Multimedia Tools and Applications (2019) 78:16345–16361
https://doi.org/10.1007/s11042-018-7012-3

Multimodal biometric scheme for human authentication technique based on voice and face recognition fusion

Anter Abozaid[1] · Ayman Haggag[1] · Hany Kasban[2] · Mostafa Eltokhy[1]

Received: 6 March 2018 / Revised: 31 October 2018 / Accepted: 30 November 2018 /
Published online: 15 December 2018

Abstract
In this paper, an effective multimodal biometric identification approach for human authentication tool based on face and voice recognition fusion is proposed. Cepstral coefficients and statistical coefficients are employed to extract features of voice recognition and these two coefficients are compared. Face recognition features are extracted utilizing different extraction techniques, Eigenface and Principle Component Analysis (PCA) and the results are compared. Voice and face identification modality are performed using different three classifiers, Gaussian Mixture Model (GMM), Artificial Neural Network (ANN), and Support Vector Machine (SVM). The combination of biometrics systems, voice and face, into a single multimodal biometric system is performed using features fusion and scores fusion. The computer simulation experiments reveal that better results are given in case of utilizing for voice recognition the cepstral coefficients and statistical coefficients and in case of face, Eigenface and SVM experiment gives better results for face recognition. Also, in the proposed multimodal biometrics system the scores fusion performs better than other scenarios.

Keywords Multimodal biometrics · SVM · ANN · GMM · Voice identification · Face recognition

1 Introduction

Biometric techniques are utilized for human identification and security. Biometrics means the technology that is applied for measuring the human physical characteristics and is considered to be very promising tools in human authentication. Most of the used biometric authentication techniques now take the advantage of a unimodal biometric authentication scheme for executing the authenticating process. The unimodal biometric authentication technique differentiates the person

✉ Anter Abozaid
anter19731973@gmail.com

[1] Electronics Technology Department, Faculty of Industrial Education, Helwan University, Cairo, Egypt

[2] Engineering Department, Nuclear Research Center, Atomic Energy Authority, Cairo, Egypt

based on only one sensor of biometric data such as fingerprint, face, voice, hand, palm print, walk, ear, retina, iris,, or signature. Many researchers presented state of the art, surveyed and differentiated between different unimodal biometric methods [1, 14, 21].

Unimodal biometric faces many confrontations such as: the noise in the captured raw data that hails from the natural resources encircling the sensor may cause mistaken labeled by the person and increases the false negative rate. The authentication utilizing the unimodal biometric may not be able to hold significant biometric data from some persons due to the defeat of enroll error. The authentication through biometric system may be suffering from spoofing attacks when an impostor attempts to impersonate the trait matching to a validly enrolled subject [3]. To overcome the challenges of a unimodal biometric authentication system, a combination of different biometric systems can be used by employing an approach that combines numerous sources of biometric input into a single decision; in this case, the scheme is denoted by multimodal biometric authentication.

Authentication using multimodal biometric schemes has been presented in [25]. Good study about diverse systems and architectures linked to multimodal idea is presented in [25]. The biometric authentication by multimodal schemes promotes the matching accuracy of the authentication process and accomplishes more reliability and security than the unimodal biometric system because it takes combination from different behavioral or physiological characteristics of the person into account to distinguish that person. This model likes data or image security using merging and combining different multi-level security techniques [11].

The most important challenge that faces the implementing of the multimodal biometric scheme is the different modality inputs fusion such as the face image and the voice signal for example, because the fusion process should be performing considering the particular modality of the biometric inputs. The information fusion of the multimodal biometric schemes may be performed before classification or after classification. In the before classification fusion, the information integrated before applying the matching algorithm, while in the after classification fusion, the information integrated after application of the matching algorithm [18].

The rest of this paper is arranged as follows: in section 2, the related work is presented. The voice recognition technique is introduced in section 3. In section 4, the face recognition techniques are discussed. The multimodal Biometric Fusion is presented in section 5. The section 6 presents the computer simulation experiments results. Finally, the last section gives the conclusions remarks.

2 Related work

In this section, previous research work and recently published papers are presented. Many researchers have presented different multimodal biometric schemes for person verification using voice and face. The reasons of integration of the voice and the face are that they are easy to acquire in a short time with acceptable accuracy using low cost technology. Srinivas Halvi et al. in 2017 presented proposed face recognition model based on transform domain and fusion technique. The proposed model is given in Fig. 1.

In this proposed model, two transform domain techniques are utilized which as the DWT and FFT techniques as shown in Fig. 1. In this proposed technique, the extracted feature from the DWT and FFT are compare utilizing Euclidian Distance (ED) for computing the parameters performance [16]. Biometric security is utilized to enhance the Wireless Body Area Network (WBAN) security. WBAN is wireless network for medical applications; it can be considered special branch form the Wireless Sensor Networks (WSNs) [9, 10, 29].

