A Performance Analysis of Face and Speech Recognition in the Video and Audio Stream using Machine Learning Classification Techniques

Chetan Sharma
M Tech Scholar
Department of Computer Science Engineering
Sagar Institute of Research and Technology Excellence, Bhopal, India

Rajdeep Singh
Assistant Professor
Department of Computer Science Engineering
Sagar Institute of Research and Technology Excellence, Bhopal, India

ABSTRACT
Biometric authentication is an emerging technology that utilizes biometric data for the purpose of person identification or recognition in security applications. A number of biometrics can be used in a person authentication system. Among the widely used biometrics, voice and face traits are most promising for pervasive application in every life, because they can be easily obtained using unobtrusive and user-friendly procedures. The low-cost audio and visual capture sensors on smart phones, laptops, and tablets has made the advantages of voice and face biometrics more outstanding compared with others. For quite a long time, the use of acoustic information alone has been a great success for speaker authentication applications. Meanwhile, the last decades or two also witnessed great advancement in face recognition technologies. Object detection and tracking is usually the first step in applications such as video surveillance. The static camera face recognition and tracking system's main purpose is to estimate the speed and distance parameters. We propose a general detection and tracking method for motion based on the visual system and using the image difference algorithm. Then recognize the person's voice to get feedback from the corresponding person. The process focuses on detecting people on stage and then completes the voice signal processing. We propose a new person recognition technology that uses face and voice fusion Compared to a single biometric recognition, and this technology can greatly improve the recognition speed. Development of security systems uses the Viola-Jones face recognition algorithm. The proposed method uses the Local Binary Pattern (LBP) as a function extraction technique to calculate local functions. Our project uses Mel Frequency Divergence Coefficient (MFCC) extraction technology for speech recognition. The extracted functions are used as input to the multi-SVM classifier to provide a gender to identify individuals and display the results. The new system can be used in various areas, such as identity verification and other potential commercial applications.

Keywords
SVM, KNN, LBP, Machine Learning, Viola Jones

1. INTRODUCTION
Biometrics is a technique of using unique features of a person to determine his identity. When a single biometric feature is used, the chance of compromise results in the growth of multimodal biometrics. With the development of science and technology, the transition from unimodel biometric (one single trait at a given an example) to multimodal biometric techniques (combos of two or more qualities) has been observed to increase safety. The most favorable biometric system is personality, durability, acceptability, collectability, and safety. However, no unique biometric identifier has all these properties. As an explanation, several biometric identifiers are used in a single organization, which is usually a known multimodal biometric system. As support, a multimodal system can use both face recognition and iris gratitude to validate people. Due to reliable or effective security explanation in security-critical submission, multimodal biometric recognition association has recently emerged in biometric society to replace established unethical systems[1]. The system works by first taking feature samples or taking digital color images for face recognition. The sample is then converted to a biometric template using a specific mathematical function. The biometric template provides a standardized, efficient, and highly differentiated format of functions, which can then be compared objectively with other templates to determine identity. Most biometric systems agree on two modes of operation. The registration mode is used to insert the template into the database. The verification mode where the template generates for the device, and then the matching element of the database of pre-registered templates searches for overview of Biometric identification systems can be used for personal verification or personal identification. Personal verification answers the following questions: "Do I claim to be me?" It determines the validity of the alleged identity by comparing the verification template with the registration template[2]. Verification requires a declaration of identity or finds the individual's registration template and compare it with the verification model. Therefore, the comparison required for verification is called a one-to-one comparison. During the verification, some knowledge of the system's identity usually provides along with the biometric identifier. These additional factors present the unique registered identity or remove biometric functions to the classification database, thus providing the relevant biometric machine presentation. In everyday life, most people who trade with us or trade with us use verification to confirm our identity[3][4]. Therefore, in the recognition system, matching is mentioned in one-to-many ways. These are two types of recognition systems: positive recognition and negative recognition. The closest neighbor is one of simplest machine learning algorithms support on control knowledge. The K-NN algorithm considers the new case / data similarity to the existing case and places the new case in the same category as the existing category. The K-NN algorithm stores all existing data or organize new data points based on connection. This

means that it can easily be categorized into drill kit categories using the K-NN algorithm when new data appears. The K-NN algorithm can be used for organization and classification, but in most cases it is used for organization problems. K-NN is a non-parametric algorithm, which means that it does not predict basic data[5][6]. It is also called non-lazy learning algorithm because it does not immediately learn from the education response, but stores the data and performs the work on the data stored during the classification. The KNN algorithm in the training field stores the data sets and classifies them into the same category as the new data when new data is received.[7-9]

Objective: The process's overall goal is to produce a classification that can pre-preview feeds in video examination or help users recognize reaction actions. To attain this goal, we have residential an automatic resolution that can perform three key functions: identify contact, monitor relationships, and characterize statement behaviors.

