IJITKM 0973-4414 Volume-8, Number-2 Jan-Jun 2015 pp. 88-93 available online at www.csjournalss.com

A Prototype for a Multimodal Biometric Security System Based on Face and Audio Signatures

Garima Yadav

Abstract—Any automatically measurable, robust and distinctive physical characteristic or personal trait that can be used to identify an individual or verify the claimed identity of an individual, referred to as biometrics, has gained significant interest in the wake of heightened concerns about security and rapid advancements in networking, communication and mobility. Multimodal biometrics is expected to be ultra-secure and reliable, due to the presence of multiple and independent - verification clues. In this study, a multimodal biometric system utilising audio and facial signatures has been implemented and error analysis has been carried out. A total of one thousand face images and 250 sound tracks of 50 users are used for training the proposed system. To account for the attempts of the unregistered signatures data of 25 new users are tested. The short term spectral features were extracted from the sound data and Vector Quantization was done using K-means algorithm. Face images are identified based on Eigen face approach using Principal Component Analysis. The success rate of multimodal system using speech and face is higher when compared to individual unimodal recognition systems.
Keywords: Principal Component Analysis, Multimodal Biometrics, Spectral density, Feature vector.

1. INTRODUCTION

The establishment of the identity of a person in a reliable and time-efficient manner is of paramount importance as the society is becoming increasingly dependent on the use of information technology for everyday tasks. Of the many automatic identification technologies, the methods based on biometrics have gained considerable attention due to its robustness as well as reliability. Biometrics is a measurable distinctive physical characteristic or personal trait that can be used to identify an individual or to verify the claimed identity of an individual.

Even though, the biometric identification systems out-perform peer technologies, the unimodal biometric systems have to contend with a variety of problems, namely, noisy data, intra-class variations, restricted degrees of freedom, non-universality and spoof attacks. Many of these limitations can be addressed by deploying multimodal biometric systems that integrate the evidences presented by multiple sources of information.

Using human face as a key to security, the biometrics face recognition technology has received significant attention in the past several years. Face biometrics is used for a wide variety of applications in both law enforcement as well as non-law enforcement. Facial recognition records the spatial geometry of distinguishing features of the face. As compared with other biometrics systems using fingerprint/palmprint and iris, face recognition has a distinct advantage because face images can be captured from a distance without touching the person being identified.

Voice of a person holds certain unique characteristics which can be utilized for personal authentication. Voice is a very intuitive behavioural and ubiquitous biometric which can be captured by modern personal computer.

Multimodal biometric technology uses more than one biometric identifier to compare the identity of the person. A video image of a person speaking a pass phrase can be used in combination for authentication, which is going to make the identification system more reliable. In the present study, a prototype has been implemented which incorporates both face as well as speech for identification purposes.

2. SPEAKER RECOGNITION

Speaker recognition is referred as recognizing person from their voices. Two individuals sounds are not identical because of the physical differences such as their vocal tract shapes, larynx sizes etc and also due to characteristic manner of speaking like the accent, rhythm, pronunciation pattern etc.

The process of speaker identification can be divided into two main phases, namely, the training phase or enrolment phase and the testing phase or identification phase. During the speaker enrolment phase or the training mode, speech samples that contain the discriminating features are collected from the speakers and feature vectors are formed which are used to train the model. In the recognition phase, the feature vectors extracted from the unknown person's utterance are compared against the model in the system database to find the similarity score, for taking a decision. Feature selection is of great importance in speech recognition, as accuracy is highly dependent on the type and number of features used. Features have been computed from the spectrum of the speech signal and relates directly to some perceptual characteristics of sound, such as loudness, pitch, etc. Most of the features are generated from the spectrogram on a frame-by-frame basis.

2.1. Feature Extraction: Short term spectral features

The short-term spectral feature as the name suggest are computed from short frames of about 20-30 millisecond duration of continuously changing speech signal due to articulatory movements. Within this interval, the signal is assumed to remain stationary and the spectral feature vector is extracted from each frame. Typical standard features considered in this prototype for speaker recognition includes Spectral Centroid, Spectral Roll off, Spectral Flux and MFCCs.

