ISSN:0975-9646
Ishwar S. Jadhav et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (3) , 2011, 1248-1252

Human Identification using Face and Voice Recognition

[1]Ishwar S. Jadhav, [2]V. T. Gaikwad, [3]Gajanan U. Patil

[1, 2] Department of Electronics & Telecommunication Engineering, Sant Gadge Baba Amravati University, Amravati, Maharashtra, India

[3] Department of Electronics & Communication Engineering, North Maharashtra University, Jalgaon, Maharashtra, India

ABSTRACT ***In this paper we proposed new technique for human identification using fusion of both face and speech which can substantially improve the rate of recognition as compared to the single biometric identification for security system development. The proposed system uses principal component analysis (PCA) as feature extraction techniques which calculate the Eigen vectors and Eigen values. These feature vectors are compared using the similarity measure algorithm like Mahalanobis Distances for the decision making. The Mel-Frequency cestrum coefficients (MFCC) feature extraction techniques are used for speech recognition in our project. Cross correlation coefficients are considered as primary features. The Hidden Markov Model (HMM) is used to calculate the like hoods in the MFCC extracted features to make the decision about the spoken wards.***

KEY WORDS: PCA- Principal Component Analysis, Mahalanobis, MFCC- Mel-Frequency cestrum coefficients, HMM- Hidden Markov Model, ASR-Automatic Speaker Recognition system.

1. INTRODUCTION

Biometric identity authentication systems are based on the biological characteristics of a person, such as face, voice, fingerprint, iris, gait, hand geometry or signature. Identity authentication using the face or the voice information is a challenging research area that is currently very active, mainly because of the natural and non-intrusive interaction with the authentication system. An identity authentication system has to deal with two kinds of events: either the person claiming a given identity is the one who he claims to be called client or if it is not then it is an impostor. Moreover, the system may generally take one decision: either accept the client or reject him and decide he is an impostor.

Among the behavioral biometrics, speech is the most convenient parameter. Automatic speaker recognition (ASR) systems identify people utilizing the utterances. Depending upon the nature of the application, speaker identification or speaker verification systems could be modeled to operate either in text dependent or text-independent modes. For text-dependent ASR, the user is required to utter a specific password, while for text-independent ASR, there is no need for such a constraint. Success in both cases depends on the modeling of speech characteristics which distinguish one user from the other. Text-dependent ASR is used for applications where the user is willing to cooperate by memorizing the phrase or password to be spoken which could be inconvenient to some users. Therefore, in this project, our focus is on text-independent ASR which is considered to be a more challenging problem.

Research in the field of speaker recognition traces back to the early 1960s when Lawrence Kersta at Bell Labs made the first major step in speaker verification by computers. He proposed the term voiceprint for a spectrogram, which was generated by a complicated electro-mechanical device. Since then, there has been a tremendous amount of research in the area. Starting from spectrogram comparisons, passing through simple template matching, dynamic-time wrapping, to more sophisticated statistical approaches like Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) and neural networks. Mel Frequency Cepstral Coefficients or MFCC features have also been used for text-independent speaker recognition.

Face recognition research has also been about for over three decades with comprehensive surveys like Zhao et al. Classical face recognition research was based on matching single pair's holistic facial images. Later, multiple independent images per individual were used to train a Linear Discriminate Analysis (LDA) classifier and recognition was performed on a single test image. These techniques did not cope well with changes in illumination, pose and facial expressions. Face recognition has become popular over the past few years because of the common availability of cameras and their ability to capture more information. Moreover, motion helps in the recognition of faces.

Multi-modal biometric research has recently gained popularity. Biometrics from independent methodology complements each other and increase the accuracy and robustness of the system. Speech and face are natural choices for multimodal biometric applications because they can be simultaneously acquired with camera and microphone. A review of audio-visual person. Identification and verification is given by Sanderson and Paliwal. A detailed book on the subject including fusion techniques is also available.

The speaker identification module gets the input from the microphone the preprocessing like voice activity detection is used to detect the start and stop of the voice sample. The MFCC is calculated as the extracted feature and the decision is made using the Hidden Markov Model to calculate the likelihoods.

