Improving User Identification Accuracy in Facial and Voice Based Mood Analytics using Fused Feature Extraction

Dolly Reney, Neeta Tripathi

Abstract—User identification involves a lot of complex procedures including image processing, voice processing, biometric data processing and other user specific parameters. This can be applied to various fields including but not limited to smartphone authentication, bank transactions, location based identity access and various others areas. In this work, we present a novel approach for uniquely identifying users based on their facial and voice data. Our approach uses an intelligent and adaptive combination of facial geometry and mel frequency analysis (via Mel Frequency Cepstral Co-efficient or MFCC) of user voice data, in order to generate a mood based personality profile which is almost unique for each user. Combination of these features is given to a machine learning based classifier, which has proven to produce more than 90% accuracy with a false positive rate of less than 7%. We have also compared the proposed approach with various other standard implementations and observed that our implementation produces better results than most of them under real time conditions.

Keywords: Identification, authentication, facial, geometry, MFCC, machine learning

I. INTRODUCTION

User identification has emerged as a trending area for authentication and uniquely representing the user's body features in digital form. The user identification process consists of the followingsteps,

- User specific data acquisitionphase
- Denoising andpre-processing
- Segmentation of various areas from the acquireddata
- Extraction and selection offeatures
- Fusion of selectedfeatures
- Classification andidentification

The data acquisition phase is of utmost importance. In this phase, the images, audio, video, finger prints and other user specific parameters are captured from the data capturing devices. These devices generate output data in their specific formats, the system designers need to convert this data as per their requirements but for analysis we need to combine these channels and produce data in single channel format. Accurate acquisition of data is a direct indication of increased accuracy of user identification, thus this step should be performed very carefully. In some applications like face detection, the user's face should be in frontal format, so that the accuracy of detection can beimproved. The acquired data are given to the pre-processing block, where the unwanted noise from the data is reduced. Furthermore, the denoised data can be given to an enhancement block, wherein any hidden features of the data can be enhanced, thus resulting in better segmentation. While pre-processing enhances the quality of the data, there is a need of extraction of intelligent information from this data, which is done by the segmentation block. This block extracts only the required regions from the data and removes all the unwanted regions. Methods for segmentation include thresholding, region growing, clustering, saliency maps, Viola Jones, etc. These methods are selected based on the application being developed for the user identification.

Segmentation results in the data with regions of interest, these regions vary w.r.t. color, shape, texture, audio, contours and other parameters. The segmented regions are given to a feature extraction block, this block defines the features of the overall data, by which the color, texture, shape, frequency components, contour information and other parameters of the data are described. Feature extraction methods include, mel frequency cepstral component extraction (MFCC), histogram evaluation, extended color maps, edge maps, texture maps and various others. These features are needed due to the fact that while classification, we cannot classify input data directly, because this data might vary in size, color, shape and texture, thus we need a constant sized feature vector, which can be used directly for the classification purpose.

Selection of features is generally an optional step in the process of user identification. In this process, the extracted features are checked for repetition, and any other ambiguities. These ambiguities and repetitions are removed from the feature vector, and an optimized feature vector is produced at the output.

The classification process involves comparison of the extracted features against a pre-defined training set. This comparison is usually done with the help of neural networks, support vector machines, nearest neighbour classifier and naive bayes classifier. Via this comparison, the data and it's regions are classified as per the application.

In our work, we are using MFCCs and facial geometry techniques for feature extraction, combined with standard deviation and variance for feature selection. The extracted feature maps are given to a machine learning classifier in order to identify the user uniquely. The next section describes some state of the art methods for user identification, followed by the methods used in this paper, and finally we conclude with some interesting observations from our results.

Revised Manuscript Received on October 23, 2019.

Dr. Dolly Reney, Electrical and Electronics Engg.Deptt, Oriental University, Indore,India.

Dr. Neeta Tripathi, tronics and Telecommunication Engg. Deptt.,shankaracharya Group of Instiute, Bhilai,India.

