Hybrid Biometric Person Authentication Using Face and Voice Features

Norman Poh, Jerzy Korczak

LSIIT, ULP-CNRS, Bld S. Brant, 67400 Illkirch, France
{poh, jjk}@dpt-info.u-strasbg.fr

Abstract In this paper, a hybrid person authentication prototype that can integrate multiple biometric devices is presented. This prototype is based on several levels of abstractions: data representations, vectors and classifiers. Frontal face and text-dependent voice biometrics are chosen to authenticate a user. For each of the biometric feature, an extractor, a classifier and a simple negotiation scheme have been designed. An extractor is made up of a sequence of operators which themselves are made up of signal processing and image processing algorithms. The face information is extracted using moments and the short speech information is extracted using wavelets. The extracted information, called vectors, is classified using two separate multi-layer perceptrons. The results are combined using a simple logical negotiation scheme. The prototype has been tested and evaluated on real-life databases.

1 Introduction

There is an increasing interest in biometric techniques for person authentication. Among these techniques are face, facial thermogram, fingerprint, hand geometry, hand vein, iris, retinal pattern, signature and voiceprint. All these methods have different degrees of uniqueness, permanence, measurability, performance, user's acceptability and robustness against circumvention [4].

A multimodal biometric system can improve the incompleteness of any unimodal biometric system. Brunelli et al. have proposed two independent biometric schemes by combining evidence from speaker verification and face recognition [1]. Dieckmann et al. have proposed an abstract level fusion scheme called "2-from-3-approach" which integrates face, lip motion and voice based on the principle that a human uses multiple clues to identify a person [2]. Kittler et al. have demonstrated the efficiency of an integration strategy that fuses multiple snapshots of a single biometric property using a Bayesian framework [5]. Maes et al. have proposed to combine biometric data, e.g. voice-print, with non-biometric data, e.g., password [6]. Jain et al. have proposed a multimodal biometric system design which integrates face, fingerprint and speech to make a personal identification [4].

The goal of this prototype is to design a hybrid biometric system that is independent of any biometric device. We propose an abstraction scheme that can combine any

biometric feature and facilitates the integration of new biometric features. By abstraction, we group these techniques into 1D, 2D or 3D recognition problems. For example, voice-print and signature acoustic is considered as a 1D pattern recognition problem, "mug-shot" face, facial thermogram, fingerprint, hand geometry, hand vein, iris, retinal pattern can be considered as a 2D or image recognition problem and face can be considered as a 3D or object recognition problem. Having classified these problems, the objective is to define a set of basic operations that work on 1D, 2D and 3D problems. These operations constitute the building blocks of extractors that can be defined to conceive a set of independent extractors either statistically (compiled) or dynamically (linked). Each extractor produces its own type of vector. The produced vector represents the biometric feature that can discriminate one person from another. The vector will be classified by its proper classifier. To combine the different results of classifiers, various negotiation strategies can be used.

The second section of this paper discusses the details of biometric authentication techniques, namely the face and the voice extractors, and neural networks as classifiers with a logical negotiation scheme. The third section discusses the databases and experiments protocol and the obtained results.

2 Biometric Authentication Methods

2.1 Face Authentication

In face recognition, problems are caused by different head orientations. So, if only the information around the eyes is extracted, then head orientations will not contribute to the errors. Of course, in doing so, other face information will be lost. Nevertheless, as a start, we opt for this simpler approach [8].

Firstly, a face image is captured using a web camera. A face is then detected using template matching. The user, however, has to move into the template rather than the template moving to search the face location. Eyes are then automatically localized using a combination of histogram analysis, round mask convolution and a peak-searching algorithm.

Moments are used to extract the eye information because it is a simple yet powerful extractor. Normalized central moments are invariant to translation, rotation and scaling. A moment of order $p+q$ of an image f_{xy} of N by N pixels with respect to a center ($\bar{x}$, $\bar{y}$) is given in Eq. 1. (More details can be obtained from [3].)

$$M_{pq} = \sum_{x}^{N-1} \sum_{y}^{N-1} f_{xy} (x - \bar{x})^p (y - \bar{y})^q \quad \textbf{(1)}$$

Instead of working on the red green blue RGB color space, we worked on the (hue saturation intensity) HSI color space as well. For each eye, a pair of moments is extracted from the green, blue, hue, saturation and intensity color space. These parameters make a vector of 10 dimensions for each eye. The magnitude of each item

in the eye vector is compressed using the logarithmic function and then normalized into the range zero and one. Fig. 1 illustrates the idea.