Low resolution camera is used to capture image for face recognition module, the preprocessing algorithm are employed like filtering to remove high frequency noise. The geometric normalization is used to remove the variation between size, orientation and location of the face in the image. The feature extraction module uses principal component analysis (PCA) decomposition on the training set, which produces the Eigen vector and Eigen values. The classification module identifies the face in a face space. The critical parameter in this classification step is the subset of eigenvector used to represent the face. The nearest neighbor classifier is used as a main classifier which ranks the gallery image by similarity measure. For similarity measure the angle between feature vector and Mahalanobis Distances is used to provide the decision.

2. METHODOLOGY

Speaker identification module

In this paper, we use Cross correlation coefficients as primary features then the MFCC features are used for text-independent speaker recognition. MFCC is a popular feature extraction technique for speech signals.

The main idea behind MFCC features is to imitate the behavior of a human ear. Psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale.

Thus for each tone with an actual frequency *f*, measured in Hertz, a subjective pitch is measured on a scale called the *Mel Scale*. The Mel-frequency scale has linear frequency spacing below 1000 Hz and a logarithmic spacing above 1 KHz as shown in Figure 2.

As a reference point, the pitch of a 1 KHz tone, 40dB above the perceptual hearing threshold, is defined as 1000 Mels. Therefore, we can use the following formula to approximate the Mels for a given frequency *f in* Hertz.

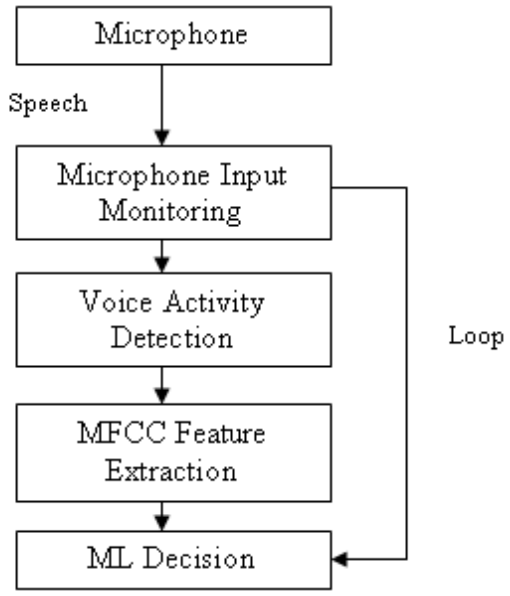

Figure 1 Process Flow of Speaker Identification Module

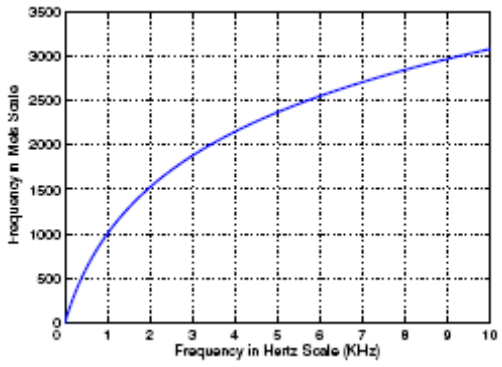

Figure 2 Relationship between Hertz and Mels scales

$$MEL(f) = 2595\ log_{10}(1 + \frac{f}{700})$$

Eq. (1)

One approach to simulate the subjective spectrum is to use a filter bank, spaced uniformly on the Mel scale as shown in Figure 3.

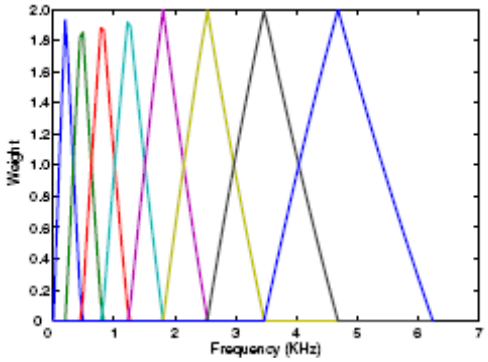

Figure 3 Mels frequency filter bank The vertical axis represents the weights of filter coefficients

This filter bank is applied to the spectrum of the speech signal to get a Mel-spectrum. The Mel-spectrum when transformed back to time domain using the Discrete Cosine Transform (DCT) gives us the MFCC coefficients. Therefore, if we denote the Mel power spectrum coefficients by S_k (where *k* is the index of the Mel-spaced filters and $k = 1, 2, \ldots, K$) then the MFCC coefficients (C_n) are calculated as

$$c_n = \sum_{K=1}^{K} (logS_k) cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right]; n = 1,2, \ldots L$$