II. LITERATURE REVIEW

In this area we will depict the different methodologies that were utilized face to face recognizable proof utilizing biometric frameworks. A biometric framework is utilizes the particular physiological or conduct highlights controlled by the client for ID and these highlights are remarkable, widespread and persevering. These frameworks incorporate face recognition, unique mark innovation, iris recognition, hand geometry, keystroke, mark and speechrecognition.

a. FaceRecognition

Facial pictures are the basic biometric include utilized for individual recognizable proof. Face recognition is primarily performed by two methodologies, they are eigen face based recognition and 3D face recognition. The eigen face based recognition works by examining face pictures and registering eigen faces which are faces made out of eigenvectors. The correlation of Eigen faces is utilized to distinguish the nearness of a face and its character. The Eigen face system is direct, productive, and yields commonly great outcomes in controlled condition. There are likewise a few confinements of Eigen faces. There is restricted strength to changes in lighting, edge, and separation. 2D recognition frameworks don't catch the real size of the face, which is a crucial issue. These impediments influence the system's application with surveillance camera. 3D face recognition frameworks make 3D models of faces and think about the 3D faces for recognition. These frameworks are increasingly exact in light of the fact that they catch the real state of faces. The procurement of 3D information is one of the principle issues for 3D frameworks. Another face recognizable proof innovation, Facial thermo grams, utilizes infrared warmth sweeps to distinguish facial attributes. This non-noisy system is light-autonomous and not powerless against masks. Indeed, even plastic medical procedure, can't prevent the system. This strategy conveys upgraded precision, speed and dependability with insignificant capacity necessities. To keep a phony face or shape from faking out the framework, numerous frameworks require the individual to grin, flicker, or generally move in a way that is human before checking Unique markinnovation

A unique mark is the example of edges and depressions on the surface of a fingertip. The fingerprints are exceedingly steady and extraordinary. The uniqueness of unique mark is controlled by worldwide highlights like valleys and edges, and by nearby highlights like edge endings and edge bifurcations, which are called details. The ongoing investigations uncover that likelihood of two people, having a similar unique mark is short of what one of every a billion.

b. Iris recognition

Iris recognition frameworks make utilization of the uniqueness of the iris examples to distinguish an individual. This framework utilizes a superb camera to catch a high contrast, high-goals picture of the iris (the hued ring encompassing the student). Iris recognition comprises of five activities; they are picture securing, iris limitation or division, iris standardization and unwrapping, include encoding, and coordinating calculation. The standardized iris locale is unwrapped into a rectangular locale. The component encoding is utilized to remove the most separating highlight in the iris design with the goal that an examination between formats should be possible. At last a choice can be made in the coordinating advance.

c. Hand geometry

Hand geometry recognition frameworks utilize various estimations taken from the human hand, including its shape, size of palm, and lengths and widths of the fingers. The procedure is exceptionally basic, moderately simple to utilize, and cheap. Hand geometry based ID comprises of following advances, picture catching and pre-handling, estimations and highlight determination lastly order and confirmation.

d. Keystroke

It is evaluated that every individual kinds on a console unmistakably. Keystroke elements is a conduct biometric; for a few people, we can watch substantial varieties in average composing designs. Further, the keystrokes recognition should be possible inconspicuously when the individual is entering in data. Social attributes estimated by keystroke recognition include: the aggregate composing speed, the time that slips by between sequential keystrokes, the time that each key is held down, the recurrence with which different keys, for example, the number cushion or capacity keys, are utilized and the arrangement used to type a capital letter

e. Mark

Mark recognition depends in transit an individual signs his or her name. Marks are a social biometric that change over some stretch of time and are affected by physical and passionate states of the people. Proficient falsifiers might almost certainly repeat marks that trick the framework. Biometric marks are utilized in keeping money and back industry so as to confine copy signature fakes. Dynamic mark check innovation is utilized, where the parson reach delicate gadgets like PDA or tablet PC. This innovation is additionally introduced in cell phones to avoid illicit access, despite the fact that the gadget is lost or stolen.

f. Speech recognition

Speech Recognition (is otherwise called Automatic Speech Recognition (ASR) or PC speech recognition) is the way toward changing over a speech flag to an arrangement of words, utilizing a PC program. Speech recognition innovation was progressively utilized inside phone networks to mechanize just as to upgrade the administrator administrations. For the most part there are three ways to deal with biometric speech recognition; they are Acoustic Phonetic Approach, Pattern Recognition Approach and Artificial Intelligence Approach. Acoustic Phonetic Approach depended on discovering speech sounds and giving fitting names to these sounds. All through the previous couple of decades there have been many face identification procedures proposed and executed. In high-

