A Robust Multimodal Biometric Authentication Scheme with Voice and Face Recognition

H. Kasban

Engineering Department, Nuclear Research Center, Atomic Energy Authority, Cairo, Egypt

Received: 15/11/2016 Accepted: 12/1/2017

ABSTRACT

This paper proposes a multimodal biometric scheme for human authentication based on fusion of voice and face recognition. For voice recognition, three categories of features (statistical coefficients, cepstral coefficients and voice timbre) are used and compared. The voice identification modality is carried out using Gaussian Mixture Model (GMM). For face recognition, three recognition methods (Eigenface, Linear Discriminate Analysis (LDA), and Gabor filter) are used and compared. The combination of voice and face biometrics systems into a single multimodal biometrics system is performed using features fusion and scores fusion. This study shows that the best results are obtained using all the features (cepstral coefficients, statistical coefficients and voice timbre features) for voice recognition, LDA face recognition method and scores fusion for the multimodal biometrics system.

Keywords: Biometrics/Multimodal/ GMM/ Voice Identification/ Face recognition

INTRODUCTION

Biometric techniques are used for human identification and security. Biometrics means the technologies that are used for measuring the human physical characteristics and they are considered very promising tools in human authentication. Most of the used biometric authentication systems now utilize a unimodal biometric authentication system for carrying out the authentication process. The unimodal biometric authentication system recognizes the person based on a single sensor of biometric information such as fingerprint, face, voice, hand, gait, ear, retina, iris, palm print, or signature. Many researchers presented state of the art, surveys and comparisons between the different unimodal biometric methods (1-4). Unimodal biometric faces many challenges such as:

- The noise in the captured raw data that comes from the natural resources surrounding the sensor may cause incorrect labeling of the person and increases the false negative rate.
- The unimodal biometric authentication system may not be able to capture meaningful biometric data from some persons due to the failure of enroll error.
- The biometric authentication system may be suffering from spoofing attacks when an impostor attempts to impersonate the trait matching to a validly enrolled subject.

To overcome the limitations of a unimodal biometric authentication system, a combination of biometric systems can be used by utilizing an approach that combines multiple sources of biometric information into a single decision, this is called multimodal biometric authentication scheme. A good study about several systems and architectures related to the multimodal biometric authentication schemes presented by Lumine et al. (5). The multimodal biometric authentication scheme improves the matching accuracy of the authentication process and performs more reliability and security than the unimodal biometric authentication system because it takes more than one behavioral or physiological characteristics of the person into account to identify that person.
The most important challenge that faces the implementation of the multimodal biometric authentication scheme is the fusion of the different modality inputs such as the voice signal and the face image, because the fusion process should be perform considering the specific modality of the biometric inputs. The multimodal biometric system information fusion may be performed before classification or after classification. In the before classification fusion, the information are combined before applying the matching algorithm, while in the after classification fusion, the information is combined after application of the matching algorithm (6).

Many researchers have presented different multimodal biometric schemes for person verification using voice and face. The reasons of combination of the voice and the face are that they are easy to acquire in a short time with acceptable accuracy using low cost technology. Poh and Korczak presented a hybrid prototype for person authentication using face and text dependent voice biometrics (7). In this prototype, the features vector is extracted from the face information using the moments and the speech information extracted using the wavelets. The extracted features are classified using two separate multilayers. The results achieved an Equal Error Rate (EER) for the face recognition equal to 0.15%, while the EER for the voice recognition equal to 0.07% (7).

Chetty and Wagner (8) proposed a robust multilevel fusion strategy involving a hybrid cascaded multimodal fusion of audio, Two Dimensional (2D) lip face motion, Three Dimensional (3D) face correlation & depth, and tri-module (audio, lip motion, and correlation & depth) for biometric person authentication. The extracted audio features vector is the Mel Frequency Cepstral Coefficients (MFCC) features, while the features vector extracted from the face images consists of three types of features; Discrete Cosine Transform (DCT) features, the explicit grid based lip motion (GRD) features and the contour based lip motion (CTR) features. The features vector extracted from the 3D face are; 3D shape and texture features. The audio signals are degraded by additive white Gaussian noise and the visual speech degraded with JPEG compression. The results achieved an EER equal to 42.9% for audio, 32% for lip face motion, 15% for 3D face, and 7.3% for tri-module (8).

Palanivel and Yegnanarayana (9) proposed a multimodal person authentication approach based on speech, face and visual speech. The face recognition is performed using Morphological Dynamic Link Architecture (MDLA) method for The extracted features vector of the speech is the Weighted Linear Prediction Cepstral Coefficients (WLPCCs) features, while the features vector of the face extracted using the morphological operations. The extracted features are classified using autoassociative neural network (AANN). The results achieved 2.5% EER for the face, 9.2% for the voice and 0.45% for the multimodal (9).