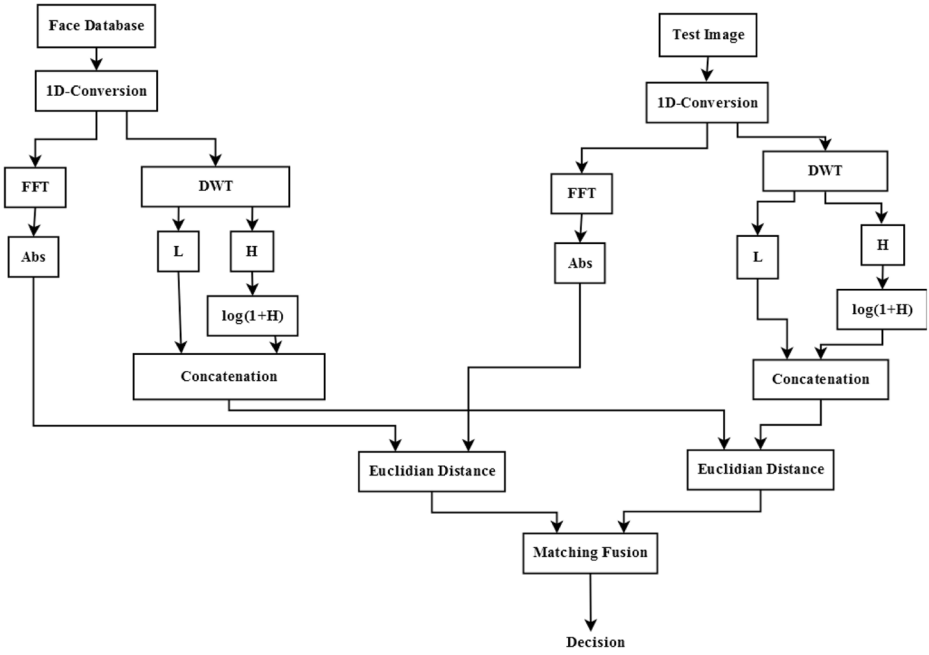

Fig. 1 Block diagram of face recognition and fusion technique model [16]

Poh and Korczak showed a hybrid prototype using combining face and text dependent voice biometrics to implement person authentication [28]. In this prototype the features vector excerpted the face information using the moments and the speech information excerpted using the wavelets. The derived features are classified using two separate multilayers. The results of this scheme achieved an Equal Error Rate (EER) equal to 0.15% and 0.07% for the face recognition and the voice recognition, respectively [28].

Chetty and Wagner suggested a powerful multilevel fusion strategy including a hybrid cascaded multimodal fusion of audio, Two-Dimensional (2-D) lip face motion, Three-Dimensional (3-D) face correlation & depth, and tri-module (audio, lip motion, and correlation & depth) for biometric person authentication. The excerpted audio features vector is the Mel Frequency Cepstral Coefficients (MFCC) features, while the features vector excerpted 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 3-D face are; 3-D shape and texture features. The audio signals are degenerated by additive white Gaussian noise and the visual speech degenerated with JPEG compression. The results of thee presented technique achieved an EER equal to 42.9%, 32%, 15% and 7.3% for audio, lip face motion, 3-D face and tri-module, respectively [6].

Palanivel and Yegnanarayana suggested a multimodal person authentication way based on speech, face and visual speech. The face differentiation is performed using Morphological Dynamic Link Architecture (MDLA) method for the excerpted features vector of the speech is the Weighted Linear Prediction Cepstral Coefficients (WLPCCs) features, while the features vector of the face excerpted using the morphological operations. The excerpted features are categorized using Auto Associative Neural Network (AANN). The result of EER for the face and the voice was 2.5% and 9.2%, respectively; the EER equals 0.45% for the multimodal [27].

Raghavendra et al. shown a person verification way based on voice and face. The features vector for the voice differentiation is the WLPCC features, while the features vector of the face differentiation is excerpted using 2D LDA. The fusion of these features has been accomplished using GMM. The results accomplished an EER for the face differentiation equal to 2.1%, for the voice differentiation equal to 2.7% and for the multimodal equal to 1.2% [30]. In [12], the person authentication by hierarchical multimodal method based on face and voice is presented. MFCCs features excerpted from the voice and the Gabor filter bank is used to establish the face features vector. The Cosine Mahalanobis Distance (CMD) is used for measuring the similarity between the planning coefficients. The results achieved an EER for the face differentiation equal to 1.02%, for the voice differentiation equal to 22.37% and for the multimodal equal to 0.39%.

In [32] biometric authentication technique based on face and voice differentiation is presented. In this research paper, the features vector for the voice is the MFCC features, while the features vector of the face differentiation is excerpted using eigenfaces. The fusion of these features has been accomplished using GMM. The results accomplished an EER for the face differentiation equal to 0.39995%, for the voice differentiation equal to 0.00539% and for the multimodal equal to 0.28125% [32]. Another proposed person authentication technique based on face and voice differentiation is presented in [19]. The features vector for the voice is the MFCCs, LPCs and LPCCs features, while the features vector of the face differentiation is excerpted using PCA, LDA and Gabor filter. The fusion of these features has been accomplished using LLR. The results achieved an EER for the face differentiation equal to 1.95%, for the voice differentiation equal to 2.24% and for the multimodal equal to 0. 64% [19]. Table 1 summarizes some multimodal biometric schemes for person verification using face and voice with/without fusion technique.

In this research paper, combined multimodal biometrics scheme is proposed for person authentication based on fusion of voice and face recognition. The proposed scheme utilizes combination of different three biometric modalities to establish reliable biometric identification system. The fusion process is carried out using two methods, feature fusion which uses the extracted feature and score fusion by using the score.

