A Hierarchical Fusion Strategy based Multimodal Biometric System

1Youssef Elmir, 2Zakaria Elberrichi and 2Réda Adjoudj
1Faculty of Sciences and Technology, University of Adrar, Algeria
2Faculty of Technology, Djillali Liabès University of Sidi Bel Abbès, Algeria

Abstract: Biometric performance improvement is a challenging task. In this paper, a hierarchical strategy fusion based on multimodal biometric identification systems is presented. This strategy relies on a combination of several biometric traits using a multi-level biometric fusion hierarchy. The multi-level biometric fusion includes a pre-classification fusion with optimal feature selection and a post-classification fusion based on the similarity of maximum of scores. The proposed solution enhances biometric recognition performances based on suitable feature selection and reduction such as principal component analysis (PCA), as much as not all of the feature vectors components support the performance improvement degree.

Keywords: multimodal biometric fusion, hierarchical fusion, PCA.

1. Introduction

With the increasing need for security in our life, biometrics has rapidly become a hot research topic because their potential application values in personal identification. Among these applications, face recognition, which is user-friendly and requires less user cooperation. However, although many methods (e.g., eigenface, manifold, Gabor, sparse representation) have been proposed, it is still difficult to achieve high accuracy due to serious appearance variations. In contrast, online signature recognition is thought to be one of the most accurate biometrics, but its performances dramatically decrease in non-cooperative situations. Furthermore, voice recognition is very useful in some situations where the voice is the only available data (e.g., phone communications), but its performances depend on the quality of the collected data. Recently, the use of multimodal biometrics [15] fusion techniques has improved the performance of the biometrics system significantly. Therefore, we thought that face; online signature and voice fusion could be a promising strategy in practical applications, where accuracy is imperative. So, inspired by this idea, we developed a face, online signature and voice fusion based biometric system for personal identification and authentication tasks.

In this paper, we propose a hierarchical face, online signature and voice fusion framework, as illustrated in Figure 1. In the training stage, features are extracted from face using Gabor filters [20, 19], from voice using Mel-Frequency Cepstral Coefficients (MFCC) and form online signature a set of samples; each sample corresponds to the point coordinates on the digitizing tablet along with the corresponding pressure Xt, Yt and Pt where t corresponds to time. In the testing stage, the probe face, online signature and voice are used to obtain a subset of reference samples in the databases of face, online signature and voice. Finally, the fusion of face, online signature and voice is performed on the candidate subset for personal identification.

The remainder of this paper is structured as follows. Section II presents the system architecture with the hierarchical strategy of fusion. Section III introduces the face, online signature and voice feature extraction. Section IV describes feature representations for face, online signature and voice and the fusion strategy. Section V shows the classification method used. Section VI presents experimental results and discussions. At last, Section VII concludes.

2. The system architecture with the hierarchical strategy

Any biometric authentication system performs the following basic tasks: [18, 13 and 5]
• data acquisition: for each of the biometrics, measurements are derived from the primary data;
• feature extraction: find a given number of characteristic features carrying information;
• feature normalization: this step is necessary for a fusion at feature level;
• feature selection: a further dimensionality reduction stage providing the most discriminatory information. Out of all possible features, find out a subset of features achieving the best generalization performance for classifier, when trained on this subset;
• data classification: the essential step of the biometric recognition; [5]
• biometric fusion: the typical task in a multimodal biometric system. The biometric data fusion could be performed at different levels (biometric sensors, features, matching score and final decision), but the most applied fusion schemes are the post-classification ones, meaning that the similarity
scores provided by different biometric classifiers are combined according to a mathematical rule in order to provide a global similarity score;

The proposed multimodal biometric system architecture is shown in Figure 1. This architecture is based on a hierarchical approach for biometric data fusion and classification. The system includes two main identification components:

- the face identification sub-system.
- the online signature identification sub-system.

3. Feature extraction

4.1. Gabor filters

A Gabor filter bank is used to construct the face vector code. The defaults parameters correspond to the most common parameters used in conjunction with face images of size $128 \times 128$ pixels. Optionally, the function returns a filter bank structure that contains the spatial and frequency representations of the constructed Gabor filter bank and some meta-data.

