M. T., & Valduriez, P. (2011). Principles of Distributed Database Systems (3rd ed.). New York, USA: © Springer Science+Business Media, LLC.

[3]. Microsoft. (2017). SQL Server Replication. Retrieved May 2017, from Microsoft Documentation: https://docs.microsoft.com/en-us/sql/relational-databases/replication/sql-server-replication

[4]. Truica, C., & Boicea, A. (2013). "Asynchronous Replication in Microsoft SQL Server, PostgreSQL and MySQL". International conference on cyber science and engineering.

[5]. Malhotra N., Chaudhary A., (2014), Implementation of Database Synchronization Technique between Client and Server, International Journal of Engineering and Computer Science, 3(7): 7070-7073.

[6]. Souri, A., Pashazadeh S. & Navin, A., H. (2014). "Consistency of data replication protocols in database systems: A review", International Journal on Information Theory (IJIT),3(4), 19-32.

[7]. Tomar, P., & Megha. (2014). "An Overview of Distributed Databases". International Journal of Information and Computation Technology, 4(2), 207-214.

[8]. Chaturvedi, N., & Prof. Jain, D., C., (2012), Analysis of Replication and Replication Algorithms in Distributed System, International Journal of Advanced Research in Computer Science and Software Engineerin, 2(5), 2277 128X.

[9]. Gudakesa, R., Sukarsa, I., & Sasmita, I., M., A., (2014). Two-ways database synchronization in homogeneous dbms using audit log approach, Journal of Theoretical and Applied Information Technology, 65(3), 1992-8645.

[10]. Agrawal, D., El Abbadi, A., & Salem,K., (2015). A Taxonomy of Partitioned Replicated Cloud-based Database Systems, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 38(1), 4-9.

[11]. Krishna, S., V., Gokul, K.& Kumar, S., (2012). Data Synchronization Using Cloud Storage, International Journal of Advanced Research in Computer Science and Software Engineering, 2(6), 2277 128X.

[12]. Kelemu, Y., & Prof. Patil, S., (2016). Hybridized Fragmentation of Distributed Databases using Clustering, International Journal of Engineering Trends and Technology (IJETT), 37(1), 2231-5381.

[13]. Salunke, D., T., & Potdar, G., P., (2014). A Survey Paper on Database Partitioning, International Journal of Advanced Research in Computer Science & Technology (IJARCST), 2(3), 2347 – 8446.

[14]. Hiremath, D., S., & Dr. Kishor, S., B., (2016). Distributed Database Problem areas and Approaches, Journal of Computer Engineering : National Conference on Recent Trends in Computer Science and Information Technology, 2278-8727.

[15]. Akshay, M., G., &Yogesh, R., G., (2016). Concurrency Control and Security Issue in Distributed Database System, International Journal of Engineering Development and Research, 4(2), 2321-9939.

[16]. Imam, A., A., Basri, S., Ahmad, R., (2015). Data synchronization between mobile devices and server-side databases: a systematic literature review, Journal of Theoretical and Applied Information Technology, 81(2), 1992-8645.

[17]. Shabani, I., Çiço B.,& Dika, A., (2012), Solving Problems in Software Applications through Data Synchronization in Case of Absence of the Network, International Journal of Computer Science Issues, 9(1): 1694-0814.

[18]. Fadoua, H., & Amel, G., T., (2015), Near Real-time Synchronization Approach for Heterogeneous Distributed Databases, The Seventh International Conference on Advances in Databases, Knowledge, and Data Applications, 978-1-61208-408-4.

[19]. Ruchi Agarwal, Sunny Singh, Satvik Vats, 'Review of Parallel Apriori Algorithm on Map Reduce Framework for Performance Enhancement', organized by Computer society of India(CSI) Delhi and NCR chapter at Bharti vidyapeeth educational complex, New Delhi during 2nd -5thDecember, 2015. Published in Springer Nature Singapore, Advances in Intelligent and Soft Computing (AISC) series, Scopus Indexed pp 403-411.

[20]. Ruchi Agarwal, Jayanthi Ranjan, 'An empirical investigation on association rule mining in Indian Retail Industry', International Journal of Intercultural Information Management, Inderscience Publishers, Vol. 3, No. 3, pp.226–241, November 2013.

[21]. Jayanthi Ranjan, Ruchi Agarwal, 'Advantages of Decision Trees Using Data Mining In Indian Retail Industry', Journal of Knowledge Management Practice, Vol. 11, Special Issue 1, January 2010.

[22]. Jayanthi Ranjan, Ruchi Agarwal, 'Application of segmentation in customer relationship management: a Data Mining perspective', International Journal of Electronic Customer Relationship Management, Inderscience Publishers, Scopus Indexed, Vol. 3, No 4, pp.402-414, 2009.

[23]. Kaushik B. et al., (2015), 'Performance evaluation of approximated artificial neural network (AANN) algorithm for reliability improvement', Journal of Applied Soft Computing, Elsevier, Vol. 26, pp.303-314.

[24]. Kaushik B. et al., (2013), 'Achieving Maximum Reliability in Fault Tolerant Network Design for Variable Networks' Journal of Applied Soft Computing, Elsevier, pp. 3211-3223.

[25]. Kaushik B. et al., (2012), 'Improved Approach for Maximizing Reliability in Fault Tolerant Networks', Journal of Advanced Computational Intelligence and Intelligent Informatics, Fuji Press, Japan, Vol. 17, 1, pp. 27-41.

IJSR CSEIT

www.ijsrcseit.com

विज्ञानं ब्रह्म

International Journal of Scientific Research in Computer Science, Engineering and Information Technology (International Journal Bimonthly Publication)
www.ijsrcseit.com

Published by :
TechnoScience Academy
www.technoscienceacademy.com

International Conference on Machine Learning & Computational Intelligence - 2017
Organised by
Department of Computer Science & Engineering
Shri Mata Vaishno Devi University, Katra, J&K 182320, India

Email: editor@ijsrcseit.com, ijsrcseit@gmail.com