EXISTING METHOD: The first method is to use the video structure to gather end-to-end candidates from the shooting range among the existing methods. Then, by using MCMC technique to choose real-time boundaries from these candidates to be selected, it is possible to divide space-bound. It should be celebrated that when given the previous probability of number of intention video shows, the MCMC method can provide more accurate distribution results. Thus, in the second method of planned method, the first probability parameters were set at most select charge by multiple regression analysis (MRA).

2. RELATED WORK

With the proper implementation of the information transmission system, based on human speech processing, there are major technical problems related to the complexity of the automatic speech detector. When the sound information is distorted due to noise and others' interference in nature, even the most reliable information will be compromised. These include noise, speech, external noise, etc. Therefore, to recover performance of the audio input data, it is recommended to examine the audio and the video stream, i.e., identify the audio in the video. To solve this difficulty, organization of objects, that is, organization of common phonetic elements is the most important.

Efim V. Zatonskikh et al. This article describes the results attain when creating a complex software prototype that implements speech recognition through the lips through neural networks. This speech appreciation based on lip movement is measured a two-step biometric verification process. This article also suggests developing a neural network model based on the LSTM layer, which is the basis for speech appreciation. To train the model, we collected and organized a record of words with words from a class. To rotate reproduce words in video, a model of lips develop and update, in which the geometric organize of main points in the lip image defined by the change of time. As a result, we get a model that can identify words in a category and has a score of 73.1%.[1]

M.A. Anusuya et al. Presents a brief review of the acceptance of automated speech and discusses key topics and advances in 60 years of research, providing technical insights and understanding of the progress made in this field. This is communication speech. Developing a speech appreciation organization involves careful consideration of the following issues: the definition of speech categories, speech expressions, feature extraction techniques, speech organizers, archives, or presentation appraisals. . The purpose of this article is to summarize or evaluate some well-known technique used at different stages of the speech system, and to recognize the research topics or function that are at front of this exciting or difficult topic.[2]

Santosh K. Gaikwad is the most important tool in human communication. The connection between the machine connections is called the machine-man interface. Voice has the potential to become an important way of communicating with a computer. This article describes the key technical concepts in perspective, appreciates the key advances in speech recognition, and identifies the visual technology developed at each speech recognition process. This article helps in choosing the technology and the advantages and disadvantages. Comparative studies of different technologies have been done gradually. This article has finally decided on the unique direction of using marathi words to promote the technology of the human sacred system.[3]

Shanthi Therese, Chelpa Lingam, and others said that speech had become a major form of human interaction. The advent of digital technology has provided us with a wide range of digital processors with high-speed, high-speed, high-power capabilities, allowing researchers to convert audio signals into digital audio signals that can use in scientific research. Achieving higher grades, lower word error rates, and solving inequality issues are key concepts in designing an effective automatic speech recognition system. In the acceptance of speech, the exploitation of features requires special consideration, as the success of the acceptance depends largely on this aspect. In this post, we focus on the progress made so far in exploiting the speech recognition system and describing technical visions of automatic speech recognition system.[4]

Sanjib Das et al. Short audio interviews are the most important and effective tool for interpersonal communication. The connection between the machine connections is called the machine-man interface. Voice has the potential to become an important way of communicating with a computer. This article outlines the key points of view, appreciates key developments in speech recognition, and provides an overview of the technologies developed at each speech recognition stage. This article helps in choosing the technology and the advantages and disadvantages. Comparative studies of different technologies have been done gradually. This article is about making decisions about the unique leadership technology of promoting human beings through other languages and discusses the various technologies used in each step. Speaking, and trying to figure out a way to design an effective speech system. This article aims to summarize or evaluate the various speech structure and identify the most recent investigate topics and purpose in this interesting and challenging field. [5]

KNN CLASSIFICATION- The principle of K-NN operation can be clarify by following algorithm: In KNN algorithm, we select each support vector as the representation point and compare the distance between the samples. Test and vector are supporting each.[10]

•Step1: Select neighborhood number K
•Step2: Calculate the Euclidean distance of neighbor K
•Step3: Take your nearest K neighbor depending on the Euclidean reserve.