A facial image of face is filtered with a bank of Gabor filters constructed using the construct Gabor filters using PhD toolbox [20, 19]. All filters are applied to the input image, magnitude responses are computed. Each of the computed magnitude responses is down-sampled. And finally, the down-sampled magnitude responses are concatenated into a single feature vector. Note that these feature vectors are produced such as those produced in [20], [9], or [11].

4.2. Mel-Frequency Cepstrum Coefficients

The signal of a voice is first processed by software that converts the speech waveform to some type of parametric representation (at a considerably lower information rate) for further analysis and processing. The speech signal is a slowly timed varying signal (it is called quasi-stationary). When examined over a sufficiently short period of time (between 5 and 100 ms), its characteristics are fairly stationary. However, over long periods of time (on the order of 1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. Therefore, short-time spectral analysis is the most common way to characterize the speech signal. A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coding (LPC), Gaussian mixture models (GMM) [11], Mel-Frequency Cepstrum Coefficients (MFCC), and others.

MFCC’s are based on the known variation of the human ear’s critical bandwidths with frequency; filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the Mel-frequency scale, which is linear frequency spacing below 1000 Hz and a logarithmic
spacing above 1000 Hz. The process of computing MFCCs is described in more detail in [12, 3].

4.3. Online signature features

Two kinds of modalities are considered when dealing with the verification of signatures, the offline modality, in which scanned copies of the signatures are available for the comparison, and the online modality, in which the signatures are acquired using digital tablets. The online modality provides more information about the signatures (trajectory, speed, pressure...etc.) and achieves consequently better verification performance than the offline modality. This is why we are more interested in this paper in the online modality.

Online signatures contain a set of samples; each sample corresponds to the point coordinates on the digitizing tablet along with the corresponding pressure. When performing the comparison, seven differences at the histogram level are considered as a linear transformation

\[ d_i = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2} \]

Angles: The angle between the X axis and the line formed with the first signature point and the current point

\[ \alpha_i = \tan^{-1}\frac{y_i - y_{i-1}}{x_i - x_{i-1}} \]

Speeds: The difference between successive distances

\[ S_i = d_i - d_{i-1} \]

Angular speeds: The difference between successive angles

\[ \Delta \alpha_i = \alpha_i - \alpha_{i-1} \]

In sum, seven features are extracted from each signature. When performing the comparison, seven differences at the signal level and seven other differences at the histogram level are computed.

4. Data fusion and reduction

4.1. Fusion at score level

A system based on fusion at the score level is proposed. The scores used are obtained from the monomodal systems used in the first proposed system (see Figure 1).

4.3. Features normalization

After feature extraction, the obtained feature vectors may exhibit significant variations both in their range and distribution. In our experiments we used min-max strategy of normalization to combine face and voice feature vectors. The goal of feature normalization is to modify the location (mean) and scale (variance) of the features values in order to ensure that the contribution of each component to the final match score is comparable [6]. Adopting an appropriate normalization scheme also helps address the problem of outliers in feature values. The simple min-max techniques were tested in this work. Let \( x \) and \( x' \) denote a feature value before and after normalization, respectively. The min-max technique computes \( x' \) as,

\[ x' = \frac{x - \min x}{\max x - \min x} \] (1)

Where \( F_i \) is the function which generates \( x \). The min-max technique is effective when the minimum and the maximum values of the component feature values are known beforehand. In cases where such information is not available, an estimate of these parameters has to be obtained from the available sample training data. The estimate may be affected by the presence of outliers in the training data and this makes min-max normalization sensitive to outliers [14].