2.1.1. Spectral Centroid

The spectral centroid[1], is the centroid of the magnitude spectrum of short time fourier transform and is a measure of spectral brightness. This simple, yet efficient parameter is estimated by summing together the product of each frequency component of the spectrum and its magnitude and then normalized by dividing with the sum of all the spectral magnitudes. Thus the spectral centroid SC which is given by

$$SC=\frac{\sum_{k=0}^{N/2-1} f_k S_k}{\sum_{k=0}^{N/2-1} S_k} \quad (1)$$

where S_k is the magnitude spectrum of the k^{th} frequency component f_k and N is the record size.

2.1.2 Spectral Roll off

Another spectral feature, which gives a measure of the spectral shape, is the spectral roll off[1] and is defined as the frequency below which 85% of the magnitude distribution of the signal is concentrated.

i.e. $RO = Minimum(R)$, such that

$$\sum_{k=0}^{R} S_k \geq 0.85 \sum_{k=0}^{N-1} S_k \quad (2)$$

2.1.3. Spectral Flux

Spectral flux[1] which is a measure of the amount of local spectral change, can be defined as the squared difference between the normalized magnitude spectra of successive frames.

$$Flux = \sum (norm_f[i] - norm_{f-1}[i])^2 \quad (3)$$

where $norm_f$ is the magnitude spectrum of the current frame, scaled to the range 0 to 1 and $norm_{f-1}$ is the normalised magnitude spectrum of the previous frame. Spectral flux is a measure of how quickly the power spectrum of the signal is changing and is computed by comparing the power spectrum of one frame with that of the previous frame.

2.1.4. Mel Frequency Cepstral Coefficient

The Mel Frequency Cepstral Coefficient (MFCC)[2] are computed with the aid of a psychoacoustically motivated filterbank, followed by logarithmic compression and discrete cosine transform(DCT). The outputs of M channel filterbank is denoted as Y(m), m=1…M the MFCCs are obtained as follows

$$C_n = \sum_{m=1}^{M} [\log Y(m)] \cos\left[\frac{\pi n}{M}\left(m-\frac{1}{2}\right)\right] \quad (4)$$

where n is index of the cepstral coefficient.

2.2. Speaker Modeling: Vector Quantization

Vector Quantization(VQ) model also known as centroid model is one of the simplest text-independent speaker model. VQ maps the large set of short term spectral feature vectors extracted in to a finite number of clusters which is represented by its centroid. The clustering is done by K-means clustering algorithm. This reduced set of feature vectors is known as codebook. A speaker database is developed consisting of N codebooks, one for each speaker[3]. In recognition phase, the test utterance features denoted as $X=\{x_1....x_T\}$ of unknown speaker is compared with all the reference vectors denoted as $R=\{r_1...r_k\}$ of known speakers in the database. The average quantization distortion is given in [2] and defined as

$$D_q(X, R) = \frac{1}{T}\sum_{t=1}^{T} \min_{1 \le k \le K} d(x_t, r_k) \quad (5)$$

where $d(.,.)$ is the Euclidean distance $\|x_t - r_k\|$.

A smaller value of equation (5) indicates higher likelihood for X and R originating from the same speaker.

3. FACE RECOGNITION

Principal Component Analysis (PCA)[4] is used to reduce the dimension of data by means of data compression techniques and reveals the most effective low dimensional structure of facial patterns. PCA extracts only components with the largest magnitudes and the dimension reduction removes the unwanted information[5]. This precisely decomposes the face structure into uncorrelated components known as eigen faces. Each face will be stored as a weighted sum of the eigen faces, which are stored in a 1D array. In eigenface approach[6], after the dimensional reduction of the face space, the distance is measured between two images for recognition. If the distance is less than some threshold value, then it is considered as a known face else it is an unknown face