•Step4: Number of your neighbors k, count number of data points per class.
•Step5: Assign new data points to the largest number of neighbors in this category.

3. PROPOSED METHODOLOGY

In this process, a facial descriptor proposed uses a local binary pattern algorithm to extract feature information from an emotion-related facial appearance by using directional information and a ternary pattern to take fine edges in the face area while the face has a smooth area. By extracting this technique performs better than other methods. Then, while sampling the information related to expression at different scales, the grid used to construct the face descriptor is classified. Reducing the number of dimensions by extracting distinctive features is based on the idea of maximizing total data distribution and minimizing differences within the class. It can be seen that the characteristic values of the six categories are highly merged, which can lead to a higher misclassification rate. Note that the actual number of functions can exceed three, but the first three functions were selected for creation for the sake of intuition. Therefore, this work uses powerful features.

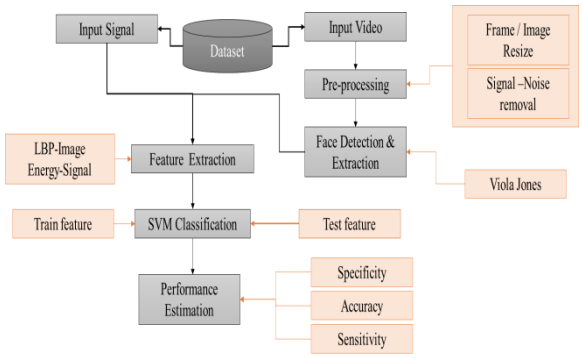


Fig.1 Proposed Flow diagram

In speech recognition, feature-by-feature features' major goal is to calculate a reduced chain of quality vectors to provide a compacted illustration of a given input indicator. Extraction of functions is regularly achieved in three phases. The first phase is described speech investigation or sound front end. It performs spectral time examination of signal or produce original functions that describe the power spectrum envelope for short speech distance. The second phase assembles the extensive function vector collected of static or dynamic functions. to finish, last phase (not always present) converts these extended function vectors into more compacted or robust vectors and then provides them to recognizer. Multi-mode systems use multiple biometric recognition systems at the same time. In this face detection process, Viola-Jones perceives the facial area and extracts the detected area. In the function extraction procedure, we can implement LBP for outline removal in images as well as MFCC used for extraction in speech signals. The Multi SVM classifier and KNN we implemented used to identify individuals and then display the results.

Face Detection using Viola-Jones Algorithm- The Viola-Jones algorithm is an extensively used instrument for article recognition. The major function of this algorithm is slow education speed but fast discovery speed. The algorithm uses Have basic function filters, so multiplication not use. By first produce, an integrated image, the competence of Viola-Jones algorithm can appreciably improve.

$$II(Y,X) = \sum_{P=0}^{Y} \sum Y(P,Q)$$

The integrated image makes it possible to calculate the integral of the H by adding only four information. For instance, picture essential of the ABCD region (Figure 1) is calculated as II

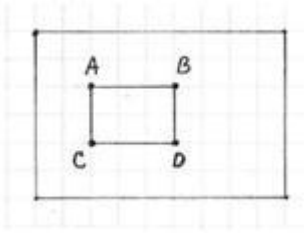
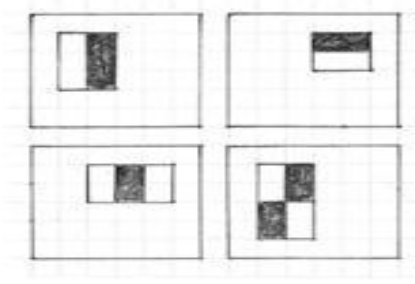


Fig.2 Image area integration using integral image

(yA, xA) -II (yB, xB) -II (yC, xC) + II (yD, xD).