3 Voice recognition

Several researchers have submitted voice differentiation as a unimodal biometric personal authentication system. The there are some advantages of using the voice as a biometric

Table 1 Differet schemes of Multimodal biometric using face and voice

Multimodal biometric scheme	Extracted features		Fusion technique	Database	Results (EER %)		
	Face	Voice			Face	Voice	Fusion
Poh et. al [28]	Moments	Wavelet	No Fusion	Persons	0.15	0.07	–
Chetty et. al [6]	DCT, GRD, CTR	MFCC	GMM	AVOZES	3.2	4.2	0.73
Palanivel et. al. [27]	MDLA	WLPCC	GMM	Newspapers	2.9	9.2	0.45
Raghavendra [30]	2D LDA	LPCC	GMM	VidTIMIT	2.1	2.7	1.2
Elmir et. al. [12]	Gabor filter	MFCCs	CMD	VidTIMIT	1.02	22.37	0.39
Soltane [32]	Eigenfaces	MFCC	GMM	eNTERFACE	0.399	0.0054	0.281
H. kasban [19]	PCA,LDA, Gabor filter	MFCCs, LPCs, LPCCs	LLR	PROPOSED	1.95	2.24	0.64

personal authentication are the voice biometric is an intuitive and natural technology because it uses the human voice, it can supply remote authentication without the need for user presence and it is low cost technology also. The disadvantages of using the voice as a biometric authentication are; the speech variability by background noises and temporary voice alterations, it is low security and poor accuracy, and it's suffering from the cross channel conditions. The block diagram of the voice recognition approach used in this research paper is shown in Fig. 2, it operates in two modes; training mode and recognition mode [17, 26].

The first step in voice training mode is the features excerption process that transforms the voice signal into features vector. In this paper, two categories of features are used. The first category is the statistical coefficients features that linked 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 excerpted in the form of the cepstral coefficients such as Mel Frequency Cepstral Coefficients (MFCCs), Linear Prediction Coefficients (LPCs), and Linear Prediction Cepstral Coefficients (LPCCs) [17]. The second step in voice training mode is speaker modeling that carried out using three classifiers, Artificial Neural Network (ANN), Support Vector Machine (SVM), and Gaussian Mixture Model (GMM) [23].

3.1 Artificial neural network (ANN)

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is made up of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is prepared for a specific application, such as pattern differentiation or data classification, through a learning process. Learning in biological systems includes adjustments to the synaptic connections that stay between the neurons [20]. The main idea of Holder-function is next. Consider $f(t) \in D_f$ Holder derived, that $|f(t+\Delta t)-f(t)| \le const(\Delta t)^{\alpha(t)}$, $\alpha(t) \in [0, 1]$

$$\begin{array}{l} \alpha = 0 \text{ means that we have break of second order} \\ \alpha = 1 \text{ means that have } O(\Delta t) \end{array}$$

This formula is a somewhat connection between "bad" functions and "good" functions. If we will look on this formula with more precise we will notice, that we can catch moments in time,

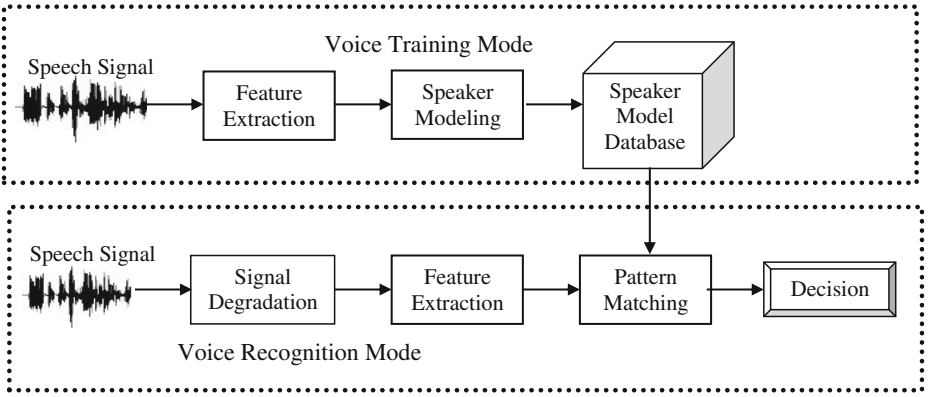

Fig. 2 Speaker identification approach

when our function knows, that it's going to change its behavior from one to another. It means that today we can make a forecast on tomorrow behavior. But one should mention that we don't know the sigh on what behavior is going to change [15].

3.2 Support vector machine (SVM)

Support Vector Machine (SVM) is a classification and regression prediction tool that uses machine learning theory to make the most of predictive accuracy while automatically avoiding over-fit to the data. Support Vector machines can be realized as systems which use hypothesis space of a linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that executes a learning bias derived from statistical learning theory.

In this we present the QP formulation for SVM classification. This is a simple representation only. SV classification in Eq. (1) as shown:

$$\min_{f,\xi_i} \|f\|_K^2 + C \sum_{i=1}^{l} \xi_i \quad y\,if\,(xi) \geq 1-\xi i, \text{for all } i\,\xi_i \geq 0 \tag{1}$$

SVM classification, Dual formulation:

$$\min_{\alpha_i} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j)\ 0 \leq \alpha_i \leq C, \text{for all i;} \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{2}$$

Variables ξ_i are called slack variables and they measure the error made at point (x_i,y_i). Training SVM becomes quite challenging when the number of training points is large. A number of methods for fast SVM training have been proposed [4].

