Feature Level Fusion of Face and Voice Biometrics Systems Using Artificial Neural Network for Personal Recognition

Cherifi Dalila, El Affifi Omar Badis and Boushaba Saddek
Institute of Electrical and Electronic Engineering, University of Boumerdes, Algeria
E-mail: da.cherifi@univ-boumerdes.dz

Nait-Ali Amine
Laboratoire Images, Signaux et Systèmes Intelligents (LISSI), Université Paris-EST, Vitry sur Seine, 94400, France
E-mail: naitali@u-pec.fr

Keywords: bioinformatiscs, face, voice, multibiometric recognition system, fusion at feature level, Artificial Neural Network (ANN).

Received: November 23, 2018

Lately, human recognition and identification has acquired much more attention than it had before, due to the fact that computer science nowadays is offering lots of alternatives to solve this problem, aiming to achieve the best security levels. One way is to fuse different modalities as face, voice, fingerprint and other biometric identifiers. The topics of computer vision and machine learning have recently become the state-of-the-art techniques when it comes to solving problems that involve huge amounts of data. One emerging concept is Artificial Neural networks. In this work, we have used both human face and voice to design a Multibiometric recognition system, the fusion is done at the feature level with three different schemes namely, concatenation of pre-normalized features, merging normalized features and multiplication of features extracted from faces and voices. The classification is performed by the means of an Artificial Neural Network. The system performances are to be assessed and compared with the K-nearest-neighbor classifier as well as recent studies done on the subject. An analysis of the results is carried out on the basis Recognition Rates and Equal Error Rates.

Povzetek: Z nevronsko mrežo so kombinirani obraz in glas za biometrično identifikacijo.

1 Introduction

Based on the fact that any biometric system has some weaknesses, it is difficult to obtain a system that accomplishes the four most desirable points for a biometric-based security system which are, Universality, Distinctiveness, Permanence and Collectability [1]. One way to overcome the limitations is through a combination of different biometric systems to reduce the classification problem which deals with the intra-class and inter-class variety [2]. Combinations of biometric traits are mainly preferred due to their lower error rates. Using multiple biometric modalities has been shown to decrease error rates, by providing additional useful information to the classifier. Fusion of these behavioral or physiological traits can occur in various levels. Different features can be used by a single system or separate systems which can operate independently and their decisions may be combined [3-6].

In this article, we have choiced Face and Voice as our biometric traits for several reasons, mainly because of their availability where people can get along with easily, regardless of gender and age. Also, because the data can be acquired simultaneously just by using a camera with an embedded microphone, this way, we avoid steps in data gathering like in the case of face and fingerprint or face and hand geometry, where the recognition algorithm might become time consuming and disables the real time functionality.

Many researchers have presented different multimodal biometric schemes for person verification using voice and face by using different fusion technique and data bases, the authors proposed different methods to extract the features from the face (Discrete Cosine Transform, grid-based lip motion, contour based lip motion, Morphological Dynamic, Link Architecture, 2D LDA, Eigenfaces, PCA, LDA and Gabor filter), and for the voice (Mel Frequency Cepstral Coefficients, Weighted Linear Prediction Cepstral Coefficients, Linear Prediction Coefficients and Linear Prediction Cepstral Coefficients) [7-14].

In this work the fusion is done at the feature level with three different schemes namely, concatenation, merging and multiplication of features extracted from faces and voices. The classification is performed by using two classifiers which are mainly K-Nearest-Neighbor and Artificial Neural Network. The first one is a classical classification method based on distance calculations, whereas the other is an intelligent system that learns in a way similar to the human brain. The complexity of the Neural Network gives it a flexibility and a capability to be tuned to better fit any type of data. In our work, we make a comparative study between the two stated classifiers to conclude whether ANN can be exploited to design better recognition systems.

The rest of the paper is structured as follow: In section 2, we dealt with feature extraction methods for face and voice used in this work. In section 3, we presented our proposed fusion method at feature level based on Artificial Neural Networks and K-nearest-neighbor classifiers. In section 4, the experimental part is described, the results are provided and discussed. Finally, a conclusion of this work is highlighted in section 4.

2 Feature extraction

2.1 Face feature extraction

Face recognition is one of the few biometric methods that possess the merits of both high accuracy and low intrusiveness. It has the accuracy of a physiological approach without being intrusive. For this reason, it has drawn the attention of researchers in fields from security, psychology, and image processing, to computer vision [15]. Numerous algorithms have been proposed and developed for the purpose of Face recognition. These algorithms can be classified into three categories: Global-Appearance-based methods, Local-feature-based methods and Hybrid methods There are methods that use the whole image of the face as a raw input to the learning process, others require the use of specific regions located on a face such as eyes, nose and mouth. There exist also methods that simply partition the input face image into blocks without considering any specific regions. In this work we mainly are going to use PCA and DCT [16].

- Principal Component Analysis (PCA) Method was developed by Turk and Pentland, it's a well-known face recognition method, known as eigenfaces, which drastically reduces the dimensionality of the original space and face detection and identification are carried out in the reduce space [17-19].
- Discrete Cosine Transform (DCT) Method is an invertible linear transform that can express a finite sequence of data points in terms of a sum of cosine functions oscillating at different frequencies [20-21]. Face recognition using DCT is divided into two stages training and classification. In the training stage, the face images are analyzed on block by block basis. The DCT coefficients with large magnitude are mainly located in the upper-left corner of the DCT matrix. Accordingly, we scan the DCT coefficient matrix in a zig-zag manner starting from the upper-left corner and subsequently convert it to a one-dimensional (1-D) vector [22].

2.2 Speech feature extraction

The speech signal conveys many levels of information to the listener (figure 1). At the primary level, speech conveys a message via words. But at other levels speech conveys information about the language being spoken and the emotion, gender and, generally, the identity of the speaker [23]. The general area of speaker recognition encompasses two more fundamental tasks.

Speaker identification is the task of determining who is talking from a set of known voices or speakers. The unknown person makes no identity claim and so the system must perform a 1:N classification. Generally, it is assumed the unknown voice must come from a fixed set of known speakers, thus the task is often referred to as *closed-set* identification.