Registration is done in the search window. Choose the minimum or maximum window size, or then select the amount of flying bread for each size. Then move the detection window to the image as shown below:

- Set window size to a minimum and adjust the step size depending on the size.
- For selected window size, window slides straight or horizontally in the same steps. At each step, an N-type verification filter use. If the censorship gives a right answer, there is a form in the present widow.
- If window size is maximum size, stop the process. or else amplify the window's size and move the part according to next selected size, or then proceed to step 2.
- The filter (from filter N) contains various categories of filters. Each category looks at the rectangular window of recognition window or conclude if it is a human shape. If so, the following classifier use. If all the racists give good answers, the censors provide the right solutions and accept the face. Otherwise, run the next filter on the smoke N filter.
- Each category contains haar miners (weak classification). Each Haar function is a combination of two-dimensional integrals in a rectangular area. The value may require a value of ± 1 Figure 2 shows an example of the Haar function associated with a closed window. The weight of the gray area is good, and the weight of the white area is bad. Hair expansion is a great way to improve your lifestyle, so have fun or treat yourself.



.

Fig.3 Example rectangle features shown qualified to the enclosing discovery window

Binary image: This is the characteristic of the simplest image with 2 gray values (0 and 1 or black and white). Each pixel is replaced by a tiny bit. These types of imagery are useful for computer vision that does not just need image information or

plans. You can create an image from a grayscale image that uses 0 to indicate pixels whose grayscale values are below the threshold and 1 to indicate other pixels. Still, this manufacturing method is useless because most of the information is lost or the image result is small.

Grayscale images: These pictures contain information about the brightness. The number of bits used to signify each pixel depends on the number of clarity levels available. The standard image is 8 frames per second. Pixels, which have 256 grayscale (Ng) or high values, range from 0 to 255.

Color image: Normally image is symbolized by the RGB model (red, green, blue), and each pixel has 24 bits. In numerous function, brightness information and color information are coupled and represented. The two pieces of information are estranged by transferring the RGB information to a numerical function.

Preprocessing: Noise diminution is a procedure of eliminating noise from a signal. All audio diplomacy, include digital or digital recordings, have a function that makes them vulnerable. Noise can be a sudden noise or a white noise with no direct or indirect noise introduced by a device mechanism or processing algorithm. In electronic recording devices, the main form of noise is the noise, which is caused by random electrons. With the force of high heat, random electrons will deviate from the direction indicated. These moving electrons will affect the signal strength of the signal, leading to detectable noise.

Face recognition: Face recognition is a computer technology that can be used in a variety of applications that recognize face in digital images. Facial recognition also refers to the person's psychological process seeing and caring for the face in the visual field. Paul Viola and Michael Jones introduced the Viola-Jones object detection system in 2001, which was the first object detection system to supply real-time article detection rates. Although it can be trained to identify dissimilar class of objects, it is due to form appreciation.

Feature extraction: pattern acceptance is a division of machine learning that focuses on pattern recognition and data inequality. In some cases it is almost considered a duplicate similar to machine learning. The tri-local state (LTP) is an extension of the local self-government (LBP). Unlike LBP, it does not set pixel entrance to 0 and 1, it uses a constant gate to set pixel level to 3 values. Using k as the threshold size, c as rate of center pixel, or use of adjacent pixel p as threshold.

MULTI SVM -In machine learning, a vector support machine (SVM, also known as a vector support network) is a standardized learning model with related learning algorithms that can classify and analyze data. To Giving a series of education, each training instance is marked for one or both of two categories. The SVM training algorithm will build a model and provide new examples for one or another category to become -P binary linear classifier (although this method has problems like Platt extraction to use SVM in the field of probabilistic classification). The SVM models represent situations such as points in space and show them the distance between each category as much as possible. Then put the new examples in the same room and guess that they fall into the class based on which side of hole they fell.

In addition to the theater random classifications, SVMs can also use the so-called core technology to successfully perform seamless classifications and seamlessly integrate their access to high-end model locations. It requires a learning approach that attempts to classify nature into groups and then maps new data maps for those groups formed. The clustering algorithm for vector support machines is called clustering vector support, which is often used in industrial applications. If no data is validated, or only a few data are marked as preprocessing the classification cards, they will be used in industrial applications.