3.1. Eigen Face Approach

Let a face image I(x,y) be a two dimensional NxN array of intensity values. An image may also be considered as a vector of dimension N^2 . Let the training set of face images be $T_1,T_2 \ldots T_m$. The average face of the set as given in [7] is defined by

$$\Psi = \frac{1}{M}\sum_{n=1}^{M} T_n \quad . \quad (6)$$

Each face differs from the average by vector $\phi_i = T_i - \Psi$ and is given in [7]. This set of very large vectors is then subjected to principal component analysis, which seeks a set of M orthonormal vectors u_n and their associated eigenvalues λ_k which best describes the distribution of data. The vector u_k and scalar λ_k are the eigenvectors and eigen values respectively, of the covariance matrix C and is given in [8]

$$C = \frac{1}{M}\sum_{n=1}^{M} \Phi\Phi_n^T = A.A^T \quad (7)$$

where the matrix A= $[\Phi_1,\Phi_2,\ldots.\Phi_M]$. The background is removed by cropping training images, so that the eigenfaces have zero values outside of the face area. Once the eigenfaces are created, identification becomes a pattern recognition task. The M' eigen vectors with the largest associated eigen values are chosen.

In face recognition phase the test face image is transformed into its eigen face component projected into face space by the operation $w_k = u_k^T(T - \Psi)$ where k=1, 2…. M' and is given in [8]. The weights obtained as above form a vector $\Omega_T = [w_1,w_2,\ldots.w_{M'}]$. The same projection is performed on training images set to obtain a collection of weights $\{\Omega_k\}$. The Euclidean distance between the new image and each training image is defined as $\varepsilon_k = \|\Omega - \Omega_k\|$ and is given in [8]. If the minimum Euclidean distance, ε_k, is less than a threshold θ_ε, then the new face is classified as the face associated with Ω_k, otherwise, it is classified as unknown.

4. METHODOLOGY

In this system, a human verification method using a combination of speech and face information is employed in order to reduce the problems of single biometric verification. In the enrolment phase, twenty face images are captured for each person. Images are of different poses[9] like front and side views, with and without spectacles and with different face expressions. For speech recognition, the system is trained with five repetitions of the passphrase. When the passphrase contains more vowel difference between them, the system will have more accurate recognition and so the pass phrases are selected according.

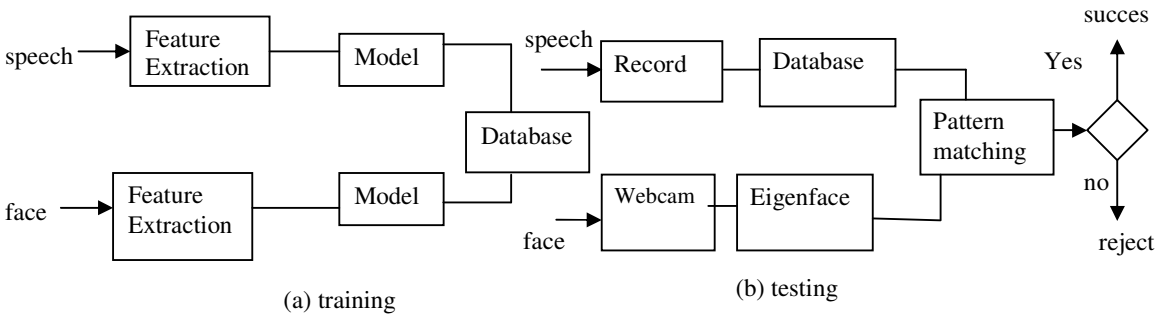

Figure 1: Block diagram of the multimodal speaker and face system

Initially, speech signal, passphrase, of the user is recorded and analyzed in real time. If it is an authorized user the pass phrase is identified from the enrolled database and the eigen faces associated with that identified user is loaded in to the face object. In the next phase, the speaker's image is captured by the webcam and is correlated with the preloaded face object. The access is permitted only if the correlation is 100 percent. In order to ensure that unauthorized personals are not granted permission, the minimum Euclidean distance calculated should be less than a preset threshold value. Thus the access is permitted based on combined results of two biometric features. Figure 1 shows the block diagram of the methodology.

5. RESULTS AND DISCUSSIONS

A total of fifty users is considered for the training of the proposed system. Altogether one thousand face images and 250 sound records are considered for the training of the system. In the test program the data of 50 users who had been considered for the training and data of additional 25 new users are used. Hence the prototype accounts for the attempts of unregistered users in the system.