3.3 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 μ_j, and the covariance matrices ε_j. These parameters estimated using the Expectation Maximization (EM) algorithm. For an input vector $X = \{X_1, \ldots, X_m\}$, the log likelihood L of the GMM can be defined by [31] as given Eq. (3):

$$L = log\ p(X/\lambda_j) - log\ p(X/\lambda_{j1}) \tag{3}$$

where $\lambda_j = (w_j, \mu_j, \varepsilon_j)$ and $\lambda_{j1} = (w_{j1}, \mu_{j1}\ \varepsilon_{j1})$ are the model of speaker j and the background model of speaker j_1.

In the voice differentiation mode, after degrading the voice signal, the features vector excerpted from the voice signal as in the training mode. After that, the pattern matching executed by measuring the probability density of the observation given by the Gaussian. The likelihood of the features vector realized by the GMM is the weighted sum over the likelihoods of the Gaussian densities that realized as shown in Eq. (4):

$$P(x_i, \lambda) = \sum_{j=1}^{M} w_j b(x_i, \lambda_j) \tag{4}$$

The likelihood of x_i given j^{th} Gaussian mixture is given in Eq. (5):

$$b(x_i, \lambda_j) = \frac{1}{(2\pi)^{D/2}|\varepsilon_j|} \exp\left\{ -\frac{1}{2} \left(x_i - \mu_j\right)^T \sum_j^{-1} \left(x_i - \mu_j\right) \right\} \tag{5}$$

Where D is the vector dimension, μ_j and ε_j are the mean vectors and covariance matrices of the training vectors respectively.

Pattern matching is executed 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 compare 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.

4 Face recognition

Face verification technique contains two main stages; face detection or localisation and face recognition. The face detection means determining the face in the whole image. As shown in Fig. 3, the block diagram of face recognition model is given. The face recognition means obtaining the similarity between the detected face image and the stored templates in a database to determine the personality of the person. Many face recognition approaches presented by many researchers [8, 24, 36]. In this paper, two face differentiation methods are used; Eigen face-based face differentiation, and Principal Components Analysis (PCA) [13, 22, 33, 35].

4.1 Eigenfaces face recognition method

Also, it is called Principal Components Analysis (PCA) based face recognition method. This method consists of two stages; training stage and 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 $i_1, i_2, \ldots\ldots, i_M$, M is the number of images, then the average face image of this set is [2]:

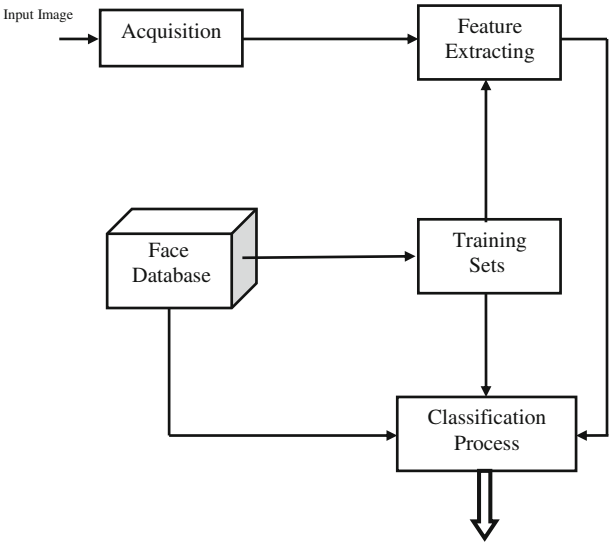

Fig. 3 Block diagram of face recognition Model

$$\bar{i} = \frac{1}{M}\sum_{j=1}^{M}(i_j) \tag{6}$$

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

$$\emptyset_n = i_n - \bar{i}$$

The covariance matrix of the image is constructed:

$$C = \sum_{j=1}^{M} \varphi_j \varphi_j^T = \mathrm{AA}^{\mathrm{T}}, \quad \mathrm{A} = [\varphi_1 \varphi_2 \ldots\ldots\ldots \varphi_M] \tag{8}$$

Then calculate the eigenvalues λ_k and the eigenvectors υ_k. The eigenvectors determine the linear combination of M difference images with Ø to form the Eigenfaces υ_i:

$$\nu_l = \sum_{k=1}^{M} \nu_{\mathrm{lk}} \varphi_k, \quad \mathrm{l} = 1, \ldots\ldots\ldots\ldots, \mathrm{M} \tag{9}$$

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 likeness between the input face image and the stored template in the database to take the final decision.

4.2 Principal components analysis (PCA)

The Eigen faces face differentiation method deals with the whole face image regardless to the structure. In principal components analysis (PCA) and factor analysis (FA) one wishes to extract from a set of p variables a reduced set of m components or factors that accounts for most of the variance in the p variables. In other words, we wish to reduce a set of p variables to a set of m underlying super ordinate dimensions [5].

These underlying factors are inferred from the correlations among the p variables. Each factor is estimated as a weighted sum of the p variables. The ith factor is thus expressed in Eq. (10).

$$F_i = W_{i1}X_1 + W_{i2}X_2 + \ldots + W_{ip}X_p \tag{10}$$

One may also express each of the p variables as a linear combination of the m factors, as shown in Eq. (11)

$$X_j = A_{1j}F_1 + A_{2j}F_2 + \ldots + A_{mj}F_m + U_j \tag{11}$$

where U_j is the variance that is unique to variable j, variance that cannot be explained by any of the common factors.