4.4. Feature selection and reduction

Principal component analysis (PCA) can be used to approximate the original data with lower dimensional feature vectors. The basic approach is to compute the Eigen vectors of the covariance matrix of the original data, and approximate it by a linear combination of the leading Eigen vectors [17]. By using PCA procedure, the test vector can be identified by first, projecting the image onto the Eigen vector space to obtain the corresponding set of weights, and then comparing with the set of weights of the vectors in the training set [2, 8]. The problem of low-dimensional feature representation can be stated as follows: Let \( \mathbf{X} = (x_1, x_2, ..., x_n) \) represents the \( n \times N \) data matrix, where each \( x_i \) is a vector of dimension \( n \), concatenated from a face and online signature feature vectors. Here \( n \) represents the total number of elements in the face and online signature feature vectors and \( N \) is the number of subjects’ references in the training set. The PCA can be considered as a linear transformation (2) from the original vector to a projection feature vector, i.e.

\[ \mathbf{Y} = \mathbf{W}^T \mathbf{X} \] (2)

Where \( \mathbf{Y} \) is the \( m \times N \) feature vector matrix, \( m \) is the dimension of the feature vector and transformation matrix \( \mathbf{W} \) is an \( n \times m \) transformation matrix whose columns are the Eigen vectors corresponding to the \( m \) largest eigen values computed according to the equation (3):

\[ \mathbf{Y} = \mathbf{W}^T \mathbf{X} \]
The Mahalanobis space is defined as:

\( S = \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T, \mu = \frac{1}{N} \sum_{i=1}^{n} x_i \) \hspace{1cm} (4)

After applying the linear transformation \( W^T \), the scatter of the transformed feature vectors \( \{y_1, y_2, ... y_N\} \) is \( W^T S W \). In PCA, the projection \( W_{opt} \) is chosen to maximize the determinant of the total scatter matrix of the projected samples, i.e.,

\[ W_{opt} = \text{arg} \max \text{det}(W^T S W) = [w_1 w_2 ... w_m] \] \hspace{1cm} (5)

Where \( \{w_i | i = 1, 2, ..., m\} \) is the set of \( n \)-dimensional Eigen vectors of \( S \) corresponding to the \( m \) largest Eigen values. In other words, the input vector in an \( n \)-dimensional space is reduced to a feature vector in an \( m \)-dimensional subspace.

5. Classification

To compute the similarity measure we adopt a bank of face, signature and voice vector codes. As in the references bank a gallery of vector codes of face images, signature and voice signals is available. Five eigen spaces are created. The similarity measure between the test vector codes and the codes in the gallery is defined as the Cosine Mahalanobis distance [1] between the projection of the test vector code and the projections of the gallery vector codes.

Let \( \Gamma_1, \Gamma_2, \Gamma_3, ..., \Gamma_N \) be the vector picked from the gallery. Let \( \mathbf{\Theta} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{t}_i \) be the average vector code. Let \( \Phi_i = \Gamma_i - \mathbf{\Theta} \) be the mean subtracted vector codes. Let the data matrix \( A \) is defined as:

\[ A = [\Phi_1 \Phi_2 ... \Phi_N]. \]

The eigen vectors of \( A^T A \) can be computed as \( A^T A \mathbf{v}_k = \lambda_k \mathbf{v}_k \). Pre-multiplying both sides by \( A \), \( A^2 \mathbf{v}_k = \lambda_k A \mathbf{v}_k \). Thus \( A \mathbf{v}_k \) are the Eigen vectors of \( A^2 \). If \( w_i \) is the projection of the mean subtracted vector code on the \( i^{th} \) eigen vector, then the projection coefficients of the vector code are:

\[ \mathbf{u} = [w_1, w_2, ..., w_N]. \]

We use the Cosine-Mahalanobis distance to measure the similarity between projection coefficients. The use of the Cosine Mahalanobis distance is motivated by the results in [1].