Eq. (2)

We represent MFCC coefficients calculated in Eq. 2 as an MFCC feature vector,

$$c = [c_1 c_2 c_3 \ldots c_L]^T$$

Eq. (3)

In this paper, we consider second order statistical modeling of the speech, assuming a wide sense stationary process (WSS). Let *J* be the index of a class and let there be a total of *N* classes so that $j = 1, 2. \ldots N$. Suppose there are *b* training utterances available per class. Each utterance from class *j* is framed using a 20ms window so that the assumption of stationary is true for each frame. Suppose after the framing of all training utterances from class *j* we end up with *hj* number of frames. We now calculate 30 MFCC features for each of the *hj* frames using 30 filter banks (i.e. $L = K = 30$ in Eq. 3) and place them in the columns of a matrix C*j*, so that the matrix is of order $30 \times hj$

$$\mathbf{C_j} = [\mathbf{c_1}|\mathbf{c_2}|\ldots\mathbf{c}_{h_j}]$$

Eq. (4)

Now we develop second order statistical model **U***j* for class *j* as,

$$\hat{\mathbf{c}} = \frac{1}{h_j}\sum_{i=1}^{h_j}\mathbf{c}_i\ ,$$

Eq. (5)

$$\mathbf{U_j} = \frac{1}{h_j}\sum_{i=1}^{h_j}(\mathbf{c}_i - \hat{\mathbf{c}})(\mathbf{c}_i - \hat{\mathbf{c}})^T\ .$$

Eq. (6)

We use the same procedure to develop the above covariance model **U**j for all classes i.e.(j = 1, 2, . . .,N). Similarly, for a test utterance z, we derive a covariance matrix **Z**. Once we have developed the second-order statistical model, we apply

an arithmetic-harmonic sphericity measure as the distance metric between all class models **Uj** (j = 1, 2, . . .,N) and **Z**, thus

$$D_s(\mathrm{U_j}, \mathrm{Z}) = \log\left[\frac{\mathrm{tr}\left(\mathrm{U_j Z^{-1}}\right)\mathrm{tr}\left(\mathrm{Z U_j^{-1}}\right)}{L^2}\right];\ j = 1, 2, \ldots, N$$

Eq. (7)

Where *L* is the dimension of the feature vector and tr (A) is the trace of matrix A. The test pattern *z* is assigned to the class having the minimum distance measure *Ds*.

3. FACE RECOGNITION MODULE

PCA is a statistical dimensionality reduction method which produces optimal linear least squares decomposition of a training set Kirby and Sirovich (1990) applied PCA to representing faces and Truk and Pentland (1991) extended PCA to recognizing faces. In PCA based face recognition algorithm, the input is training set $t_{1\ldots}t_N$ of N facial images such that ensemble mean of the training set is zero ($\sum_i t_1 = 0$). In computing the PCA representation, each image is interpreted as a point in $IR^{N\times M}$, where each image is N×M pixel.PCA finds the optimal linear least square representation in N-1 dimensional space, with the representation preserving variance .the PCA representation is characterized by a set of N-1 Eigen vectors ($e_1,\ldots e_{N-1}$) and Eigen values ($\lambda_1\ldots.\ \lambda_{N-1}$).in the face recognition literature, Eigen vectors can be refereed as a Eigen faces. We normalize the Eigen vectors so that they are orthonormal. The Eigen vectors are ordered so that $\lambda_1 \geq \lambda_{i+1}$. The λ_i are equal to the variance of the projection of the training set on to the i^{th} Eigen vector. Thus the lower order Eigen vectors encode the larger variations in the training set. The lower order refers to the index of Eigen vectors and Eigen values. The higher order Eigen vectors encode smaller variations, it is commonly assumed that they represent noise in the training set because of this assumption and empirical results, higher order Eigen vectors are excluded from the representation. Faces are represented by their projection on to a subset of M≤ N-1 Eigen vectors, which we call face space .thus a facial image is represented as a point in an M-dimensional face space.

The first step is normalization of the input image. The goal of the normalization step is to transform the facial image into a standard format that removes or attenuates variations that can affect recognition performance. This step consists of four substeps. The first substep low-pass filters or compresses the original image. Images are filtered to remove high-frequency noise. An image is compressed to save storage space and reduce transmission time. The second substep places the face in a standard geometric position by rotating, scaling, and translating the center of eyes to standard locations. The goal of this substep is to remove variations in size, orientation, and location of the face in an image.