Speaker verification (also known as speaker authentication or detection) is the task of determining whether a person is who he/she claims to be (a yes/no decision). Since it is generally assumed that imposters (those falsely claiming to be a valid user) are not known to the system, this is referred to as an *open-set* task [23]. Depending on the level of user cooperation and control in an application, the speech used for these tasks can be either *text-dependent* or *text-independent*. In a text-dependent application, the recognition system has prior knowledge of the text to be spoken and it is expected that the user will cooperatively speak this text. In the other hand, in a text-independent application, there is no prior knowledge by the system of the text to be spoken, such as when using extemporaneous speech. Text-independent recognition is more difficult but also more flexible [23], this approach is considered in our work. It is inconvenient to use the whole speech directly as an input for biometric recognition systems. We instead use the features which represent the unique distinctive characteristics that make the difference between speakers for the following reasons [24]:

- The feature extraction process transforms the raw signal into feature vectors in which speaker-specific properties are emphasized and statistical redundancies are suppressed.
- With features extracted, we can avoid the problem of the curse of dimensionality.
- The signal during training and testing session can be greatly different due to many factors such as people voice change with time, health condition (e.g. the speaker has a cold), speaking rate and also acoustical noise and variation recording environment via microphone.

There is several feature extraction approaches for speech, the most popular are: Linear Predictive Analysis (LPC), Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Predictive Coefficients (PLP), Relative Spectra filtering of log domain (RASTA), Mel-Frequency Cepstral Coefficients (MFCC).

2.2.1 Mel-frequency cepstral coefficients

The MFCC feature extraction technique is the most popular approach used in speaker recognition systems today, it has been utilized intensively in literature [25-26] and others. The Mel scale was developed by Stevens and Volkman in 1940 as a result of a study of the human auditory perception. This method is capable of capturing phonetically important characteristics of the speech. MFCC are based on the well-known variation of the human ear’s critical bandwidths with frequency. Steps of the MFCC extraction process are summarized in figure 2 [27].

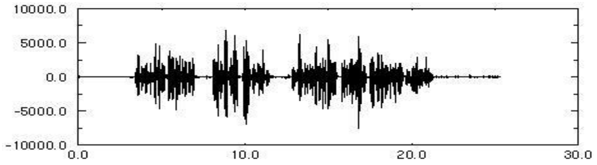

Figure 1: A sample of input speech signal[27].

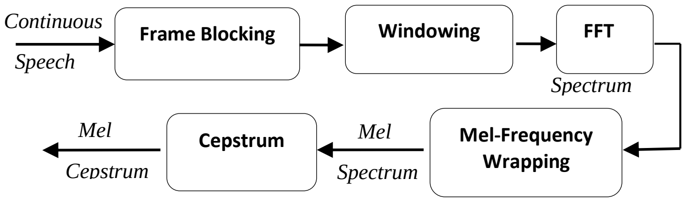

Figure 2: Block diagram of the MFCC process [27].

2.2.2 Vector Quantization (VQ)

Several state-of-the-art feature characterization and matching techniques have been developed and proposed in literature for speaker recognition. Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ). The last one was used in our project, because it is easy to implement. Vector Quantization (VQ) is a process of mapping vectors from a large vector space to some regions in that space. Each region is called a cluster that can be represented by its center which called a codeword. The set of all code words is called a codebook [27]. A speaker-specific VQ codebook is generated from any speaker by clustering his/her training acoustic vectors. The distance from a vector to the closest codeword of a codebook is called VQ-distortion.

2.2.3 Feature scaling

Since the face features vary in a scale of [0, 255], and voice features in a complete different scale [-10,14], a feature normalization must be performed to map the values from their ranges to a range of [0,1] in order to prevent one modality from contributing more than the other in the learning process. We have used the Min-Max normalization rule.

3 Data fusion at feature level

3.1 Proposed fusion schemes

In feature level fusion, feature sets originating from multiple information sensors are integrated into a new feature set. For non-homogeneous compatible feature sets, such as features of different modalities like face and speech as is presented in this article, a single feature vector can be obtained by concatenation [1, 20]. The new feature vector now has a higher dimensionality which increases the computational load. It is reported that a significantly more complex classifier design might be needed to operate on the concatenated data set at the feature level space.

The fusion at the feature level is expected to perform better in comparison with the fusion at the score level and decision level. The main reason is that the feature level contains richer information about the raw biometric data [28]. It is to be noted that a normalization may be necessary because of the **non-homogeneity** of the different traits used in the Multibiometric system. In the present work, we consider performing a data fusion at the feature level between face and voice. This is to be done in three different ways.

3.1.1 Fusion by concatenation (pre-normalized features)

In this Fusion, we concatenate features of a Face sample (Fij) with features of a Voice sample (Vij) to get one large sample, without normalization of the features, taking m samples with n features of each.

$$\begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1n} \\ F_{21} & F_{22} & \cdots & F_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ F_{m1} & F_{m2} & \cdots & F_{mn} \end{bmatrix} Concatinated\ with \begin{bmatrix} V_{11} & V_{12} & \cdots & V_{1n} \\ V_{21} & V_{22} & \cdots & V_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ V_{m1} & V_{m2} & \cdots & V_{mn} \end{bmatrix}$$

$$\text{Giving} \Rightarrow \begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1n} & V_{11} & V_{12} & \cdots & V_{1n} \\ F_{21} & F_{22} & \cdots & F_{2n} & V_{21} & V_{22} & \cdots & V_{2n} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ F_{m1} & F_{m2} & \cdots & F_{mn} & V_{m1} & V_{m2} & \cdots & V_{mn} \end{bmatrix}$$

Figure 3: Fusion by Concatenation.

This has been previously done and stated in literature [29], we apply it in order to see the impact of data normalization and its absence.

3.1.2 Fusion by merging (normalized features)

This is to be done by alternatively placing one face feature, followed by one voice feature, until all features are placed one next to the other with normalized features.

$$\begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1n} \\ F_{21} & F_{22} & \cdots & F_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ F_{m1} & F_{m2} & \cdots & F_{mn} \end{bmatrix} \text{Merged with} \begin{bmatrix} V_{11} & V_{12} & \cdots & V_{1n} \\ V_{21} & V_{22} & \cdots & V_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ V_{m1} & V_{m2} & \cdots & V_{mn} \end{bmatrix}$$

$$\text{Giving} \Rightarrow \begin{bmatrix} F_{11} & V_{11} & \cdots & F_{1n} & V_{1n} \\ F_{21} & V_{21} & \cdots & F_{2n} & V_{2n} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ F_{m1} & V_{m1} & \cdots & F_{mn} & V_{mn} \end{bmatrix}$$

Figure 4: Fusion by Merging.