Raghavendra et. al (10) presented a person verification approach based on voice and face. The features vector for the voice recognition is the WLPCCs features, while the features vector of the face recognition is extracted using 2D LDA. The fusion of these features has been carried out using GMM. The results achieved an EER for the face recognition equal to 2.1%, for the voice recognition equal to 2.7% and for the multimodal equal to 1.2% (10).

Elmir et. al (11) presented a hierarchical multimodal method for person authentication based on voice and face. MFCCs features extracted from the voice and the Gabor filter bank is used to construct the face features vector. The Cosine Mahalanobis Distance (CMD) is used for measuring the similarity between the projection coefficients. The results achieved an EER for the face recognition equal to 1.02%, for the voice recognition equal to 22.37% and for the multimodal equal to 0.39% (11).

Soltane (12) presented a person authentication approach based on face and voice recognition. The features vector for the voice is the MFCC features, while the features vector of the face recognition is extracted using eigenfaces. The fusion of these features has been carried out using GMM. The results achieved an EER for the face recognition equal to 0.39995%, for the voice recognition equal to 0.00539% and for the multimodal equal to 0.28125% (12).

Table (1) summarizes some multimodal biometric schemes for person verification using face and voice.
This paper proposes a multimodal biometric scheme for person authentication based on fusion of voice and face recognition. The rest of this paper is arranged as follows; after the introduction section the second section presents the voice recognition. The third section presents the face recognition. The fourth section presents the fusions scheme. The fifth section presents the simulation results. Finally, the last section gives the conclusions remarks.

**VOICE RECOGNITION**

Several researchers have presented voice recognition as a unimodal biometric personal authentication system (13-17). The main advantages of using the voice as a biometric personal authentication are; the voice biometric is an intuitive and natural technology because it uses the human voice, it is low cost technology, and it can provide remote authentication without the need for user presence. The limitations of using the voice as a biometric personal authentication are; the variability of the speech by background noises and temporary voice alterations, it is poor accuracy and low security, and it's suffering from the cross channel conditions. The block diagram of the voice recognition approach used in this paper is shown in Fig. (1), it operates in two modes; training mode and recognition mode.

<table>
<thead>
<tr>
<th>Multimodal biometric scheme</th>
<th>Extracted Features</th>
<th>Fusion Technique</th>
<th>Database</th>
<th>Results (EER %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poh et. al (7)</td>
<td>Moments</td>
<td>Wavelet</td>
<td>No Fusion</td>
<td>Person</td>
</tr>
<tr>
<td>Chetty et. al (8)</td>
<td>DCT, GRD, CTR</td>
<td>MFCC</td>
<td>GMM</td>
<td>AVOZES</td>
</tr>
<tr>
<td>Palanivel et. al. (9)</td>
<td>MDLA</td>
<td>WLPCC</td>
<td>GMM</td>
<td>Newspapers</td>
</tr>
<tr>
<td>Raghavendra (10)</td>
<td>2D LDA</td>
<td>LPCC</td>
<td>GMM</td>
<td>VidTIMIT</td>
</tr>
<tr>
<td>Elmir et. al. (11)</td>
<td>Gabor filter</td>
<td>MFCCs</td>
<td>CMD</td>
<td>VidTIMIT</td>
</tr>
<tr>
<td>Soltane (12)</td>
<td>Eigenfaces</td>
<td>MFCC</td>
<td>GMM</td>
<td>eNTERFACE</td>
</tr>
</tbody>
</table>

**Figure (1):** Speaker identification approach
The first step in voice training mode is the features extraction process that converts the voice signal into features vector. In this paper, three categories of features are used. The first category is the statistical coefficients features that are related to the voice signal such as the mean, the standard deviation, the median, the third quartile and the dominant. The second category is the voice features extracted in the form of the cepstral coefficients such as Mel Frequency Cepstral Coefficients (MFCCs), Linear Prediction Coefficients (LPCs), and Linear Prediction Cepstral Coefficients (LPCCs). The third category is the voice timbre features such as the bass, the baritone, the tenor, the alto, and the soprano\(^{18}\).