The eigen vectors span the vector space. The Eigen values correspond to the variance along each Eigen vectors. We need to understand the transformation between the vector space and the Mahalanobis space before computing the Cosine Mahalanobis distance. The Mahalanobis space has unit variance along each dimension. Let \( u \) and \( v \) be two vectors in the Eigen space. Let \( \Sigma_u = \Sigma_v \) be the variance along the \( i^{th} \) dimension. Let \( m \) and \( n \) be the corresponding vectors in the Mahalanobis space. The relationship between the vectors is defined as:

\[ \mathbf{m}_k = \frac{\mathbf{u}_k}{\Sigma_u}, \mathbf{n}_k = \frac{\mathbf{v}_k}{\Sigma_v} \] \hspace{1cm} (6)

Mahalanobis cosine is the cosine of the angle between the projections of the vectors on the Mahalanobis space. So, the Cosine Mahalanobis distance between \( u \) and \( v \) is computed in terms of \( m \) and \( n \).

\[ d_{\text{Cosine}(u,v)} = \cos \theta(u,v) = \frac{\mathbf{m}_k^T \mathbf{n}_k}{|m||n|} \] \hspace{1cm} (7)

6. Results and analysis

4.1. Data sets

To evaluate our proposed fusion method, The ORL face database, the QU-PRIP database and ELSDSR voice database are used. The first database contains ten different images, each of 40 distinct subjects. For some subjects, the images were taken at different times, with varying lighting, facial expressions (e.g., open or closed eyes, smiling or no smiling) and facial details (glasses or 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) [16]. The second database used for the signature verification system is available from Qatar University; it contains 138 subjects with three reference signatures each, and some of the subjects have as many as six references. The third database used for the voice based system is available from the Technical University of Denmark (DTU) [4]; it contains 23 subjects with nine reference voices each.

Considering the number of subjects in the ELSDSR-voice database, we have selected the first twenty three users from the ORL-face database including voice samples and the same number from QU-PRIP database. 23 different users from the ORL-face database have been used. From both subsets, and by taking advantage of the independence of face, signature and voice traits, 23 virtual subjects have been created from those who have face, signature and voice traits.

The following training and testing process for monomodal systems has been established:

For training purposes, each face has been modeled using three samples, and each signature and voice has been modeled using the same number of samples.

For the testing, for each client three more samples of each trait (face, signature and voice) were also selected for testing; the same 23 clients are used as impostors, except that each client claims an identity different from his own. Each client has been considered and, from each impostor, six samples have been selected.

Consequently, the sub-corpus for the experiments consists of 23 clients, and \( 22 \times 23 \times 3 = 1518 \) multimodal impostor attempts.

4.2. Results

Our experiments demonstrate that hierarchical fusion-based method improves the efficiency of the multimodal authentication compared with both score
and feature based fusion methods. For identification purposes.

Table 1 demonstrates that the best recognition rate is obtained by hierarchical and scores fusion strategies (92.75% and 94.20% respectively) but the one obtained by features fusion strategy is not so far. Additionally, for verification purposes, the lowest equal error rate is obtained by using the hierarchical fusion strategy (0.10%). Additionally, features and scores fusion strategy demonstrate respectively low equal error rates in comparison with the monomodal systems. The same observation is noted for the minimal half total error rate and for the verification rate when the false accept rate is equal to 1% or 0.1%, on the contrary, when it is equals to 0.01% the verification rate is the same for all the studied systems.

<table>
<thead>
<tr>
<th></th>
<th>RR(^2) at Rank one (in %)</th>
<th>EER(^2) (in %)</th>
<th>MHTER(^3) (in %)</th>
<th>VR(^4) at 1% FAR(^5) (in %)</th>
<th>VR at 0.1% FAR (in %)</th>
<th>VR at 0.01% FAR (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face</td>
<td>73.91</td>
<td>1.45</td>
<td>0.72</td>
<td>98.55</td>
<td>97.10</td>
<td>1.45</td>
</tr>
<tr>
<td>Signature</td>
<td>82.61</td>
<td>1.68</td>
<td>1.68</td>
<td>92.75</td>
<td>86.96</td>
<td>1.45</td>
</tr>
<tr>
<td>Voice</td>
<td>56.52</td>
<td>7.54</td>
<td>7.54</td>
<td>72.46</td>
<td>57.97</td>
<td>1.45</td>
</tr>
<tr>
<td>features fusion</td>
<td>76.81</td>
<td>1.28</td>
<td>0.56</td>
<td>98.55</td>
<td>98.55</td>
<td>1.45</td>
</tr>
<tr>
<td>scores fusion</td>
<td>94.20</td>
<td>0.10</td>
<td>0.10</td>
<td>100</td>
<td>98.55</td>
<td>1.45</td>
</tr>
<tr>
<td>hierarchical fusion</td>
<td>92.75</td>
<td>0.10</td>
<td>0.10</td>
<td>100</td>
<td>98.55</td>
<td>1.45</td>
</tr>
</tbody>
</table>

Figure 2. Cumulative match characteristic curves.