5 Multimodal biometric fusion

Fusion in multimodal biometric schemes can be done before matching or after matching, the fusion before matching may be sensors fusion or features fusion. The sensors fusion is executed if the biometric system utilizes multiple sensors for a single trait. The features fusion is done by combining the different features vectors that extracted from multiple of biometric systems. The fusion after matching may be scores fusion or decisions fusion. The scores fusion

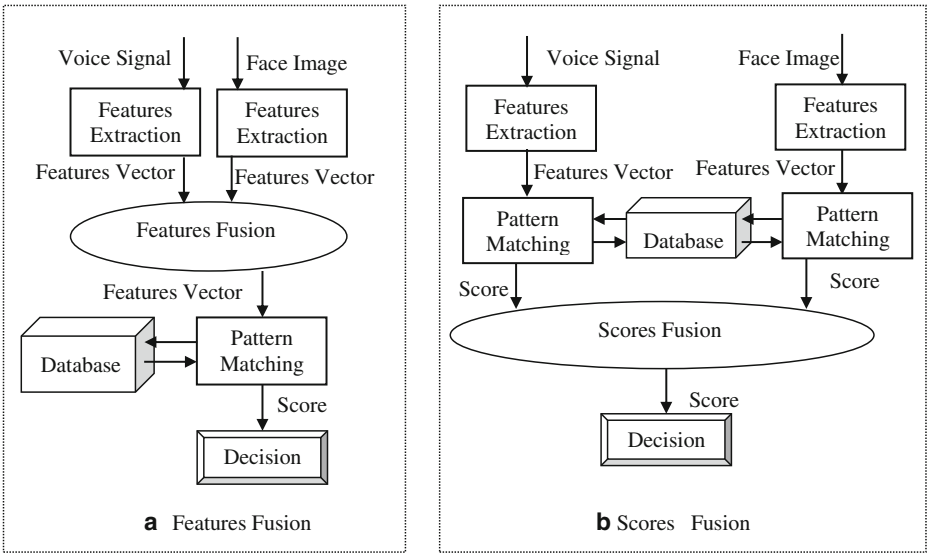

Fig. 4 The proposed block diagram of multimodal biometric fusion scheme

is carried out by combining the individual matching scores to single score according to some rules such as sum, max, and 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 [7].

In this paper, two fusion methods are used and compared; features fusion and scores fusion. Results of the features fusion are shown in Fig. 4a, the extracted features vectors 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 is given in Fig. 4b is based on LLR formula that computes the total fused score by [34], as given in Eq. (12):

$$S = \frac{p(S_{voice}|G) \ . \ p(S_{face}|G)}{p(S_{voice}|I) \ . \ p(S_{face}|I)} \tag{12}$$

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_{voice} is the matching score of the voice recognition technique, and S_{face} is the matching score of the face differentiation technique [28].

6 Results and discussions

For testing the performance of proposed multimodal biometrics system, a voice and face database are collected for 100 persons, 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 s and the faces images are resized into 512 × 512 pixels in RGB color model. The database is acquired using Lenovo tablet with camera model A3500-

Fig. 5 Some of face images from the used image database

HV and standard microphone. Figure 5 shows samples from the used images database, which is utilized in the computer simulation experiments for evaluating the proposed multimodal scheme.

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

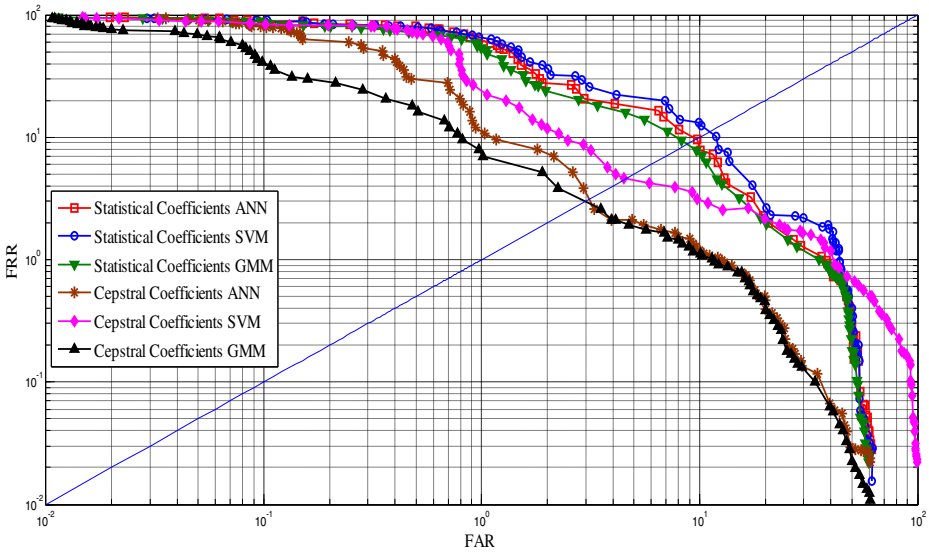

Fig. 6 ROC curves for the voice recognition using different features vectors

zero effort impostor trials. The EER refers to the point where the FAR and FRR are equal, it is defined as shown in Eq. (13):

$$EER = \frac{FAR + \mathrm{FRR}}{2}, \text{when}\,\mathrm{FAR} = \mathrm{FRR} \tag{13}$$

In the voice training mode, 300 voice signals (3 signals / person) are used. The statistical coefficients features and the cepstral coefficients features are used individual and together in order to obtain the best results of voice recognition process. The remain 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. Figure 6 shows the ROCs curves for the voice differentiation 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.