The second step in voice training mode is speaker modeling that are carried out using Gaussian Mixture Models (GMMs). The GMM model contains a finite number of Gaussian distributions defined by three parameters; the weights \( w_j \), the mean vectors \( \mu_j \), and the covariance matrices \( \Sigma_j \). These parameters estimated using the Expectation Maximization (EM) algorithm \(^{19}\). For an input vector \( X = \{X_i, \ldots, X_m\} \), the log likelihood \( L \) of the GMM can be defined by \(^{19}\):

\[
L = \log p(X/ \lambda_j) - \log p(X/ \lambda_{ij}) \tag{1}
\]

where \( \lambda_j = (w_j, \mu_j, \Sigma_j) \) and \( \lambda_{ij} = (w_{ij}, \mu_{ij}, \Sigma_{ij}) \) are the model of speaker \( j \) and the background model of speaker \( j \).

In the voice recognition mode, after degrading the voice signal, the features vector are extracted from the voice signal as in the training mode. After that, the pattern matching is carried out by measuring the probability density of the observation given by the Gaussian. The likelihood of the features vector defined by the GMM is the weighted sum over the likelihoods of the Gaussian densities that defined as:

\[
P(x_i, \lambda) = \sum_{j=1}^{n} w_j b(x_i, \lambda_j) \tag{2}
\]

The likelihood of \( x_i \) given \( j^{th} \) Gaussian mixture is:

\[
b(x_i, \lambda_j) = \frac{1}{(2\pi)^{D/2}|\Sigma_j|} \exp \left\{ -\frac{1}{2} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j) \right\} \tag{3}
\]

Where \( D \) is the vector dimension, \( \mu_j \) and \( \Sigma_j \) are the mean vectors and covariance matrices of the training vectors respectively.

Pattern matching is carried out by calculating the matching score between the stored features in the speaker model database and the given model in the recognition mode. The extracted features in the recognition mode are compared with the stored features in the speaker model database and finally the decision is made. The decision is taken using the basis of the matching score, then it accepted as a genuine speaker or it rejected as an imposter speaker.

**FACE RECOGNITION**

The face verification technique involves two main stages; face detection or localisation and face recognition. The face detection means determining the face in the whole image. The face recognition means obtaining the similarity between the detected face image and the stored templates in a database to determine the identity of the person. Many face recognition approaches were presented by many researchers \(^{19, 24}\). In this paper, three face recognition methods are used; eigenface based face recognition \(^{25}\), LDA based face recognition \(^{10}\), and Gabor filter based face recognition \(^{11, 26}\).
Eigenfaces Face Recognition Method

Also, it is called Principal Components Analysis (PCA) based face recognition method. This method consists of two stages: the training stage and the operational stage. In the training stage, a set of the training images that contain the distribution of the face images in a lower dimensional subspace (Eigenspace) is determined. Consider a set of face images \( i_1, i_2, \ldots, i_M \), \( M \) is the number of images, then the average face image of this set is:

\[
\bar{i} = \frac{1}{M} \sum_{j=1}^{M} (i_j) \tag{4}
\]

The difference between each face image and the average face image is:

\[
\Phi_i = i_n - \bar{i} \tag{5}
\]

The covariance matrix of the image is constructed:

\[
C = \sum_{j=1}^{M} \phi_j \phi_j^T = AA^T , \quad A = [\phi_1, \phi_2, \ldots, \phi_M] \tag{6}
\]

Then calculate the eigenvalues \( \lambda_k \) and the eigenvectors \( \upsilon_k \). The eigenvectors determine the linear combination of \( M \) difference images with \( \Phi \) to form the Eigenfaces \( \upsilon_i \):

\[
\upsilon_i = \sum_{k=1}^{M} \upsilon_{ik} \phi_k , \quad l = 1, \ldots, M \tag{7}
\]

Finally, select the Eigenfaces corresponding to the highest eigenvalues at \( K=M \).

In the operational stage, the face image is projected onto the same Eigenspace and then computing the similarity between the input face image and the stored template in the database to take the final decision.

LDA Face Recognition Method

The Eigenfaces face recognition method deals with the whole face image regardless of the structure. The LDA face recognition method provides a discrimination among the classes, it aims to locate the base of vector provide the best discrimination among the classes, and try to maximize the difference between the classes and minimize it within the same class. If \( m_i \) is the \( i^{th} \) class mean and \( m \) is the global mean, the difference between the classes is a corresponding to the scattering matrix \( S_b \) defined as:

\[
S_b = \sum_{i=1}^{C} (m_i - m)(m_i - m)^T \tag{8}
\]

The difference within the class is corresponding to the scatter matrix \( S_w \) defined as:

\[
S_w = \sum_{i=1}^{C} \sum_{x_k \in C_i} (x_k - m_i)(x_k - m_i)^T \tag{9}
\]