In accepting the speech, the system of classification of controlled patterns gave an example. In other words, each input mode has an additional class mark. The order arranger can also be trained without supervision. For example, in a technique called vector qualitative, some reactions to input data are grouped by searching for a solid group of data. The customized cluster center table is called a codebook, and the new vector can be identified by locating the cluster center closer to the new vector. For voting, see Figure 4a. The case of the f1 child vowels is often and F2 is shown. Representative vowels are replaced by bot (/ a /) and boot (/ u /). Note that they are in a good group.

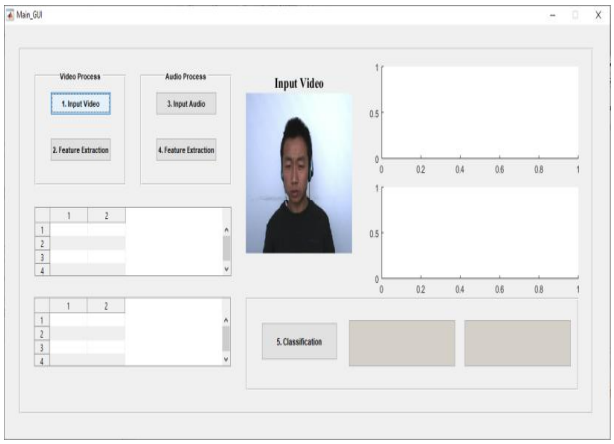


Fig.4 Input video dataset

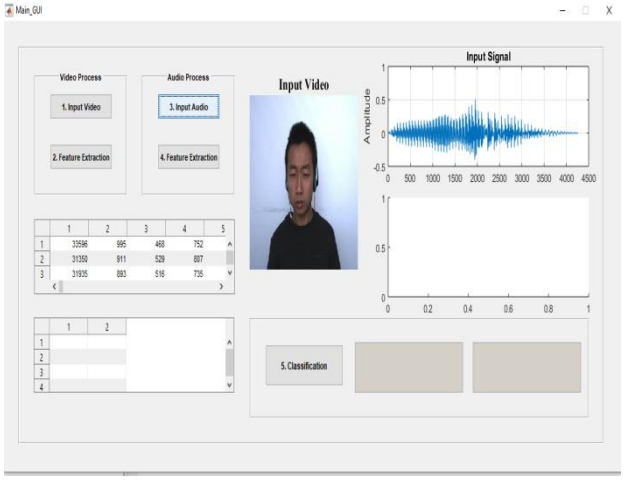


Fig.5 input audio dataset

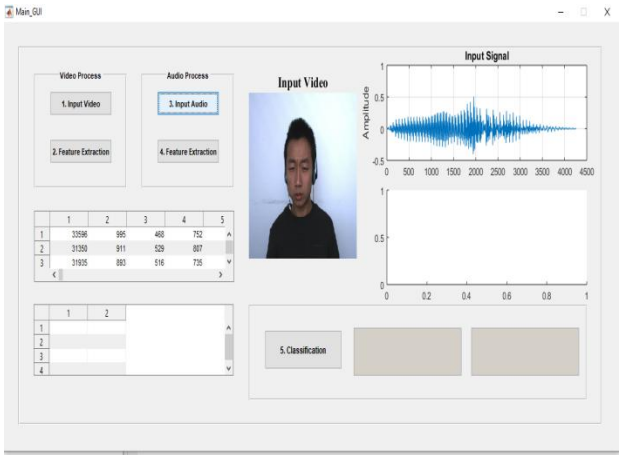


Fig.5 audio feature extraction

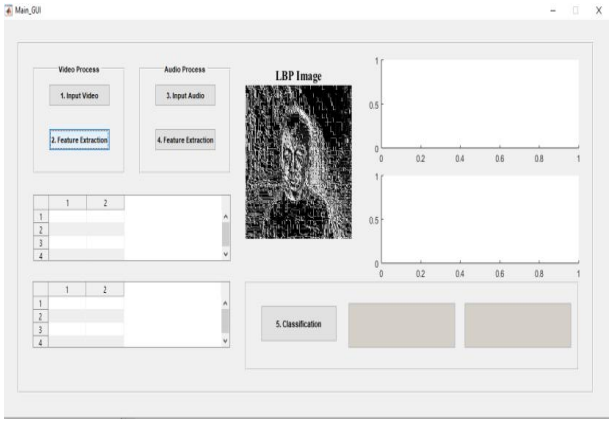


Fig.6 LBP feature extraction

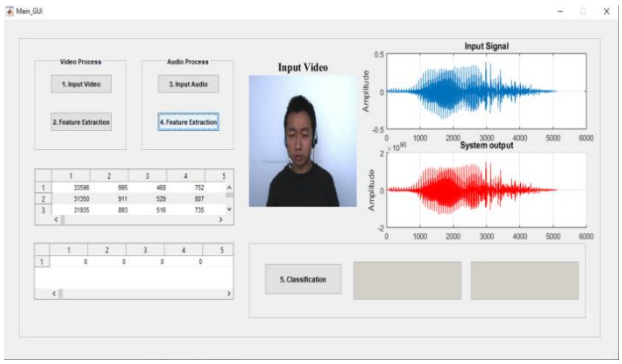


Fig.7 Audio feature extraction

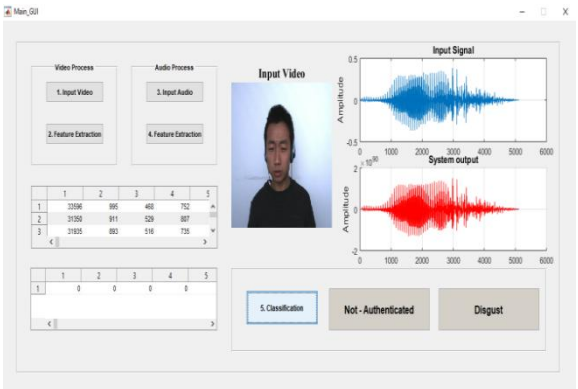


Fig.8 Person identification

Local Binary Pattern Functions (LBPs) work well in a diversity of applications, including texture cataloging or segmentation, image recovery, or exterior inspection. The task is to find the position and size of all substances belonging to a given class in the image. Examples include the upper body. The face detection algorithm focuses on perceive frontal faces. This is comparable to image detection, where imagery of people is matched little by little. The image matches image accumulates in the database. As shown in fig Predictably, the selected function will contain applicable information from input data so that this simplified representation can be use instead of the complete initial data to perform