3.1.3 Fusion by multiplication (normalized features)

This is to be done by multiplying pre-normalized Face features with pre-normalized Voice features element-wise. Then we normalize the resulting product matrix. We did not find a theoretical background for this fusion scheme except considering that features multiplication can be some sort of polynomial terms [30].

$$\begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1n} \\ F_{21} & F_{22} & \cdots & F_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ F_{m1} & F_{m2} & \cdots & F_{mn} \end{bmatrix} .* \begin{bmatrix} V_{11} & V_{12} & \cdots & V_{1n} \\ V_{21} & V_{22} & \cdots & V_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ V_{m1} & V_{m2} & \cdots & V_{mn} \end{bmatrix}$$

$$\text{Giving } \Rightarrow \begin{bmatrix} F_{11} * V_{11} & F_{12} * V_{12} & \cdots & F_{1n} * V_{1n} \\ F_{21} * V_{21} & F_{22} * V_{22} & \cdots & F_{2n} * V_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ F_{m1} * V_{m1} & F_{m2} * V_{m2} & \cdots & F_{mn} * V_{mn} \end{bmatrix}$$

Figure 5: Fusion by Multiplication.

Each of the resulting fused data will be fed to our designed Neural Network system for classification. The results will be compared with the performance of a K-NN classifier as shown in Table 1.

	Face	Voice	Fusion	Feature scale
Method 1	Raw Pixels	MFCC + VQ	Concatenation	Pre-Normalized
Method 2	Raw Pixels	MFCC + VQ	Merged	Normalized
Method 3	Raw Pixels	MFCC + VQ	Multiplied	Normalized
Method 4	PCA	MFCC + VQ	Concatenation	Normalized
Method 5	DCT	MFCC+VQ	Concatenation	Normalized

Table 1: Proposed fusion methods to be experimented.

3.2 Classifiers

3.2.1 K-nearest neighbor algorithm

The idea in k-Nearest Neighbor methods is to identify ***k*** samples in the training set whose independent variables ***x*** are similar to ***u***, and to use these k samples to classify this new sample into a class, v. If all we are prepared to assume is that ***f*** is a smooth function, a reasonable idea is to look for samples in our training data that are near it (in terms of the independent variables) and then to compute ***v*** from the values of y for these samples. When we talk about neighbors, we are implying that there is a distance or dissimilarity measure that we can compute between samples based on the independent variables. One way to perform this task is to use the most popular measure of distance: Euclidean distance.

3.2.2 Artificial neural networks

Neural networks are algorithms that are patterned after the structure of the human brain. They contain a series of mathematical equations that are used to simulate biological processes such as learning and memorizing. In a neural network, the goal as in all modeling techniques (such as Linear regression, Logistic regression, Survival analysis or time-series analysis ...), is predicting an outcome based on the values of some input variables stated that ANNs could be used as alternatives to the foregoing techniques. Neural networks can have one or multiple outputs. In this work, we are dealing with multi-class classification problem, where each person (Face and Voice) is a distinct class, hence the use of a multiple output Neural Network. Although many different types of neural network training algorithms have been developed, we preferred to stick with the famous "back-propagation" algorithm, which is the most popular used technique [31-34] and we have considered the Logistic activation function in our network design represented in Figure 6.

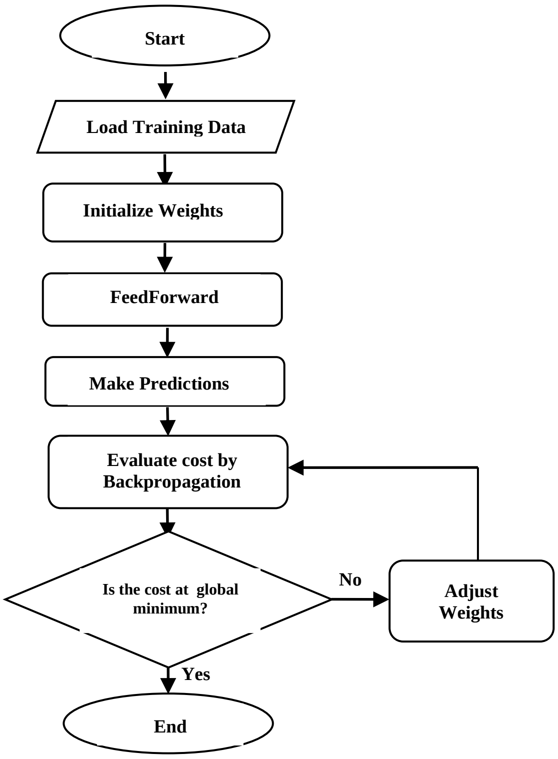

Figure 6: Neural network training.

4 Experimental & results

In the present work, the major aim is to realize a multibiometric system based on a fusion of two main modalities, Face and Voice. This is to be done on the feature level. An Artificial Neural Network is to be designed for the sake of classification. The performance of this system is then compared to the k-NN classification approach involving the use of some classical methods (PCA and DCT) for the face, and MFCC with Vector Quantization for voices.

A fusion is done at the feature level for each of the following systems (Table 1), and then fed to a k-NN classifier. We consider applying this approach on many databases, and compare performances with respect to the Artificial Neural Network design.

4.1 Database description

In this work, we have run into the problem of missing a database that contains both the face and voice of the same person, because it is unlikely for a subject to give away two or three of his identity modalities at once for the sake of a bare scientific experiment. This is generally justified by security and anonymity reasons. In order for us to approach this issue, we have followed some works in literature, in which the authors have combined two or three datasets. The first set is for one modality, taken from a group of subjects at some circumstances, the other set is for another modality taken from a dissimilar group of people at completely different circumstances. Then each modality from set 1 is assigned to the other modality from set 2, thus the fusion is performed by concatenation. The database formed by the procedure just described is usually referred to as a virtual database [25-26].

4.1.1 Face databases: ORL database (AT&T)

There are ten different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). A preview of the faces is in (Figure 7). The files are in PGM format. The size of each image is 112x92 pixels, with 256 grey levels per pixel.

Figure 7: Preview of the ORL database images.

4.1.2 Face databases: FEI database

The FEI face database is a Brazilian face database that contains a set of face images taken between June 2005 and March 2006. There are 14 images for each of 200 individuals, a total of 2800 images. All images are colorful and taken against a white homogenous background in an upright frontal position with profile rotation of up to about 180 degrees. Scale might vary about 10% and the original size of each image is 640x480x3 pixels. All faces are mainly represented by students and staff at FEI, between 19 and 40 years old with distinct appearance, hairstyle, and adorns. The number of male and female subjects are exactly the same and equal to 100. Figure 8 shows a sample of image variations from the FEI face database.