The third substep masks background pixels, hair, and clothes. This prevents image variations that are not directly related to the face from interfering with identification process. The fourth substep attenuates illumination variation among images, which is a critical factor in the algorithm performance. The second step performs the PCA decomposition on the training set, which produces the eigenvectors and eigenvalues. The third step identifies the face in a normalized image the face in a normalized image, and consists of two substeps.

The first substep projects the image into face space. The critical parameter in this substep is the subset of eigenvectors used to represent the face. The second substep identifies faces with a nearest-neighbor classifier. Or, more precisely, the classifier ranks the gallery images by similarity to the probe. The critical design decision in this step is the similarity measure in the classifier. We presented performance result using L1 distance, L2 distance, angle between feature vectors, and Mahalanobis distance. Additionally, we created three new similarity measures by combining the distance with the L1, L2 and angle similarity measures.

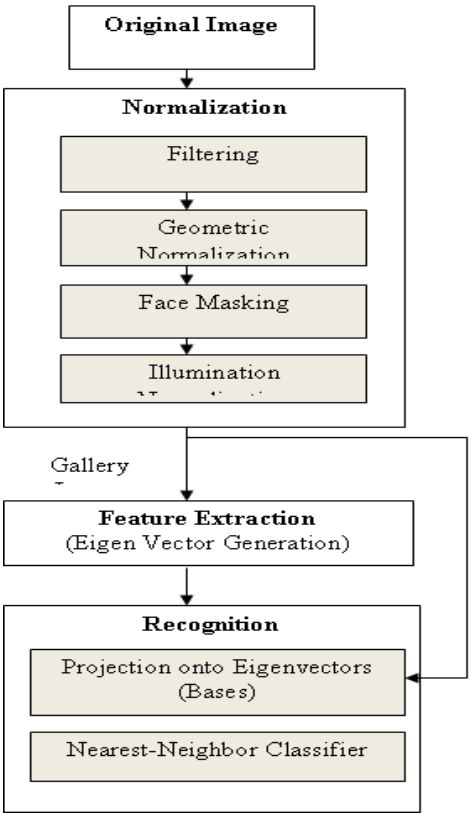

Figure 4 Process Flow of the Face recognition module

4. RESULTS OF EXPERIMENT

Face recognition based on PCA is implemented by using MATLAB. Speaker identification is incorporated with for better recognition. The GUI is created as shown Figure 5.

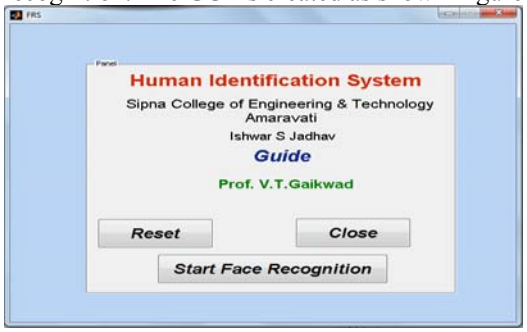

Figure 5 GUI for Face Recognition

Faces are captured and tested using this GUI. Recognition results are visualized in GUI. Speaker Identification is implemented with MATLAB. The Results are shown as in Figure 6-9.

Figure 6 Trained databases.

Figure 7 Test database

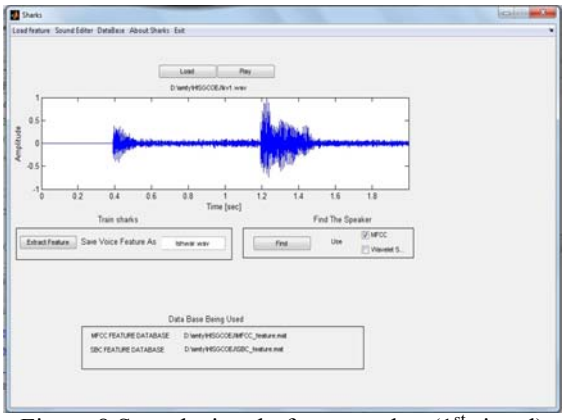

Figure 8 Speech signal of test speaker (1st signal)