the required tasks The local binary mode (LBP) function works very well in different submission counting texture organization or segmentation, image retrieval or exterior examination shown Multi-SVM performs the mapping from input space to function space to support non-linear classification problems. Kern tricks can help achieve this by tolerating the lack of accurate representations of mapping skin texture that can cause the curse of dimensionality.

PERFORMANCE-MEASURE-Process performance measure against presentation indicators such as accuracy, sensitivity or specificity.

Terms associated to presentation indicators:

- TP-True (correct identification)
- TN true negative (reject correctly)
- FP false positive (identification error)
- UN false negative (wrong rejection)

Accuracy: Accuracy in a categorization difficulty is the number of correct predictions from the model in dissimilar calculations.

Accuracy = (TP+TN) / (TN+TP+FN+FP)

Sensitivity: The aptitude of test to suitably identify people with the disease (true positive rate). compute authentic positive rate correctly identified.

Sensitivity = TP / (TP + FN)

Specificity: ability of the test to correctly identify people without the disease (true negative speed). Measure proportion of original negatives properly identified.

Specificity = TN / (TN + FP)

F-measure: F-measure (F1-score or F-score) calculates test exactness or is defined as the weighted harmonic mean of exactitude or recall.

f_measure = 2 * ((precision * recall rate) / (precision + recall rate));

The data in Table 1 show Comparison work with previous work

Table I. Comparison of the Results Obtained With Existing Solutions

	Classification	Accuracy	Specificity	Sensitivity
Proposed Work	Multi -SVM	99.68	99.93	99.63
	KNN	99.36	99.90	99.60
Previous Work	Neural Network Model	73.1	-	-