Figure 8: Preview of the FEI database images.

4.1.3 Speech database

We have collected samples that are 12 minutes long from different people reading books from the internet. The utterances were text-independent. Then we adjusted the sampling frequency of every sample to 11025 Hz using audio enhancement software (Audacity). After that, we cropped the long samples at a length of less than 14 seconds making 48 samples per person.

4.1.4 Neural network design

Since there is no rule of thumb for choosing the number of hidden layers as well as the number of neurons contained inside them, we tried a set of configurations with multiple numbers of layers and neurons, and analyzed the behavior of the networks designed at each time.

5 Experiments

In order to evaluate the performance of the proposed methods, we have used some standard indices for assessment. The false acceptance rate (FAR), is the measure of the likelihood that the biometric security system will incorrectly accept an access attempt by an unauthorized user. The false rejection rate (FRR) is the measure of the likelihood that the biometric security system will incorrectly reject an access attempt by an authorized user. The Equal Error Rate (EER) is defined as the point where the value of FAR equals the value of FRR in the Receiver Operating Curve (ROC) which plots FAR versus FRR.

5.1 Experiment I: (ORL without external effects + voice)

We have downsized the ORL images to 40x40 pixels, in order to minimize the amount of calculations as compared to 112x92. The number of subjects is 40. We used ORL images without any external effects, samples of speech are assigned to those face samples for each subject. A detailed description of this database used for this experiment, containing the matrices dimensions before and after fusion (Table 2). The results of this experiment are shown in (Table 3) describing the recognitions rates and equal error rates.

In terms of recognition rate, when trained and tested without external effects, ANN gave better results than K-NN with an average of (96.33 vs 92.66%) still with an insignificant difference ($p=0.081>0.05$). The proposed method 2 (Raw Faces & MFCC + VQ) merged and normalized hit the best accuracy (99.16%). This is because the configuration of the network enables it to fit well trained data and generalize to the test data. In terms of equal error rate, method 5 (DCT of Faces & MFCC + VQ) on K-NN outperformed all the methods (1.73 %) followed by proposed method 2 on ANN (2.5 %) which are close and both very good.

Databases	Details	Training		Testing	
		Face	Voice	Face	Voice
ORL without external effects + Voice	Samples	280x1600	280x1600	120x1600	120x1600
	Fused	280x3200		120x3200	
	Authorized	40 subjects / 7 samples each		40 subjects / 3 samples each	
	Unauthorized	/		160 subjects / 10 samples each	

Table 2: Description of Experiment I databases.

	Features	Classifier	RR (%)	EER (%)	Th (%)	AUC
Raw Faces & MFCC + VQ	Proposed Method 1: Concatenated$^{(pn)}$	ANN	95.83	7.5	51	0.9515
		K-NN	89.16	5.24	60	0.719
	Proposed Method 2: Merged$^{(n)}$	ANN	99.1667	**2.5**	34.4	0.9947
		K-NN	96.66	3.706	60	0.8505
	Proposed Method 3: Multiplied$^{(n)}$	ANN	95.83	14.1667	32.2	0.9137
		K-NN	90	9.237	60	0.7374
PCA for Faces & MFCC+VQ	Concatenated$^{(n)}$	ANN	94.1667	13.32	27.3	0.9351
		K-NN	90.83	3.7125	33.4	0.8905
DCT for Faces & MFCC+VQ	Concatenated$^{(n)}$	ANN	96.667	8.35	33.9	0.9714
		K-NN	96.66	**1.73**	40	0.923

$^{(pn)}$Pre-normalized features.$^{(n)}$ Normalized features.

Table 3: Results with different schemes of fusion and classification.

5.2 Experiment II: (ORL with external effects + voice)

We had to introduce some effects in order to enrich the data, because it is a necessity for the neural network to have different and versatile features to enhance the way it learns the variety of appearances and details. Each image had undergone 5 effects thus the 35 samples per subject. As for voice, we took 35 samples of speech for each subject and assigned them to the faces of the corresponding person. A description is in Table 4. The results of this experiment are shown in Table 5 describing the recognitions rates and equal error rates.

In terms of recognition rate, when trained and tested with external effects, ANN dominated K-NN in every fusion scenario, with an average of (97.66 vs 90.99%) with a significant difference (p=0.027<0.05). A total test recognition was reached by proposed method 2 (Raw Face & MFCC + VQ merged and normalized) on ANN (100%). The next competing system is the classical method 5 (DCT of Faces & MFCC + VQ) and was on ANN as well (99.37vs 96.26% for K-NN). As for EER, proposed method 2 has attained the lowest error on ANN (1.67%) against K-NN (8.1 %) which is a very good result, mainly attained by enriching the system by more samples with external effects.

In methods 3, 4 and 5, K-NN performed better than ANN, with an EER of 13.11 ± 1.9 % in average compared to ANN with EER of 26.91 ± 7.6 % with a significant difference (p=0.03<0.05). These methods are either highly sensitive to noise where features could be altered significantly, or the neural network configuration was not suitable for this kind of data after effects were involved stating the change of illumination by Gaussian noise as well as eyes cover which can prevent the system from recognizing one's identity if it depended on those features.

Even though the eigenvectors on Method 4 were sorted in a descending order with respect to their corresponding eigenvalues, this method gave the worst EER on ANN (35.67%), the same thing with Method 5 (23.14%), this is basically related to the unbalance of the system where face features dimensionality was much less than voice features dimensionality (100 vs 1600) and (144 vs 1600) respectively.

Databases	Details	Training		Testing	
		Face	Voice	Face	Voice
ORL with external effects + Voice	Samples	1400x1600	1400x1600	480x1600	480x1600
	Fused	1400x3200		480x3200	
	Authorized	40 subjects / 35 samples each		40 subjects / 12 samples each	
	Unauthorized	/		160 subjects / 45 samples each	

Table 4: Description of Experiment II databases.