The system is also tested to find out the rate of False Acceptance (FAR) and False Rejection (FRR). The False Acceptance means acceptance of impostors. The False Rejection means rejection of a true claimant. Thus the performance is measured in terms of FAR and FRR.

If I_a is the number of impostors classified as true claimant and I_t is the total number of impostors in classification test, then FAR can be calculated as the ratio of I_a to I_t . If C_r is the total number of true claimant classified as impostors, C_t is the total number of true claimant classification test then FRR is the ratio of C_a to C_t.

The success rate, false acceptance and false rejection rate of speaker and face unimodal system are computed and given in fig 2 & 3 respectively. In speaker unimodal system, the maximum success rate is obtained when the speaker threshold is assigned a value between 1.0 and 2.2. The average value of success rate in this regime for speaker unimodal system is found to be 0.59.

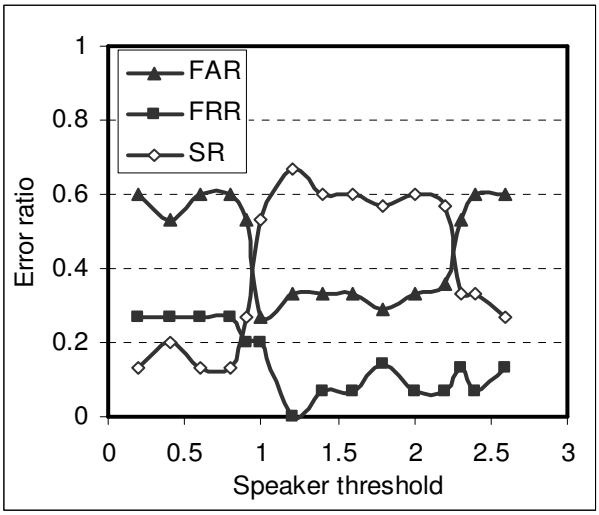

Figure 2: Influence of Speaker Threshold in the proposed System

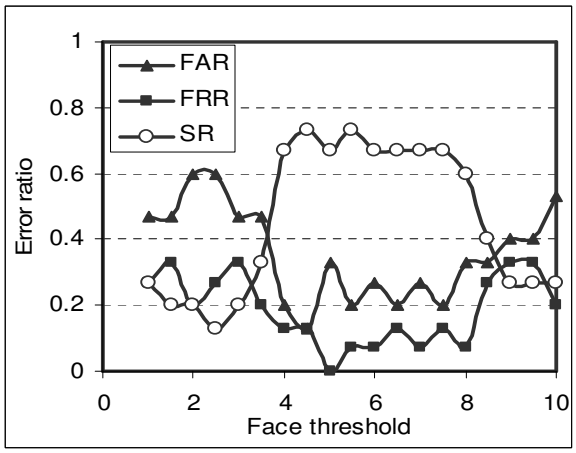

Figure 3: Influence of Face Threshold in the proposed system

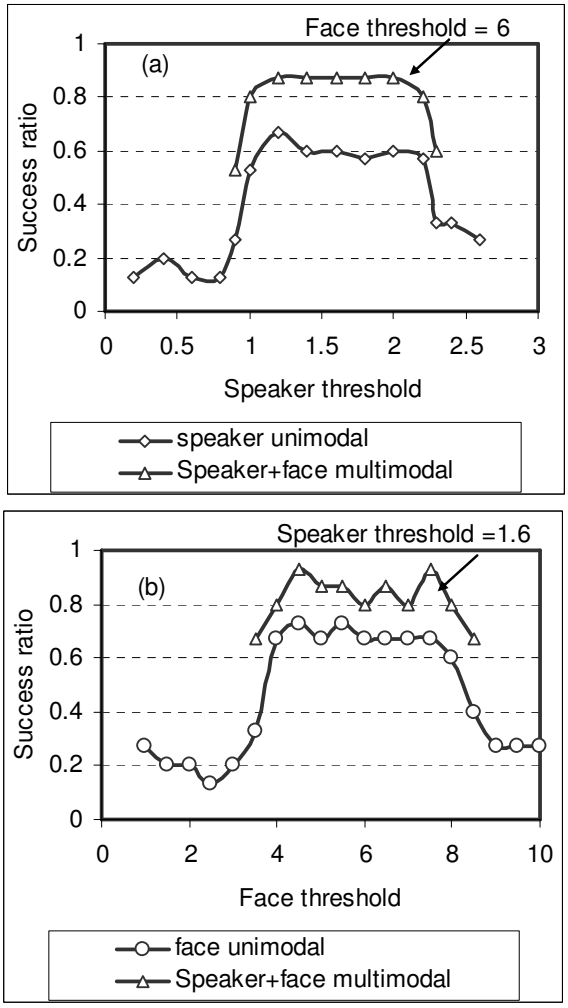

Figure 4: Success rate of face speaker multimodal

In the case of face unimodal system, the magnitude of success rate is found to be maximum when the face threshold is assigned a value between 4 and 8. The average value of success rate in this regime for face unimodal system is found to be 0.67.

The success rate with the multimodal system with speaker and face is computed considering the data of 75 users and is given in fig 4a & 4b. The success rate of face _ speaker multimodal system, when the magnitude of face threshold takes a value of 6.0 is compared with that of speaker unimodal system and is given in Fig 4a. Fig 4b compares the success rate of face_ speaker multimodal system, when the speaker threshold takes a value of 1.6, with the success rate of face unimodal system. The maximum value of success rate for face_speaker multimodal system is obtained when the speaker threshold has a magnitude between 1.0 and 2.2 and face threshold between 4.0 and 8. The average value of success rate over the said region of thresholds is 0.85, which is found to be higher in magnitude when compared to the corresponding data of unimodal system based on speaker or face. This emphasizes the importance and relevance of face_speaker mutimodal system over face or speaker unimodal system.