dimensional information, PCA strategy is intended to display straight variety. Its will likely locate a lot of commonly symmetrical premise works that catch the bearings of most extreme difference in the information and for which the coefficients are pairwise decorrelated [3]. Eigenfaces was presented early [4] on as amazing utilization of principal components analysis (PCA) to take care of issues in face recognition and discovery. PCA is an unsupervised strategy, so the technique does not depend on class data. In our execution of eigenfaces, we utilize the nearest neighbor (NN) way to deal with arrange our test vectors utilizing the Euclidean distance[2].One augmentation of PCA is that of applying PCA to tensors or multilinear exhibits which results in a technique known as multilinear principal components analysis (MPCA) [5]. Fisherfaces is the immediate utilization of (Fisher) straight discriminant analysis (LDA) to face recognition [6]. ICA is a speculation of PCA in that it attempts to distinguish high-arrange factual connections between pixels to shape a superior arrangement of premise vectors. In [8], where the pixels are treated as arbitrary factors and the face pictures as results. To demonstrate our method for perceiving faces is imitated fairly by utilizing neural network. This is practiced with the point of creating recognition frameworks that joins man-made consciousness for thinking of a framework that is keen. The utilization of neural networks for face recognition has been appeared by [9] and [10]. In [11], we can see the proposal of a semi-administered learning strategy that utilizes bolster vector machines for face recognition. There have been numerous endeavors in which notwithstanding the regular systems neural networks were actualized.

A basic content ward speaker distinguishing proof methodology, consolidating spectrograms and Discrete Cosine Transform (DCT) was introduced by Kekre et al [11] dependent on DCT use to find similitudes between free example spectrograms. A direct based element extraction calculation where highlight depends on late time-recurrence change and modules to reproduce cochlea flag preparing was exhibited by Li et al [12]. An exploration Agenda - Short Utterance Speaker Recognition - was a vital speaker recognition are when just restricted speech information was accessible for testing/preparing was proposed by Fatima and Zheng [13]. This recorded normally utilized best in class speaker recognition strategies and noteworthiness of prosodic speaker recognition. Speaker recognition utilizing dynamic neurotransmitter based neural networks.[NN] with wavelet pre-handling was proposed by George et al [14] which included a framework equipped for speaker check in a shut speakers set utilizing a wavelet preparing process allowing speaker subordinate list of capabilities extraction. A speaker distinguishing proof framework utilizing Wavelet Transform and NN was proposed by Daqrouq [15] .

III. PROPOSED USER IDENTIFICATION ALGORITHM

The proposed system can be represented using the following block diagram,

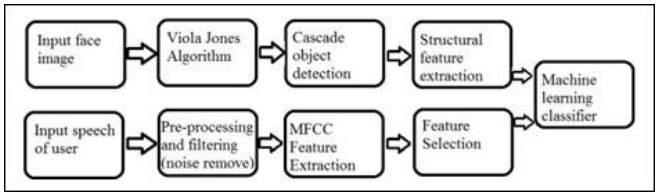

Figure 1. Proposed system

In the proposed system, the user's speech and face image are taken at the input. The two input signals are processed very differently, this processing on the individual signals results into a combined feature vector, which is given to a machine learning classifier. The details of each block are explained as follows,

a. Input face image

This block is responsible for capturing the images, and storing them for further processing. We used a standard camera for this purpose, so that the system is inexpensive to re-create for any researcher.

b. Viola Jones Algorithm

This algorithm is based on rescaling the detector instead of input image and run the detector many times through the image every time with different size and this is covered in following steps

1. Integral images calculation
2. Adaboost algorithm
3. Cascade object detection
4. Structural feature extraction.

Finally speech processing is done as given below

c. Input speech from user & pre-processing

The input speech was taken from pre-recorded samples, sampled at 8kHz, and were pre- processed by a median and gaussian filter inorder to remove any noise present in the sound. This denoised sound is then given to a silent regions removal unit, which removes all the silent regions from the sound and produces only speech at theoutput.

d. Mel Frequency Cepstral Component extraction

The voice signal is given to the MFCC extraction block,this block extracts the cepstral component from the voice samples and describes the voice in terms of features of fixed size length, This is needed as the voice of the user can be of varying length,andthgus might not be useful while comnparsion by machine learning classifier,as the classifier needs fixed length inputs for processing,thus MFCC is used.

e. Feature selection

MFCC gives a lot of features after processing. These features are generally repetitive in nature a and cause accuracy reduction and an increase in classification delay if taken directly. Thus, we need to reduce the features and select only those features which are important for the system. In order to do this, we apply a variance calculationformula,

$$var(x) = \frac{\sum \sqrt{x_i^2 - mean(x)^2}}{N}$$

where, x is the feature vector, and N is the number of samples in the feature vector. If the sample value is more than the variance, then it is accepted and stored for classification, else it is discarded. Using this step, reduces the feature size drastically and helps to optimize the accuracy of the overall system.