	Features	Classifier	RR (%)	EER (%)	Th	AUC
Raw Faces & MFCC + VQ	Proposed Method 1: Concatenated (pn)	ANN	94.37	**9.02**	62.4	0.9989
		K-NN	85.41	10.84	20	0.8570
	Proposed Method 2: Merged (n)	ANN	100	**1.67**	54.1	0.9960
		K-NN	95.62	8.1	20	0.9213
	Proposed Method 3: Multiplied (n)	ANN	96.45	21.94	48.7	0.8597
		K-NN	86.45	**14.83**	40	0.8408
PCA for Faces & MFCC+VQ	Concatenated (n)	ANN	98.12	35.67	40.2	0.7056
		K-NN	91.25	**13.44**	54.6	0.8010
DCT for Faces & MFCC+VQ	Concatenated (n)	ANN	99.37	23.14	46.3	0.8480
		K-NN	96.26	**11.07**	63.7	0.8700

(pn) Pre-normalized features.**(n)** Normalized features.

Table 5: Results with different schemes of fusion and classification.

5.3 Experiment III: (FEI + voice)

We have downsized the FEI images to 40x40 gray pixels, in order to minimize the amount of calculations as compared to the original colored 640x480x3. The number of subjects is 100. Since the FEI images are varying by degrees from left to right, we decided to take random dispositions for each subject. 10 random positions were taken for each person. As for voice, we took 10 samples of speech for each subject and assigned them to the faces of the corresponding person. Totally, the database contains 1000 samples for training. We took the remaining 4 images for testing and assigned4voice samples to them, this makes 400 samples for testing with authorized subjects (Table 6). From the obtained results in this experiment as shown in (Table 7), all fusion methods were better in recognition on ANN than K-NN in average (94.05 vs 79.65%) with a high significance of ($p = 0.007 < 0.01$), because the network system has fit well the data and generalized to the testing images despite the changes in degrees of rotation from left to right. In terms of EER, we ignore *method 1* from discussion because it reports high errors, proposed *method 3* gives same errors on both classifiers, same as *method 5*. Proposed *method 2* on ANN gave the lowest

Databases	Details	Training		Testing	
		Face	Voice	Face	Voice
FEI + Voice	**Samples**	1000x1600	1000x1600	400x1600	400x1600
	Fused	1000x3200		400x3200	
	Authorized	100 subjects / 10samples each		100 subjects / 4 samples each	
	Unauthorized	/		100 subjects / 10 samples each	

Table 6: Description of Experiment III databases.

	Features	Classifier	RR (%)	EER (%)	Th (%)	AUC
Raw Faces & MFCC + VQ	Proposed Method 1: Concatenated (pn)	ANN	86.5	23.15	3.9	0.8489
		K-NN	74.5	15.75	33.4	0.7060
	Proposed Method 2: Merged (n)	ANN	97	**9.25**	53.4	0.9435
		K-NN	81.25	13	33.4	0.7771
	Proposed Method 3: Multiplied (n)	ANN	90.5	19.45	33.1	0.8620
		K-NN	72.25	19.55	20	0.7130
PCA for Faces & MFCC+VQ	Concatenated (n)	ANN	97.75	15	24.2	0.9326
		K-NN	79	**11.25**	14.3	0.7991
DCT for Faces & MFCC+VQ	Concatenated (n)	ANN	98.5	13.75	39.3	0.9395
		K-NN	91.25	13.2	42.9	0.8173

(pn)Pre-normalized features.(n) Normalized features.

Table 7: Results with different schemes of fusion and classification.

	Method 1	Method 2	Method 3	Method 4	Method 5
ORL& Voice	0.2325	0.1442	0.1763	0.044	0.0484
ORL with effects &Voice	0.1419	0.0747	0.0189	-0.095	-0.022
FEI& Voice	0.1429	0.1725	0.149	0.1335	0.1222

Table 8: AUC differences for Experiments I, II, III.

EER (9.25%) against the best EER of K-NN on *method 4* (11.25%), the difference is insignificant however ANN was better.

This may be related to the way ANN learns from features containing rotation contrary to K-NN which is a bare distance computation within a predefined radius. The area under the ROC Curve is a good measure of the system performance, basically, the greater is the area the greater is the ratio TPR/FPR meaning a capability to get more correct classification for less incorrect ones (1 is the maximum value). The Table 8 shows the differences between AUCs of ANN and KNN (subtracting AUC of KNN from the AUC of ANN) for each fusion method. The results have been obtained from the three experiments for the three virtual databases. It is noticeable that most differences are positive. If the used virtual bases are considered separately, ANN is better than KNN in at least 2 out of 5 results. Considering an analysis based on each fusion method, ANN gave better results than KNN in at least 2 out of 3 results. Finally, taking all methods and bases into account, ANN out performed KNN (13 out of 15). These observations lead to the deduction that ANN has an undeniable (outstanding) potential to perform better than KNN for all experimented fusion methods.

5.4 Experiment IV: (ORL + Voice) on PCA

In this Experiment, the idea is to train the database without external effects and test it with effects. In order to avoid the curse of dimensionality and have some flexibility in the training, as well as avoiding the system unbalance found in Experiment II for Method 4 and 5, we used PCA for the whole database Raw Faces & Voice with features normalized and merged because it was found to be the best system in the previous Experiments I & II& III. The major aim of this experiment is to evaluate the response to noise and external effects. The results of this experiment are tabulated in Table 9. A comparison of the recognition rates and EERs with and without effects between ANN and K-NN is tabulated in Table 10. In an intra-classifiers comparison of recognition rates, it is remarkable that external effects and noise have affected ANN with a high significance (a drop of 10%), but still behaved better than K-NN (a drop of 20%).

In inter-classifiers comparison, ANN outperformed K-NN with and without effects significantly as well. As for EERs, the error rates have increased significantly in both classifiers when noise was involved, (2.5 vs 20.48% ANN and 3.68 vs 14.3% K-NN). Even though there is an insignificant difference in averages between ANN and K-NN (20.48 vs 14.3 %), K-NN still reached a low error rate of (10.12%) while ANN kept a high EER (19.37%). For neural networks, this is an under fitting problem where the network is highly biased and generalizes too much to the point of reaching a high uncertainty whether to accept authentic subjects or reject imposters. This problem can be approached by tuning the network with other parameters as will follow in the next section proceeding.

	RR(%)		EER (%)		Eigenvectors
	No effects	Effects	No Effects	Effects	
ANN	99.16	87.9	2.5	**22.7**	80 eig
K-NN	95	76.45	3.01	**20.69**	
ANN	99.12	89.58	2.5	**19.37**	200 eig
K-NN	95.38	74.58	4.36	**12.09**	
ANN	99.16	89.97	2.5	**19.37**	280 eig
K-NN	96.66	76.04	3.67	**10.12**	

Table 9: Comparison between ANN and K-NN tested with and without effects.