6. CONCLUSIONS

The magnitude of average success rate in the favourable regime of threshold for face system is higher than the speaker system, which can be attributed to the fact that face system accounts for higher number of biometric traits than the speaker system. The joint analysis of acoustics and visual improves the robustness, as they provide complementary secondary clues that can help in the analysis of the primary biometric signals. In this paper a multimodal biometric human verification method is used to improve the verification rate and reliability in real time. In this system for real time personal verification PCA is used for face recognition and short term spectral features for speech recognition. As a result the proposed system can provide stable verification rate and it overcomes the limitation of single mode system.

REFERENCES

[1]. M.H. Supriya, K.Shaheer, M.G. Mahendran, and P.R.S. Pillai, “Towards Improving the Target Recognition Using a Hierarchical Target Trimming Approach,” WSEAS Transactions on Signal Processing, Vol.3, pp. 340-345, 2007.

[2]. T. Kinnunen, and H. Li, “An overview of text independent speaker recognition from features to supervectors,” J Speech Comm, Vol. 52, pp.12-40, 2009.

[3]. M. Shahneh, and A. Taheri, “Voice Command Recognition System Based on MFCC and VQ Algorithms,” Wor Acad of Sc, Eng and Tech, Vol. 57, pp.534-538, 2009.

[4]. L. Sirovitch, and M. Kirby, “Low-Dimensional Procedure for the Characterization of Human Faces,” J. Opt Soc. of Am, Vol.2, pp. 519-524, 1987.

[5]. H. Moon, and P.J. Phillips, “Computational and Performance aspects of PCA-based Face Recognition Algorithm,” Perception, Vol.30, pp.303-321, 2001.

[6]. M. Turk, and A. Pentland, “Eigenfaces for Recognition,” J of Cognitive Neuros, Vol. 3, pp.71-86, 1991.

[7]. S. Gupta, O.P. Sahu, R. Gupta, and A. Goel, “A Bespoke Approach for Face-Recognition using PCA,” Int J on Comp Sc and Eng, Vol.2, pp.155-158, 2010.

[8]. Y.V. Lata, C.K.B. Tungathurthi, H.R.M. Rao, A. Govardhan, and L.P. Reddy, “Facial Recognition using Eigenfaces by PCA,” Int J. Recent Trends in Eng, Vol.1, pp.587-590, 2009.

[9]. D. Beymer, “Face Recognition Under Varying Pose,” in: Proc. Computer Vision and Pattern Recognition, 1994. pp. 756-761.