The LDA subspace spanned by a set of vectors, \( W \) satisfies the following:

\[
W = \arg_{w} \max \frac{|W^T S_b W|}{|W^T S_w W|} \tag{10}
\]
Gabor Filter Based Face Recognition Method

Gabor filter based face recognition has achieved success in face recognition because it exploits the visual properties such as; the orientation, the selectivity, the spatial localization and spatial frequency characteristics. It has the following formula in the spatial domain:

\[ G(y, z; f, \theta) = \exp \left\{ -\frac{1}{2} \left[ \frac{y^2}{\sigma_y^2} + \frac{z^2}{\sigma_z^2} \right] \right\} \cos(2\pi f y') \]  \tag{11}

\[ y' = y \sin(\varphi) + z \cos(\varphi) \]  \tag{12}

\[ z' = y \cos(\varphi) - z \sin(\varphi) \]  \tag{13}

where \( f \) is the sinusoidal wave frequency along the direction \( \varphi \) from the \( y \)-axis, \( \delta_y \) and \( \delta_z \) are the space constants of the Gaussian envelope along \( y \) and \( z \) axes, respectively.

Gabor filter bank construct the face vector by adapting the parameters correspond to the localized face images determining by its frequency representations. Gabor filter is applied to the face image and then compute the magnitude responses. The computed magnitude response is sampled and after that, the sampled response is concatenated into single features vector. The features vector is used for training and testing the GMM classifier.

Multimodal Biometric Fusion

Fusion in multimodal biometric schemes that can be done before matching or after matching. The fusion before matching may be sensors fusion or features fusion. The sensors fusion is carried out if the biometric system uses multiple sensors for a single trait. The features fusion is done by combining the different features vectors that are extracted from multiple of biometric systems. The fusion after matching may be scores fusion or decisions fusion. The scores fusion is carried out by combining the individual matching scores to single score according to some rules such as sum, max, min rule or by using a formula such as Likelihood Ratio (LLR). The decisions fusion is carried out when the outputs by different matching techniques are available and it considers the weakest fusion.

In this paper, two fusion methods are used and compared; features fusion and scores fusion. For the features fusion shown in Fig. (2-a), the extracted features vectors are extracted from the voice signal and from the face image are combined in a single features vector, which compares to the enrollment template and assigned the final matching score as a single biometric system. The scores fusion as shown in Fig. (2-b) is based on LLR formula that computes the total fused score by:

\[ S = \frac{p(S_{\text{voice}}|G) \cdot p(S_{\text{face}}|G)}{p(S_{\text{voice}}|I) \cdot p(S_{\text{face}}|I)} \]  \tag{14}

where \( p(.|G) \) is the matching scores probability density function of the genuine person, \( p(.|I) \) is the matching scores probability density function of the impostor person, \( S_{\text{voice}} \) is the matching score of the voice recognition technique, and \( S_{\text{face}} \) is the matching score of the face recognition technique.
RESULTS AND DISCUSSIONS

For testing the performance of the proposed multimodal biometrics system, a voice and face database are collected for 100 persons and for every person, five pictures are taken (500 faces images) and every person say the same word five times (500 voices signals). The voices signals are sampled at 8 kHz over 3 seconds and the faces images are resized into 512 x 512 pixels in RGB color model. The database is acquired using Lenovo tablet with camera model A3500-HV and standard microphone.

The performance of the proposed scheme has been evaluated using the Receiver Operating Characteristic (ROC) curve and the Equal Error Rate (EER). The ROC curve is a plot of the False Acceptance Rate (FAR) against the False Rejection Rate (FRR). FAR reflects the proportion of zero effort impostors trials misclassified as genuine trials, while FRR reflects the proportion of the genuine trials misclassified as zero effort impostor trials. The EER refers to the point where the FAR and FRR are equal, it is defined as:

$$EER = \frac{FAR + FRR}{2}, \text{ when } FAR = FRR \quad (15)$$

In the voice training mode, 300 voice signals (3 signals / person) are used. The statistical coefficients features, the cepstral coefficients features and the voice timbre features are used individual and together in order to obtain the best results of voice recognition process. The remaining 200 voice signals (2 signals / person) are used for testing in the voice recognition mode. During the testing, the voice signals are degraded with Additive White Gaussian Noise (AWGN) in order to test the robustness of the proposed scheme. Fig. (3) shows the ROCs curves for the voice recognition using different features vectors. The equal error line (EEL) curve shows the values of EERs at intersecting with ROCs curves. Table (2) compares the values of the EER for the voice recognition using different features vectors.