		No effects	Effects	Significance
RR (%)	**ANN**	**99.14**	**89.15**	$p = 9.52.10^{-5} < 0.001$
	K-NN	95.68	75.69	$p = 1.23.10^{-5} < 0.001$
	Significance	$p = 0.002 < 0.01$	$p = 9.37.10^{-5} < 0.001$	/
EER (%)	**ANN**	**2.5**	20.48	$p = 8.5.10^{-5} < 0.001$
	K-NN	3.68	**14.3**	$p = 0.03 < 0.05$
	Significance	$p = 0.03 < 0.05$	$p = 0.14 > 0.05$ (NS)	/

Table 10:Comparison of average RR% intra and inter classifiers with and without effects.

Neural Networks	RR (%)		EER (%)		λ
	No effects	Effects	No effects	Effects	
ANN1	99.16	**91.25**	2.5	18.12	1
ANN2	99.16	**93.75**	2.5	14.79	0.1
ANN3	99.16	**94.58**	1.66	11.25	0.01
ANN4	99.16	**95.41**	1.66	**9.1**	0.001
ANN5	99.16	**95.62**	1.66	**7.7**	0.0001
ANN6	99.16	**96.45**	1.66	**8.95**	0.00001
ANN7	99.16	**96.87**	1.66	**5.83**	**0.000001**

Table 11: Results of tuning the neural network when tested with and without effects.

	Before Tuning (with effects)	After Tuning (with effects)	Significance
RR (%)	89.15	**94.84**	$p = 4.22.10^{-6}< 0.001$
EER (%)	20.48	**10.82**	$p = 6.61.10^{-6}< 0.001$

Table 12: Comparison of recognition rates and EERs pre and post tuning when testing with effects in average.

5.5 Experiment V: dependency of the neural network

In order to assess the dependency of the system either on face or voice or both of them, and to avoid the problem of over fitting as well as under fitting, we designed some more complex systems containing from 1 to 4 hidden layers and tested them with black faces (Faces features= 0), white faces (Faces features = 1) and without voice (Voice features =0). A description of the configurations is in test 1 (Tables 13) and test2 (Table 14). Since recognition rates were low in test 1, we tried to change the configurations in order to confirm the results. In the both tests, when faces were made black, recognition rates have dropped to averages of 4.69±2.55 % and 4.69±2%. This is because a great number of zeros in the test features will zero so many connection weights in the prediction model by multiplication which affects significantly the recognition.

For white faces, rates were much better than with black faces 8.35±3.66% (p<0.001) and 8.54±4.35% (p<0.001) for test 1 and 2 respectively, this confirms the first hypothesis. However without voice, accuracies were high in layer 1 (test1 gave 55.5±4.5 %, test2 gave 55.5±4.05 %). We understood that neural networks were relying on face features more than voice features. This can be related to the difference of ranges and variances

Test 1

<table>
<tr><td rowspan="2"><b>1 Layer</b></td><td><b>Input</b></td><td colspan="12"><b>Hidden Layers</b></td><td><b>Output</b></td></tr>
<tr><td>280</td><td colspan="12">500</td><td>40</td></tr>
<tr><td rowspan="2"><b>2 Layers</b></td><td><b>Input</b></td><td colspan="6"><b>Layer 1</b></td><td colspan="6"><b>Layer 2</b></td><td><b>Output</b></td></tr>
<tr><td>280</td><td colspan="6">250</td><td colspan="6">250</td><td>40</td></tr>
<tr><td rowspan="2"><b>3 Layers</b></td><td><b>Input</b></td><td colspan="4"><b>Layer 1</b></td><td colspan="4"><b>Layer 2</b></td><td colspan="4"><b>Layer 3</b></td><td><b>Output</b></td></tr>
<tr><td>280</td><td colspan="4">200</td><td colspan="4">150</td><td colspan="4">150</td><td>40</td></tr>
<tr><td rowspan="2"><b>4 Layers</b></td><td><b>Input</b></td><td colspan="3"><b>Layer 1</b></td><td colspan="3"><b>Layer 2</b></td><td colspan="3"><b>Layer 3</b></td><td colspan="3"><b>Layer 4</b></td><td><b>Output</b></td></tr>
<tr><td>280</td><td colspan="3">200</td><td colspan="3">100</td><td colspan="3">100</td><td colspan="3">100</td><td>40</td></tr>
</table>

Table 13:Characteristics of 4 complex configurations of neural networks in terms units.

Test 2

Since recognition rates were low in test 1, we tried to change the configurations in order to confirm the results.

<table>
<tr><td rowspan="2"><b>1 Layer</b></td><td><b>Input</b></td><td colspan="12"><b>Hidden Layers</b></td><td><b>Output</b></td></tr>
<tr><td>280</td><td colspan="12">500</td><td>40</td></tr>
<tr><td rowspan="2"><b>2 Layers</b></td><td><b>Input</b></td><td colspan="6"><b>Layer 1</b></td><td colspan="6"><b>Layer 2</b></td><td><b>Output</b></td></tr>
<tr><td>280</td><td colspan="6">500</td><td colspan="6">300</td><td>40</td></tr>
<tr><td rowspan="2"><b>3 Layers</b></td><td><b>Input</b></td><td colspan="4"><b>Layer 1</b></td><td colspan="4"><b>Layer 2</b></td><td colspan="4"><b>Layer 3</b></td><td><b>Output</b></td></tr>
<tr><td>280</td><td colspan="4">500</td><td colspan="4">300</td><td colspan="4">100</td><td>40</td></tr>
<tr><td rowspan="2"><b>4 Layers</b></td><td><b>Input</b></td><td colspan="3"><b>Layer 1</b></td><td colspan="3"><b>Layer 2</b></td><td colspan="3"><b>Layer 3</b></td><td colspan="3"><b>Layer 4</b></td><td><b>Output</b></td></tr>
<tr><td>280</td><td colspan="3">500</td><td colspan="3">400</td><td colspan="3">300</td><td colspan="3">200</td><td>40</td></tr>
</table>

Table 14:Characteristics of 4 complex configurations of neural networks in terms of units.

between faces and voices in our database taking into consideration that our normalization rule was not linear. Unfortunately, a homogeneity test was not performed to assess our databases. In the other hand, as the system started containing more hidden layers, the accuracies dropped to the level of faces, this means that the system started leaning from voices same as faces approximately, however the recognition was still bad which is not a good point.