\(^1\) RR: Recognition Rate,
\(^2\) EER: Equal Error Rate,
\(^3\) MHTER: Minimal Half Total Error Rate,
\(^4\) VR: Verification Rate,
\(^5\) FAR: False Accept Rate.
Furthermore, as shown in Figure 2, the Cumulative Match Characteristic (CMC) curve is used to evaluate the identification performance as a comparison. Figure 2 shows the experimental results.

It is clear that the recognition rate obtained by fusion at score level is the better one at rank one (10^3), but starting from rank two, the hierarchical fusion strategy has increased its recognition rates comparing to other strategies.

To evaluate the performance of verification, Receiver Operating Characteristic (ROC) curve and Detection Error Trade-off (DET) curve are used and the results are shown in Figure 3. The ROC and the DET curves of face, signature, voice, features fusion, scores fusion and hierarchical fusion based verification as comparisons are given in these figures.

![ROC curve on the ORL face, Els city voice and QJ signature databases](image1)

![DET curve on the ORL face, Els city voice and QJ signature databases](image2)

Figure 3. Receiver operating characteristic and Detection error trade-off curves.

On the contrary of CMC curves, where hierarchical and scores fusion strategies obtained roughly the same identification performance, Figure 3 shows that the hierarchical fusion strategy produces less errors than scores fusion strategy, it is clear that the hierarchical fusion based multimodal systems can combine the advantages of the other fusion strategies to improve the method’s overall efficiency.

From these figures, the hierarchical fusion of face, signature and voice improves the performance, which indicates the high effectiveness of multimodal biometrics. This strategy significantly improves performances of both features fusion and scores fusion strategies.

7. Conclusion

In this paper we introduced a new vision for a highly accurate biometric system which combines face, signature and voice authentication systems in order to optimize the accuracy and performance. The proposed approach is based on hierarchical multilevel biometric fusion integration: feature-level fusion and matching score-level fusion. The hierarchical biometric fusion provides most of the overall performance improvement for the whole multimodal biometric system.

Finally, further researches should be performed especially on feature-level biometric fusion, and its impact on biometric recognition accuracy.

Acknowledgements

The online signature database and its features extraction block were offered by Qatar National Research Fund (QNRF) grant through the National Priority Research Program (NPRP) No. 09 - 864 - 1 – 128.

References


Youssef Elmir received his BSc in Computer Science from Djillali Liabes University-Sidi Bel Abbès, Algeria in 2005, he received his MSc in Computer Science from Mohamed Boudiaf University of Sciences and Technology-Oran, Algeria in 2007, currently he is a PhD candidate in Computer Science at Djillali Liabès University-Sidi Bel Abbès, Algeria. Following his MSc, he is now working as an assistant professor in Mathematics and Computer Science department at University of Adrar.

Zakaria Elberrichi received his Master degree in computer science from the California State University, and received his PhD in computer science from the university Djillali Liabes, Sidi-Belabbes, Algeria. He is the head of the Data Mining and Web Intelligent team of the EEDIS laboratory and a professor at the same University.

Read Adjoudj earned his Ph.D in computer science from Djillali Liabes University, Sidi Bel Abbes, Algeria, 2006, and earned his Master (Magister) from the same university in 2002, and his Bachelor degree (engineer) from the same university in 2000. His major area of research is biometrics, pattern recognition, image processing, and paradigms of artificial intelligence. He is a member of the Evolutionary Engineering and Distributed Information Systems Laboratory EEDIS, Computer Science Department, University of Sidi Bel-Abbes.