f. Machine learning classifier

This is the most crucial step for the system. In this step, the input dataset is first divided into training and testing sets, and the classifier is trained with the training set. The training set is selected to contain all possible combinations of speech and faces for the user, so that the training of the algorithm is done optimally, and the results are generated with utmost accuracy. The machine learning classifier is based on a modified version of the k-nearest neighbour classifier and works as follows,

- For each feature set, evaluate the number of features Nf, Nv. Where Nf are the number of face features, and Nv are the number of voicefeatures
- Go from k = 1 to Nf, and 1 to Nv, & follow the givensteps,
  - Use euclidean distance and classify the input sample with knn classifier with the given value ofk
  - Use city-block distance and classify the input sample with knn classifier with the given value ofk
  - Use cosine distance and classify the input sample with knn classifier with the given value ofk
  - Use correlation and classify the input sample with knn classifier with the given value ofk
  - Store the classes given by each of the steps into an arrayP
  - Repeat the steps for each value ofk
- Find the most frequently occurring entry in the array P, and use that as the class for the given featureset

If the classes for both face and voice matches, then identify the user, else mark the input samples as un-identified. Repeat this process for all images and voice samples at the input and evaluate the accuracy. The next section shows the accuracy and delay comparison of our algorithm with standardtechniques.

IV. RESULTS AND ANALYSIS

We compared our results with existing standard techniques like neural networks and support vector machines, and observed the following results on our customized dataset of real time users with their actual recorded voice samples, this paid dataset is available on request,

Number of trained samples	Number of tested samples	Accuracy (%) NN+LDA+MFCC	Accuracy (%) SVM+LDA+MFCC	Accuracy (%) NN+PCA+MFCC	Accuracy (%) SVM+PCA+MFCC	Accuracy (%) Proposed
10	10	70	70	80	70	90
20	20	75	70	80	75	90
30	30	75	75	82	80	88
50	75	77	75	84	82	89
75	100	78	76	84	83	91
100	120	79	76	85	84	90
150	175	79	77	86	86	93
200	250	81	78	86	87	93
300	375	81	79	87	88	93
400	500	82	80	87	89	94
500	600	82	81	88	90	94

Table 1. Accuracy comparison

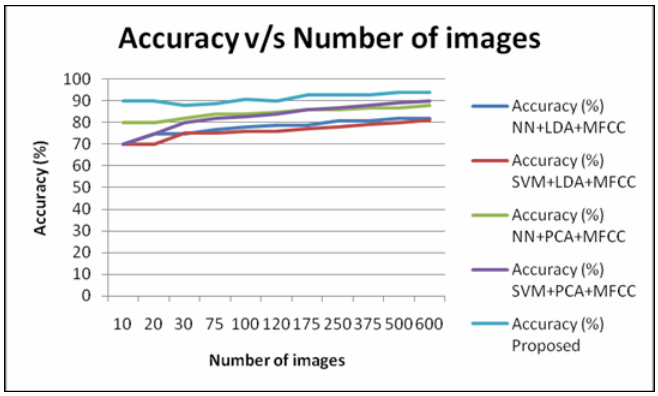

Figure 3. Accuracy v/s Number of images

From the graph, we can observe that the proposed system outperforms the standard systems in terms of accuracy by more than 20% in most cases. Thus the system can be used for accurate classification of users from both voice and image data. The next comparison was made for delay analysis, which can be shown as follows,

Number of trained samples	Number of tested samples	Delay (ms) NN+LDA+MFCC	Delay (ms) SVM+LDA+MFCC	Delay (ms) NN+PCA+MFCC
10	10	0.60	0.70	0.65
20	20	0.75	0.82	0.79
30	30	0.82	0.96	0.89
50	75	0.91	1.23	1.07
75	100	1.26	2.64	1.95
100	120	1.58	3.88	2.73
150	175	1.92	4.57	3.25
200	250	2.88	5.99	4.44
300	375	3.96	6.72	5.34
400	500	5.60	7.94	6.77
500	600	6.80	8.10	7.45

Table 2. Delay comparison

The delay is reduced by nearly 25% due to the fact that the feature selection unit reduces the number of features needed for comparison, and thus it improves the speed of classification, thereby requiring less delay.