5.6 Discussions

- In Experiment I, we used ORL & Voice fused using five different schemes, we trained and tested without external effects. Proposed *method 2* behaved very well on ANN and performed better than others. Proposed *method 3* has given a good recognition rate but is was the faultiest method.
- In Experiment II, we repeated Experiment I introducing external effects in training and testing databases. *Proposed method 2* implemented on ANN gave again the best RR and EER. *Methods (3,4,5)* were unacceptable with ANN contrary to their performance on K-NN.
- In Experiment III, we tested the capability of neural networks to generalize to unseen modals containing degrees of rotations for faces fused with voice. *Proposed method 2* reached the best results in terms of recognition rates and equal error rates. In contrast, *proposed method 3* was totally unacceptable. The AUC analysis was run on both classifiers performing on five methods of the study, with all the previous experiments, and has shown that from this criterion point of view, neural networks were much better than K-NN.
- In Experiment IV, we got back to ORL & voice and applied PCA on the whole database with normalized and merged features. We trained without external effects and tested with noise and effects. This has been done to assess the response to noise when not trained with. In terms of recognition rate, ANN performed well, in contrast with EER where it failed to give a low error. A tuning protocol was set up and applied in order to adapt the system to the type of data and solve the problem encountered consisting of under fitting. This was done mainly by varying the regularization parameters of the networks. The procedure of tuning gave good and promising results and confirmed the flexibility of neural networks.
- In Experiment V, we have done a dependency test on different configurations with a variety of regularization parameters, we found the system to be depending on face features over voice features. We could lower this dependency by designing more complex configurations, however, the recognition rates kept very bad telling that the system could not perform well in absence of one of the modalities.

6 Conclusion

In this work, we have introduced the concept of data fusion and explained why Multibiometric systems perform better than Unimodal systems. Our experimental part contained four experiments mainly done on two virtual databases, ORL & Voice, and FEI & Voice. Throughout Experiments I and II, *proposed method 2* gave the best recognition rates (99.16 and 100 %) and realized the least faulty systems (2.5 and 1.67%). We understand ultimately that ANN trained with merged and normalized data features from different modalities can be very effective. In experiment III where the database was much larger than the first and second trial, recognition rates diminished slightly and the equal error rate has increased significantly (9.25 %) but it maintained its position yielding the best performance since all other schemes have deteriorated as well. Although the *proposed method 1* was relatively good in experiments I and II, it was remarkably defective on the FEI database with Voices (23.11%), we concluded that normalizing features could have a powerful impact on the behavior of the neural network especially when the feature ranges are not approximate. *Proposed method 3* led to the conclusion that multiplying non-homogenous features as face and voice could alter unexpectedly the distinctive characteristics of different classes thus result in a completely unreliable system in comparison to the *proposed method 2*.

It is to mention that the classical methods 4 and 5 involving PCA and DCT for faces and MFCC & VQ for voices were much more effective in K-NN than ANN, this says basically that when features fed to a neural network are dimensionally unbalanced, the performance of the system could drop badly. In contrast with K-NN which is a simple distance measure that would not be affected by this problem. In experiment IV, we showed how neural networks could be influenced by noise and external effects simulating real-life scenarios. This has been done by training without effects and testing with them. Even though the results between K-NN performing better than ANN against noise were insignificant, we decided to set up a diagnosis protocol aiming to approach this problem. This has been done by discovering whether the modal of the neural network was under fitting the data, just well-fitting the data or over fitting it. The problem in hand was under fitting, it was resolved by changing the configurations in a convenient manner (Tuning the network) citing the layers and the regularization parameters. Using this perspective could lead to very promising and adaptive performances.

Finally, it is to be emphasized that we were able to achieve two major purposes of this study, first, was validating an effective data fusion method at feature level (proposed method 2 merging and normalizing features with equal dimensions), and second, consists of taking a good grasp of the concept of neural networks to the point of controlling its behavior as wanted to achieve good and better results.

As for further works, we hope applying this study on a better database where voices are recorded in an anechoic chamber. Also to apply a homogeneity test on this database in order to have a good statistical understanding of the features being fed to the recognition systems in hand.

References

[1] Faundez-Zanuy M. *Data fusion in biometrics*. IEEE A&E Systems Magazine, 20(1):34-38. January, 2005. https://doi.org/10.1109%2Fmaes.2005.1396793

[2] Almahafzah H., Imran M., and Sheshadri H.S. *Multi-algorithm Feature Level Fusion Using Finger Knuckle Print Biometric.* In *FGIT-FGCN/DCA*, pp. 302-311, 2012. https://doi.org/10.1007%2F978-3-642-35594-3_42

[3] Ross A., Jain A.K. *Mutlimodal Biometrics: an Overview*. Proceeding of 12th IEEE European Signal Processing Conference, pp.1221-1224, Austria, September 2004. ISBN: 978-320-0001-65-7.

[4] Nada Alay and Heyam H. Al-Baity. *A multimodal biometric system for personal verification based on different level fusion of iris and face traits*. Bioscience Biotechnology Research Communications. publisher by Society for Science and Nature, 12(3):565-576. September 2019. https://doi.org/10.21786%2Fbbrc%2F12.3%2F3

[5] Oloyede, M. and Hancke, G. *Unimodal and Multimodal Biometric Sensing Systems: A Review*. IEEE Access; 4:7532--7555, 2016. https://doi.org/10.1109%2Faccess.2016.2614720

[6] Fernandes, S.L. & Josemin Bala, G. Analyzing State-of-the-Art Techniques for Fusion of Multimodal Biometrics. Proceedings of the Second International Conference on Computer and Communication Technologies, pp.473–478, 2015. http://dx.doi.org/10.1007/978-81-322-2526-3_49.

[7] Chetty G. and Wagner M. *Robust face-voice based speaker identity verification using multilevel fusion*. Image and Vision Computing. 26(9): 1249–1260, 2008. https://doi.org/10.1016%2Fj.imavis.2008.02.009

[8] Palanivel, S. & Yegnanarayana, B. *Multimodal person authentication using speech, face and visual speech*. Computer Vision and Image Understanding, 109 (1):44–55, 2008. http://dx.doi.org/10.1016/j.cviu.2006.11.013.