The results in Fig. 6 and Table 2 show that, the cepstral coefficients features give the lowest EER among the other features extracting method. The reason is, in the cepstral features, any periodicities, or frequented 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

Table 2 EER for the voice recognition using different features vectors

Voice recognition method	EER (%)
Statistical Coefficients + ANN	10.55
Statistical Coefficients +SVM	10.73
Statistical Coefficients + GMM	9.48
Cepstral Coefficients + ANN	3.15
Cepstral Coefficients + SVM	4.47
Cepstral Coefficients + GMM	2.98

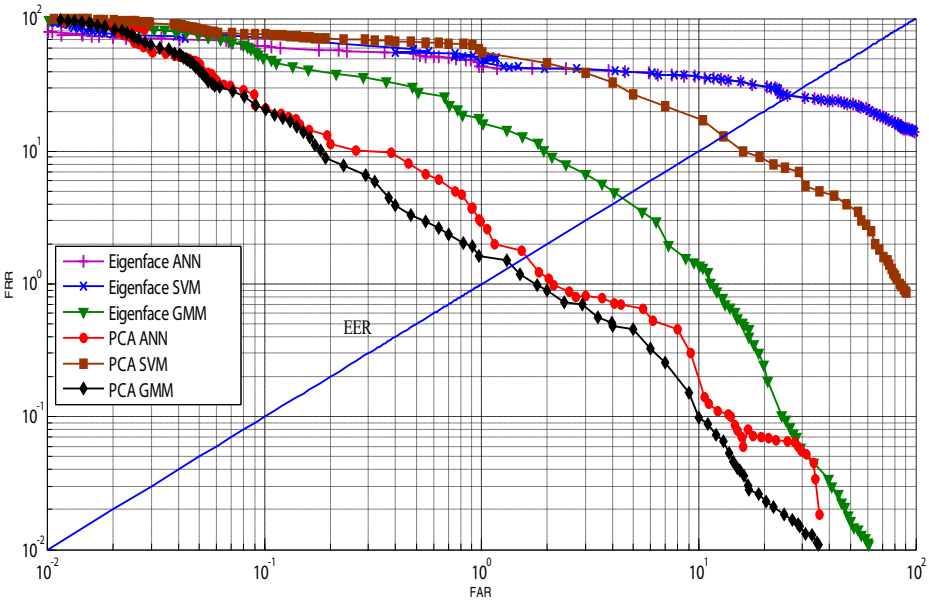

Fig. 7 ROCs curves for the different face recognition methods

separates repetitive time patterns in the waveform. Using the statistical coefficients and the voice timbre features make efficient the performance of the voice recognition technique, so that, in this paper, all features are used training and testing in the voice differentiation technique.

In the face recognition computer experiments, there are 300 faces images (3 images / person) are used for training and the remain 200 faces images (2 images / person) are used for testing the three recognition methods; Eigenface and PCA so as to select the method that gives the best results of face recognition process. During the testing, some of faces images are degenerated with JPEG compression in order to test the robustness of the face differentiation approaches. Figure 7 Shows the ROCs curves for the different face recognition methods. Table 3 compares the values of the EER for the three face recognition methods.

The results of face recognition experiments are shown in Fig. 7 and Table 3. These results clear that, the FAR and FRR of the PCA with GMM 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 PCA face recognition method is used for the face recognition. The PCA with GMM is more robust because it finds the optimal projective direction by maximizing the difference between

Table 3 EER for the different face recognition methods

Voice recognition method	EER (%)
Eigenface + ANN	26.83
Eigenface +SVM	27.12
Eigenface + GMM	4.45
PCA + ANN	1.71
PCA + SVM	13.05
PCA + GMM	1.43