4. CONCLUSION

Biometric recognition is related to human recognition through physiological individuality, fingerprints, iris, voice, face, etc. The biometric system can use for classification or verification of people. This paper detects the human face and voice

signals. Therefore, the LBP function descriptor extracts the exact functions in pixels in the image. MFCC technology is used for speech recognition. We use Multi-SVM and KNN classifier for classification. Therefore, compared to other classifiers, the MUTI SVM and KNN classifier require only a few positive examples. Therefore, the performance of the classifier also assesses. The selected functions are introduced into the organization to achieve best results. Since the procedure is iterative, results attain filtered through iteration. Therefore, future enhancements will make using PSO (Particle Swarm Optimization) knowledge. This technique will achieve a better degree of recognition. The result is better in terms of recognition speed, correctness or best numeral of functions produce.

5. REFERENCES

[1] V. Zatonskikh, Georgii I. Borzunov, Konstantin Kogos Development of Elements of Two-Level Biometric Protection Based on Face and Speech Recognition in the Video Stream Efim Department of Cryptology and Cybersecurity National Research Nuclear University MEPhI (Moscow Engineering Physics Institute) Moscow,

[2] M.A.Anusuya and S.K.Katti ,Department of Computer Science and Engineering,Sri Jaya chamarajendra College of Engineering, Mysore, India, (IJCSIS) International Journal of Computer Science and Information Security,2009.

[3] Santosh K.Gaikwad, Dr.Babasaheb Ambedkar Marathwada, Bharti W.Gawali, 2011, A Review on Speech Recognition Technique.pp1561-1569

[4] Shanthi Therese ,Chelpa Lingam, International Journal of Scientific Engineering and Technology, June 2013.,Review of Feature Extraction Techniques in Automatic Speech Recognition.

[5] Speech Recognition Technique: A Review Sanjib Das Department of Computer Science, Sukanta Mahavidyalaya, (University of North Bengal), India, International Journal of Engineering Research and Applications (IJERA) MayJun 2012.

[6] Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank SeideMichael L. Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, Yifan Gong, and Alex Acero Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA 2009

[7] Nidhi Desai1, Prof.Kinnal Dhameliya2, Prof.Vijayendra Desai3, International Journal of Emerging Technology and Advanced Engineering, December 2013, Feature Extraction and Classification Techniques for Speech Recognition: A Review.

[8] Li Deng and John C. Platt, Microsoft Research, One Microsoft Way, Redmond, WA, USA, November 2010, Ensemble Deep Learning for Speech Recognition.

[9] Samy Bengio and Georg Heigold, Google Inc, Mountain View, CA, USA, feb. 2007, Word Embeddings for Speech Recognition. Rubi, International Journal of Computer Science and Mobile Computing, Vol.4 Issue.5, May- 2015, pg. 1017-1024 © 2015, IJCSMC All Rights Reserved 1024

[10] Chalapathy Neti, Member, IEEE, Guillaume Gravier,, Ashutosh Garg, Audio-Visual Speech Gerasimos Potamianos, Member, IEEE, Student Member, IEEE, and Andrew W. Senior, Member, IEEE 2006, Recent Advances in the Automatic Recognition.

[11] Dandan Mo, December 4, 2012, A survey on deep learning: one small step toward AI. 11. Aalto University publication series, Foundations and Advances in Deep Learning, Kyunghyun Cho, 2014.

[12] Abboud, A. J., Sellahewa, H. and Jassim, S. A. "Quality approach for adaptive face recognition", in Proc. Mobile Multimedia/Image Processing Security, and Applications, SPIE Vol. 7351, 73510 N, 2009.

[13] Aloysius G., "Efficient High Dimension Data Clustering using ConstraintPartitioning KMeans Algorithm," the International Arab Journal of Information Technology, Vol. 10, No. 5, pp. 467-476, 2013.

[14] Alsaade.F and Zahrani.M, "Enhancement of Multimodal Biometric Verification Using a Combination of Fusion Methods",5th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications March 22-26, 2009.

[15] Amoli.G, Thapliyal.N, Sethi.N: Iris Preprocessing. International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, No. 6, pp. 301-304, 2012.

[16] Ang.R. Safavi-Naini.R, McAven.L:. Cancelable Key-based Fingerprint Templates. In C. Boyd and J. Gonzalez Nieto (Eds.), Australasian Conference on Information Security and Privacy, pp. 242-252, 2005.