[9] Raghavendra, R., Ashok Rao, and G. Hemantha Kumar. *Multimodal Person Verification System Using Face and Speech*. Procedia Computer Science 2: 181–187, 2010. https://doi.org/10.1016%2Fj.procs.2010.11.023

[10] Elmir Y, Elberrichi Z., Adjoudj R. *Multimodal biometric using a hierarchical fusion of a person ' s face, voice, and online signature*. Journal of Information Processing Systems, 10(4): 555-567; 2014. https://doi.org/10.3745/jips.02.0007

[11] Soltane M. Figueiredo-Jain (FJ) Tune Algorithm for Gaussian Mixture Modal (GMM) Based Face and Signature Multi-Modal Biometric Verification Fusion Systems. Journal of Computational Intelligence and Electronic Systems. 4(1): 27-36, 2015. https://doi.org/10.1166%2Fjcies.2015.1110

[12] Kasban H. *A robust multimodal biometric authentication scheme with voice and face recognition*. Arab Journal of Nuclear Sciences and Applications 50(3):120–130, 2017. ISSN 1110-0451.

[13] Gad R, El-Fishawy N, El-Sayed A, Zorkany M. *Multi-biometric systems: a state of the art survey and research directions*. International Journal of Advanced Computer Science and Applications, 6(6):128-138, 2015. https://doi.org/10.14569%2Fijacsa.2015.060618

[14] Balaka Ramesh Naidu, P.V.G.D Prasad Reddy. Fusion of Face and Voice for a Multimodal Biometric Recognition System. International Journal of Engineering and Advanced Technology (IJEAT), 8(3), February 2019. ISSN: 2249–8958.

[15] Shang-Hung Lin."An introduction to face recognition technology". Informing Science: The International Journal of an Emerging Transdiscipline, (3):001-007, 2000. https://doi.org/10.28945%2F569

[16] Cherifi D., Radji N., Nait Ali A.*, "Effect of Noise, Blur And Motion On Global Appearance Face Recognition Based Methods Performance"*, International Journal of Computer and Applications,16(6):0975–8887, February 2011. https://doi.org/10.5120/2019-2723

[17] Draper B.A., Baek K., Bartlett M.S., and Beveridge J.R.."*Recognizing Faces with PCA and ICA"*. Computer Vision and Image Understanding, 91(1-2):115-137, 2003. https://doi.org/10.1016/s1077-3142(03)00077-8

[18] Turk M. and Pentland A., "*Face recognition using eigenfaces"*. Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 586-591, Maui, HI, USA, June, 1991. https://doi.org/10.1109/cvpr.1991.139758

[19] Turk M.and Pentland A. *Eigenfaces for recognition.* Journal of Cognitive Neuroscience,3(1): 71-86, 1991. https://doi.org/10.1162/jocn.1991.3.1.71

[20] Chen Y. and Zhao Y. *Face Recognition Using DCT and Hierarchical RBF Model*. Intelligent Data Engineering and Automated Learning (IDEAL), pp. 355–362, 2006. https://doi.org/10.1007%2F11875581_43

[21] Nagil J., Khaleel Ahmed S. and Nagi F. *Pose Invariant Face Recognition using Hybrid DWT-DCT Frequency Features with Support Vector Machines*. In Proceedings of the 4th International Conference on Information Technology and Multimedia at UNITEN (ICIMU), Malaysia, pp. 99-104, November 2008. https://www.researchgate.net/publication/228699251

[22] Hajiarbabi M., Askari J., Sadri S., and Saraee M. *Face Recognition using Discrete Cosine Transform plus Linear Discriminant analysis*. Proceedings of the World Congress on Engineering, I:652-655, London, U.K, July 2007.

http://www.iaeng.org/publication/WCE2007/WCE2007_pp652-655.pdf

[23] Reynolds A.D. *An Overview of automatic speaker recognition technology*. Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 4:4072–4075, Orlando, FL, USA, 2002. https://doi.org/10.1109%2Ficassp.2002.5745552

[24] Majeed, S. A., Husain, H., Samad, S. A., and Idbeaa, T. F. Mel Frequency Cepstral Coefficients (MFCC) Feature Extraction Enhancement in the Application of Speech Recognition: a Comparison Study. Journal of Theoretical and Applied Information Technology 79(1):38-56, September 2015. http://www.jatit.org/volumes/seventynine1.php

[25] Kaur M., Girdhar A., Kaur M. *Multimodal biometric system using speech and signature modalities.* International Journal of Computer Applications. 5(12):13-16, August 2010. https://doi.org/10.5120%2F962-1339

[26] Elmir Y., Elberrichi Z., Adjoudj R. *A hierarchical fusion strategy based multimodal biometric system*. The International Arab Conference on Information Technology (ACIT'2013). Khartoum, December 17-19, 2013. https://doi.org/10.13140/RG.2.1.4675.1842

[27] Illinois Image Formation and Processing (IIFP).*DSP mini-project: An automatic Speaker Recognition System.* http://minhdo.ece.illinois.edu/teaching/speaker_recognition

[28] Cherifi D., Hafnaoui I., Nait Ali A.,*"Multimodal Score-Level Fusion Using Hybrid GA-PSO for Multibiometric System"*, International journal of computing and informatics, 39(1): 209–216, 2015. http://www.informatica.si/index.php/informatica/article/download/837/622

[29] Liestol K., Anderson PK., Anderson U., *Survival analysis and neural nets*, Statistics in Medicine, 13(12):1189-1200, June 1994. https://doi.org/10.1002%2Fsim.4780131202

[30] Andrew Ng, *"Machine Learning Online Course".* University of Stanford, 2011. https://freevideolectures.com/course/2257/machine-learning.

[31] Williams D. and Hinton G. *Learning representations by back-propagating errors. Nature*, 323(6088): 533-538, October 1986. https://doi.org/10.1038%2F323533a0

[32] Hinton GE. *How neural networks learn from experience*. Scientific American 267(3):145–151, September 1992. https://doi.org/10.1038%2Fscientificamerican0992-144

[33] Ba Lathika, D Devaraj. *Artificial Neural Network Based Multimodal Biometrics Recognition System*. International Conference On Control, Instrumentation, Communication and Computational Technologies (ICCICCT). Kanyakumari, India 10-11 July 2014. https://ieeexplore.ieee.org/document/6993100

[34] Soleymani S. , Dabouei A., Kazemi H., Dawson J. and Nasrabadi N. M., *Multi-Level Feature Abstraction from Convolutional Neural Networks for Multimodal Biometric Identification*, 24th International Conference on Pattern Recognition (ICPR), pp. 3469-3476, Beijing, 2018. https://doi.org/10.1109/icpr.2018.8545061