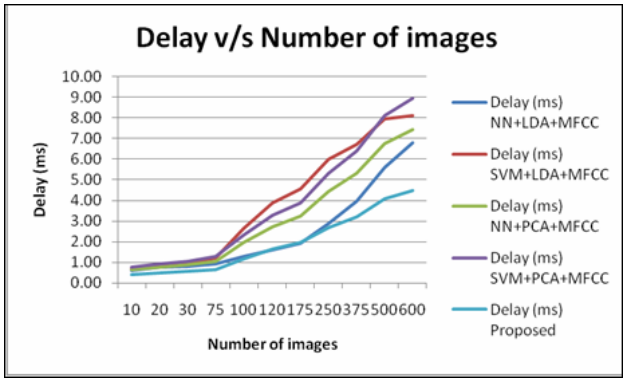

Figure 4. Delay v/s Number of images

Other methods use the feature set directly, thereby requiring more delay as compared to the proposed method. Thus, the proposed system can be used in real time scenarios where the rate of change of input images is fast and processing has to be done in real time, like for smartphone authentication.

V. CONCLUSION

From the observed results we can conclude that the system outperforms the existing systems in terms of accuracy and delay, and thus it can be used for real time applications in smartphones and other areas. The system improves the accuracy and delay mainly due to the feature selection unit and the machine learning unit, which adaptively select features for classification, and results in optimum system performance. The system can further be improved with the help of artificial intelligence algorithms like deep nets, so that the accuracy can be further improved.

REFERENCES

1. D. Pullella, R. Togneri, "Speaker identification using higher order spectra," Dissertation of Bachelor of Electrical and Electronic Engineering, 2006.
2. R. Jaksa, M. Katrák, "Neural network model of the backpropagation algorithm," I-st. Slovak-Japanese seminar on intelligent systems, Herl'any. 2005.
3. U. Shrawankar, V. M. Thakare, "Techniques for feature extraction in speech recognition system: A comparative study", arXiv preprint arXiv:1305.1145, May 2013.
4. Li Q, Huang Y., "Robust speaker identification using an auditory-based feature", In IEEE In. Conf. on Acoustics, Speech and Signal Processing, March 2010,pp. 4514-4517.
5. D. Reney, N. Tripathi, "Human face and expression recognition with kernel fisher analysis", IJEAST, Vol. 1, Issue 6, pp. 152-156, 2016.
6. P. Viola, M. J. Jones, "Robust real-time face detection", International journal of computer vision, vol. 57, pp. 137-154, May 2004.
7. S. A. Khayam, "The discrete cosine transform (DCT): theory and Coding, Michigan State University, pp. 602-802, March 2003. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.184.6102&rep=rep1&type=pdf
8. R. D. Peacocke, D. H. Graf, "An introduction to speech and speaker recognition", In Readings in Human–Computer Interaction, pp. 546-553, Jan 1995.
9. K. Gibert, I. Pinyol, L. Oliva, et al, "Pseudobagging: Improving class discovery by adapting bagging techniques to clustering algorithms", Proceedings 'V Taller Nacional de Minería de Datos y Aprendizaje'in: Ferrer-Troyano, FJ, pp.157-166.
10. S. Gaikwad, B. Gawali, P. Yannawar, S. Mehrotra, "Feature extraction using fusion MFCC for continuous marathi speech recognition", India Conference (INDICON), IEEE, pp. 1-5, Dec 16.
11. K. Daqrouq, T. A. Hilal, M. Sherif, S. El-Hajjar, A., Al-Qawasmi, "Speaker identification system using wavelet transform and neural network", International Conference on Advances in Computational Tools for Engineering Applications, ACTEA, pp. 559-564, Jul 15.
12. S. Malik, F.A. Afsar, "Wavelet transform based automatic speaker recognition", Multi Topic, IEEE International Conference, pp. 1-4, Dec 2009.
13. P. Motlıcek, "Feature extraction in speech coding and recognition", Technical Report of PhD research internship in ASP Group, OGI-OHSU, Sep. 2002.
14. D. Reney, N.Tripathi, "An Efficient Method to Face and Emotion Detection", Fifth International Conference on Communication Systems and Network Technologies, pp. 493-497, Apr 2015.
15. D. Avci, "An expert system for speaker identification using adaptive wavelet sure entropy", Expert Systems with Applications, vol. 36, pp. 6295-6300, April 2009.