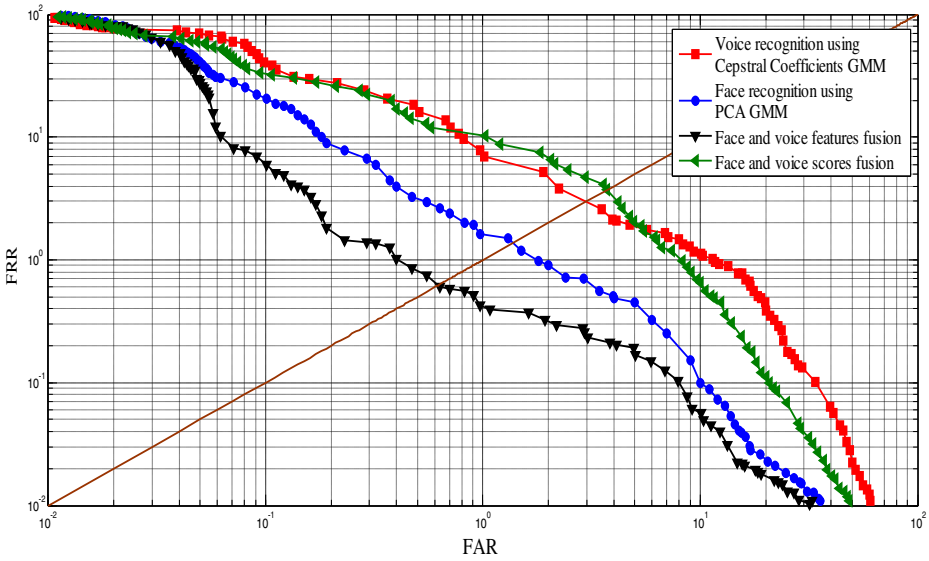

Fig. 8 ROCs curves for the proposed multimodal fusion approach

class scatter and minimizing it within the same class scatter. The two fusion processes results are given in Fig. 8. This figure shows the ROCs curves for the individual voice recognition, individual face recognition and after the fusion using features fusion and scores fusion.

The results of the features fusion and scores fusion experiments clear that the features fusion gives EER equal to 2.81, and scores fusion gives the lowest EER equal to 0.69. Scores fusion gives the best results because it takes in consideration the different biometric traits based on their strength and weaknesses for different users, then the collected information will lead to the right identification of the user. In addition to the LLR between the genuine and impostor distribution reduces the probability of error. Furthermore, the obtained results are compared with some published results as displayed in Table 4. The results reveal the ability of the proposed approach as a promising multimodal fusion approach.

As shown in the final computer simulation experiment results in Fig. 8 and Table 4, the proposed multimodal scheme gives lower EER. Also, it performs better than the previous

Table 4 Comparison between the obtained EER of the proposed scheme with the other published results

Authentication method	Results (EER %)		
	Voice	Face	Fusion
Poh and Korczak [28]	0.07	0.15	–
Chetty and Wagner [6]	4.2	3.2	0.73
Palanivel and Yegnanarayana [27]	9.2	2.9	0.45
Raghavendra et al. [30]	2.7	2.1	1.2
Elmir et al. [12]	2.37	1.02	0.39
Soltane [32]	0.01	0.39	0.28
H. kasban [19]	2.24	1.95	0.64
Proposed scheme	2.98	1.43	0.62

related work as shown in Table 4 which tabulates the EER values of the proposed scheme and the previous work results.

In the future work, triple multimodal will be studied using iris, face and voice. Three biometrics will be combined in one multimodal biometric scheme. Also, biometric security for WBAN using unimodal and combined model will be studied with power consumption and complexity consideration. The third research point in the future work will focus on design and testing wireless combined biometric person authentication system.

7 Conclusions

In this research paper, a fusion scheme for voice and face differentiation as a multimodal biometrics system for human authentication is proposed. Both voice and the face recognition are performed using different feature extraction tools to choose the best for the recognition process. Results of voice recognition process showed that the best results are obtained by simulation of the Cepstral Coefficients using GMM classifier scenario. Results of face recognition process showed that the PCA with the GMM classifier based face differentiation method is the best face recognition method among the other tested methods. The fusion results showed that, the scores fusion gives the lowest EER and considers a promising multimodal fusion approach. The proposed scheme performs better than other biometric schemes. The computer simulation experiments reveal the superiority of the proposed modal for the proposed face recognition modal and the proposed fusion scenarios.

References

1. Abdel Karim N, Shukur Z (2015) Review of user authentication methods in online examination. Asian Journal of Information Technology 14(5):166–175
2. Abhishree TM, Latha J, Manikantan K, Ramachandran S (2015) Face recognition using Gabor filter based feature extraction with anisotropic diffusion as a pre-processing technique. Procedia Computer Science 45: 312–321
3. Agashe NM, Nimbhorkar S (2015) A survey paper on continuous authentication by multimodal biometric. International Journal of Advanced Research in Computer Engineering & Technology 4(11):4247-4253
4. Baken RJ, Orlikoff RF (2000) Clinical measurement of speech and voice – second edition. Singular Publishing Group, San Diego
5. Burges C (1998) A tutorial on support vector machines for pattern recognition. In: Data mining and knowledge discovery (Volume 2). Kluwer Academic Publishers, Boston, pp 1–43
6. Chetty G, Wagner M (2008) Robust face-voice based speaker identity verification using multilevel fusion. Image Vis Comput 26:1249–1260
7. Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge
8. De A, Saha A, Pal MC (2015) A human facial expression recognition model based on Eigen face approach. Procedia Computer Science 45:282–289
9. Dodangeh P, Jahangir AH (2018) A biometric security scheme for wireless body area networks. Journal of Information Security and Applications 41:62–74. https://doi.org/10.1016/j.jisa.2018.06.001
10. El-Bendary MAM (2015) Developing security tools of WSN and WBAN networks applications. Springer, Japan