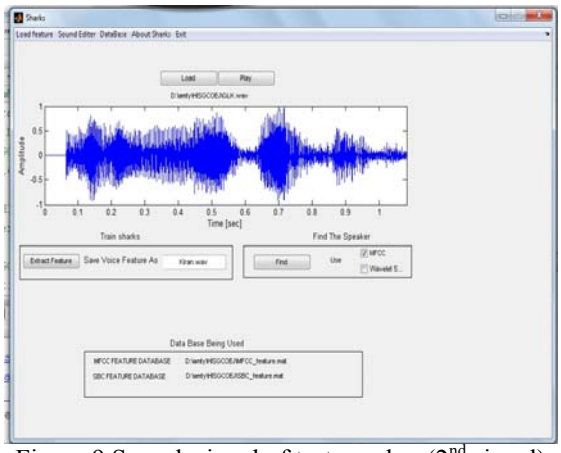

Figure 9 Speech signal of test speaker (2nd signal)

5 CONCLUSION AND FUTURE WORK

The Eigen face approach for Face Recognition and MFCC for voice recognition process is fast and simple which works well under constrained environment. From this project, we increase the level of security. If any one of face and voice is not match with the database then the person is denied for the next step or he was not recognized.

Online Human Identification using face and voice recognition is still a very challenging topic after decades of exploration. A number of typical algorithms are presented separately, being categorized into appearance-based schemes.

6 REFERENCES:

1. Furui, S.: An Overview of Speaker Recognition Technology. In: ESCA Workshop on Automatic Speaker Recognition, Identification and Verification (1994).

2. Pawlewski, M., Jones, J.: Speaker Verification: Part 1. Biometric Technology Today 14(6), 9–11 (2006).

3. Reynolds, D.: A Gaussian Mixture Modeling Approach to Text-independent Speaker Identification. PhD Thesis, Georgia Institute of Technology (1992).

4. McLachlan, G.: Mixture Models, vol. Wright, J. and Yang, A. and Ganesh, A. and Sastri, S, S. and Ma, Y. Marcel Dekker, New York (1988).

5. Tishby, N.: On the Application of Mixture AR Hidden Markov Models to Text independent Speaker Recognition. IEEE Trans. on Signal Proc. 39, 563–570 (1991).

6. Poritz, A.: Linear Predictive Hidden Markov Models and the Speech Signal. In: Proceedings of IEEE ICASSP, pp. 1291–1294 (1982).

7. Rosenberg, A.: Sub-word Talker Verification using Hidden Markov Models. In: Proceeding of IEEE ICASSP, pp. 269–272 (1990).

8. Levinson, D.: A Perspective on Speech Recognition. Communication Magazine 28 (1990)

9. Kohata, M.: Interpolation of LSP Coefficients using Recurrent Neural Networks. Electronics Letters 32 (1996).

10. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face Recognition: A Literature Survey. ACM Computing Survey 35(4), 399–458 (2003)

11. Turk, M., Pentland, A.: Eigenfaces for Recognition. Journal of Cognitive Neuroscience 3, 71–86 (1991).

12. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. IEEE Trans. on PAMI 19, 711–720 (1997).

13. Sanderson, C., Paliwal, K.: Identity Verification Using Speech and Face Information. Digital Signal Processing 14(5), 449–480 (2004).

14. Sanderson, C.: Biometric Person Recognition: Face, Speech and Fusion. VDM Verlag (2008)

15. Turk, Matthew A. and Pentland, Alex P., Face Recognition Using Eigenfaces. *Proc. IEEE Conference on Computer Vision and Pattern Recognition,* Maui, Hawaii, 1991.

16. Etemad, K. and Chellappa, R., Discriminant Analysis for Recognition of Human Face Image. *Journal of Optical Society of America A*, Vol. 14, pp. 1724-1733, 1997.

17. Zhou, S. and Chellappa, R., Multiple-exemplar discriminant analysis for face recognition, *Proc. of the 17th International Conference on Pattern Recognition, ICPR'04*, 23-26 August 2004, Cambridge, UK, pp. 191-194.
18. Moghaddam, B., Jebara, T. and Pentland, A., Bayesian Modeling of Facial Similarity, *Advances in Neural Information Processing Systems 11*, MIT Press, 1999.
19. Moore, B.: Information Extraction and Perceptual Grouping in the Auditory System.Human and Machine Perception: Information Fusion (1997).
20. Haung, X., Acero, A., Hon, H.: Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall PTR, New Jersey (2001).
21. Moore, B.: Frequency Analysis and Masking. Academic Press, USA (1995).