![Fig. (2): Multimodal biometric fusion](image-url)
The results in Fig. (3) and table (2) show that, the cepstral coefficients features give the lowest EER among the other features extracting method. The reason is that in the cepstral features, any periodicities, or repeated patterns in the spectrum is mapped to one or two specific components in the cepstrum and leads to separate the harmonic series such as the spectrum separates repetitive time patterns in the waveform. Using the statistical coefficients and the voice timbre features improves the performance of the voice recognition technique, so that, in this paper, all features are used training and testing in the voice recognition technique.

For face recognition, 300 faces images (3 images / person ) are used for training and the remaining 200 faces images (2 images / person ) are used for testing the three recognition methods; Eigenface, LDA, and Gabor in order to select the method that gives the best results of face recognition process. During the testing, some of faces images are degraded with JPEG compression in order to test the robustness of the face recognition approaches. Fig. (4) shows the ROCs curves for the different face recognition methods. Table (3) compares the values of the EER for the three face recognition methods.

**Table (2):** EER for the voice recognition using different features vectors

<table>
<thead>
<tr>
<th>Features vector</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical coefficients</td>
<td>8.754</td>
</tr>
<tr>
<td>Cepstral coefficients</td>
<td>4.583</td>
</tr>
<tr>
<td>Voice timbre</td>
<td>5.412</td>
</tr>
<tr>
<td>All features (statistical coefficients + Cepstral coefficients + Voice timbre)</td>
<td>2.983</td>
</tr>
</tbody>
</table>

**Table (3):** EER for the different face recognition methods

<table>
<thead>
<tr>
<th>Face recognition method</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eigenface</td>
<td>4.451</td>
</tr>
<tr>
<td>LDA</td>
<td>1.373</td>
</tr>
<tr>
<td>Gabor</td>
<td>2.99</td>
</tr>
</tbody>
</table>
The results in Fig. (4) and table (3) show that, the FAR and FRR of the LDA face recognition method are less than the FAR and FRR and it gives the lowest EER among the other methods, so that, in this paper, the LDA face recognition method is used for the face recognition. The LDA is more robust because it finds the optimal projective direction by maximizing the difference between class scatter and minimizing it within the same class scatter.

Fig. (5) shows the ROCs curves for the individual voice recognition, individual face recognition and after the fusion using features fusion and scores fusion.
The results show that the features fusion gives EER equal to 2.81, and scores fusion gives the lowest EER equal to 0.64. Scores fusion gives the best results because it takes into consideration the different biometric traits based on their strength and weaknesses for different users, then the collected information will lead to the correct identification of the user. In addition to the LLR between the genuine and impostor distribution minimizes the probability of error. Furthermore, the obtained results are compared with some published results as shown in table (4). The results reveal the ability of the proposed approach as a promising multimodal fusion approach.

**Table (4):** Comparison between the obtained EER of the proposed scheme with the other published results

<table>
<thead>
<tr>
<th>Authentication method</th>
<th>Results (EER %)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Voice</td>
<td>Face</td>
<td>Fusion</td>
<td></td>
</tr>
<tr>
<td>Poh and Korczak (7)</td>
<td>0.07</td>
<td>0.15</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Chetty and Wagner (8)</td>
<td>4.2</td>
<td>3.2</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>Palanivel and Yegnanarayana (9)</td>
<td>9.2</td>
<td>2.9</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>Raghavendra et. al. (10)</td>
<td>2.7</td>
<td>2.1</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>Elmir et. al. (11)</td>
<td>22.37</td>
<td>1.02</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>Soltane (12)</td>
<td>0.01</td>
<td>0.39</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>Proposed</td>
<td>2.24</td>
<td>1.95</td>
<td>0.64</td>
<td></td>
</tr>
</tbody>
</table>

**CONCLUSIONS**

The paper presented a scheme that fused the voice and face recognition as a multimodal biometrics system for human authentication. Both the voice and the face recognitions were carried out using different methods in order to obtain the method that gives the best results of the recognition process. The results of voice recognition process show that the best results are obtained using three categories of features (cepstral coefficients, statistical coefficients and voice timbre). The results of face recognition process show that the LDA based face recognition method is the best face recognition method among the other tested methods. The fusion results show that the scores fusion gives the lowest EER and that it is considered a promising multimodal fusion approach. The proposed method achieved EER equal to 0.64.

**REFERENCES**

(6) A. Jain, K. Nandakumar and A. Ross, Pattern Recognition, 38, 2270, (2005).
(22) S. Xuan, S. Xiang and H. Ma, Computer Vision, 10(6), 493, (2016).