11. El-Bendary MA (2017) FEC merged with double security approach based on encrypted image steganography for different purpose in the presence of noise and different attack. Multimed Tools Appl 76(24): 26463–26501
12. Elmir Y, Elberrichi Z, Adjoudj R (2014) Multimodal biometric using a hierarchical fusion of a person's face, voice, and online signature. J Inf Process Syst:555-567
13. Fookes C, Lin F, Chandran V, Sridharan S (2012) Evaluation of image resolution and super-resolution on face recognition performance. J Vis Commun Image Represent 23(1):75–93
14. Gad R, El-Fishawy N, El-Sayed A, Zorkany M (2015) Multi-biometric systems: a state of the art survey and research directions. Int J Adv Comput Sci Appl 6(6):128-138
15. Galka J, Masior M, Salasa M (2014) Voice authentication embedded solution for secured access control. IEEE Trans Consum Electron 60(4):653–661
16. Halvia S, Ramapurb N, Rajac KB, Prasadd S (2017) Fusion based face recognition system using 1D transform domains. Procedia Computer Science 115:383–390
17. Inthavisas K, Lopresti D (2012) Secure speech biometric templates for user authentication. IET Biometrics 1(1):46–54
18. Jain A, Nandakumar K, Ross A (2005) Score normalisation in multimodal biometric systems. Pattern Recogn 38:2270–2285
19. Kasban H (2017) A robust multimodal biometric authentication scheme with voice and face recognition. Arab Journal of Nuclear Sciences and Applications 50(3):120–130
20. Kinnunen T, Karpov E, Franti P (2006) Real time speaker identification and verification. IEEE Trans Audio Speech Lang Process 14(1):277–288
21. Kumar HCS, Janardhan NA (2016) An efficient personnel authentication through multi modal biometric system. International Journal of Scientific Engineering and Applied Science 2(1):534–543
22. Li H, Suen CY (2016) Robust face recognition based on dynamic rank representation. Pattern Recogn 60: 13–24
23. Liu Z, Wang H (2014) A novel speech content authentication algorithm based on Bessel–Fourier moments. Digital Signal Process 24:197–208
24. Liu T, Mi JX, Liu Y, Li C (2016) Robust face recognition via sparse boosting representation. Neurocomputing 214:944–957
25. Lumini A, Nanni L (2017) Overview of the combination of biometric matchers. Information Fusion 33:71–85
26. Morgen B (2012) Voice biometrics for customer authentication. Biom Technol Today 2012(2):8–11
27. Palanivel S, Yegnanarayana B (2008) Multimodal person authentication using speech, face and visual speech. Comput Vis Image Underst 109:44–55
28. Poh N, Korczak J (2001) Hybrid biometric person authentication using face and voice features. International Conference, Audio and Video Based Biometric Person Authentication, Halmstad, Sweden, pp 348–353
29. Qia M, Chena J, Chen Y (2018) A secure biometrics-based authentication key exchange protocol for multi-server TMIS using ECC. Comput Methods Prog Biomed 164:101–109
30. Raghavendra R, Rao A, Kumar GH (2010) Multimodal person verification system using face and speech. Procedia Computer Science 2:181–187
31. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted Guassian mixture models. Digital Signal Processing 10:19–41
32. Soltane M (2015) Greedy expectation maximization tuning algorithm of finite GMM based face, voice and signature multi-modal biometric verification fusion systems. International Journal of Engineering & Technology 15(03):41–52
33. Turk M, Pentland A (1991) Eigenfaces for recognition. J Cogn Neurosci 3(1):71–86
34. Vapnik V (1998) Statistical learning theory. Wiley, New York
35. Xuan S, Xiang S, Ma H (2016) Subclass representation-based face-recognition algorithm derived from the structure scatter of training samples. Comput Vis 10(6):493–502
36. Zheng CH, Hou YF, Zhang J (2016) Improved sparse representation with low-rank representation for robust face recognition. Neurocomputing 198:114–124

Anter Abozaid was born in Elbehaira Egypt in 1981. He received his B.Sc. degree from Bani Sweif University, Egypt, in June 2003, Since March 2012, he has been with the Electronics Technology Dept., Faculty of Industrial Education, Helwan University, Egypt. His current research interests are in the fields of Security and Authentication.

Ayman Haggag was born in Cairo, Egypt in 1971. He received his B.Sc. degree from Ain Shams University, Egypt, in June 1994, M.Sc. degree from Eindhoven University of Technology, The Netherlands, in December 1997, and Ph.D. degree from Chiba University, Japan, in September 2008. Since March 1996, he has been with the Electronics Technology Department, Faculty of Industrial Education, Helwan University, Egypt. His current research interests are in the fields of Network Security, Wireless Security, Software Defined Network and Wireless Sensor Network.

Hany Kasban received the B. Sc., M. Sc. and Ph. D. degrees in Electrical and Electronic Engineering from Menoufia University, Egypt in 2002, 2008 and 2012, respectively. He is currently an Associate Professor in the Department of Engineering and Scientific Instruments., Nuclear Research Center (NRC), Egyptian Atomic Energy Authority (EAEA), Cairo, Egypt. He is a co-author of many papers in national and international conference proceedings and journals. His current research areas of interest are in the fields of electronics, digital signal processing, communication systems and nuclear applications in industry and medicine.

Mostafa Eltokhy was born in Kaluobia, Egypt, in 1970. He received his B.Sc.degree from Zagazig university, Banha branch, Egypt, and M.Sc. degree from Technical University, Eindhoven, The Netherlands in 1993 and 1998, respectively. He received his Ph.D. degree from Osaka University, Osaka, Japan in 2003. Presently, he is an Associate Professor of Electronics Engineering at Department of Electronics Technology, Faculty of Industrial Education, Helwan University, Cairo, Egypt. His current research interests are high performance digital circuits and analog circuits. He is a member of the IEEE.