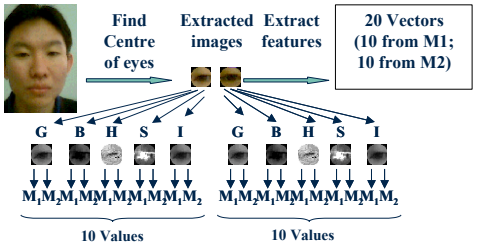

Fig. 1. Eye feature extraction using moments

2.2 Text Dependent Speaker Authentication

The front end of the speech module aims to extract the user dependent information. It includes three important steps: speech acquisition, detection and extraction. In general, the user's vocal password is sampled via a microphone at 8 kHz over a period of 3 seconds. In the second step, the presence of speech is then detected and then extracted using the Morlet wavelet [7].

By convoluting the wavelets with a speech signal, several scales of wavelet coefficients are obtained. The magnitude of wavelet coefficients is proportionate to the variation of signals. High magnitude of wavelet coefficients at a scale means a high variation change. Based on this information, it is possible to segment the speech signal and then used the segmented wavelet coefficients as a vector feature.

In our experiments, a wavelet transform on a speech signal of 3 seconds gives 8 analyzable scales. By using signal-to-noise analysis on the wavelet coefficients scale, we were able to determine that wavelets of scale-1, 2, 3 and 4 are more significant than other scales. Each of these scales is then truncated, normalized and then sampled before being merged to form a vector of 64 values (see Fig. 2). Through this sampling process, some important data could be lost. Such data reduction is necessary to make sure that the final vector is small enough to train the neural network.

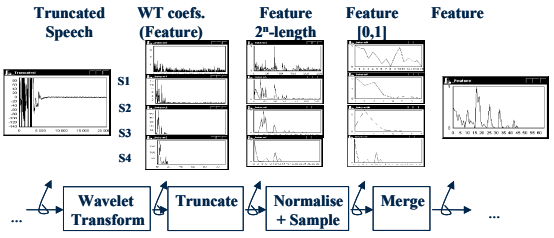

Fig. 2. Voice feature extraction using wavelet transform

2.3 Classifier and negotiation

A Multi-Layer Perceptron (MLP) is used for each type of face and voice vectors because it is robust against noise and efficient. It is considered as a universal classifier. An authorized person has his proper face and voice MLPs. Each of the MLPs is trained using the classical back-propagation algorithm. The output of each of the MLPs is true if the output neuron activation is more than an optimized threshold.

Each of the MLPs is trained several times and only the "best" observed MLP is retained. The auto-selection scheme, in general, is performed by evaluating the minimum cost error committed by the MLP evaluated on a validation set [8] (see experiment protocol below). As for the fusion of decision of two classifiers, instead of using complicated fusion scheme, we opt for a logical AND operation.

3 Test results

3.1 The database

Two databases have been created. The first database contains 30 persons. Each person has 10 face images and 10 sessions of speech recordings. The second database has 4 persons and each person has 100 face images and 100 sessions of speech recordings. The same amount of vectors is extracted using the raw biometric data. For the first and second databases, there are, therefore, 30 persons × 10 sessions × 2 biometric types = 600 vectors and 4×100×2=800 vectors respectively. The first database is aimed at simulating the real situation whereby an access point provides biometric-enabled check for 30 people. The second database is created so that more data is made available to train the MLP.

The database was acquired using a Creative WebCam and a standard PC microphone. The front view face image captured has a dimension of 150×225 pixel in RGB color and the voice is sampled at 8 kHz over 3 seconds. The biometric data was captured within one visit of the person to minimize the cost needed to capture large amount of data.

3.2 The experiments protocol

The database of features is divided into training set, validation set and test set randomly several times. This is in accordance to the cross-validation protocol. The training set is the data used directly to train the MLPs, i.e., changing the weight connections, while the validation set is used to calibrate the threshold and control the training, i.e., to determine the stopping condition and select the best trained MLP from a population of MLPs, and the test set is used *exclusive* to test the trained MLP.

The training:validation:test ratio are 3:1:1 and 5:2:3 for the first and second databases respectively.

Two error measures are used: *False Acceptance Rate* (FAR) and the *False Rejection Rate* (FRR). FAR and FRR are functions of a threshold that can control the trade-off between the two error rates [8].

The performance of the authentication system can be measured by plotting a Receiver Operating Characteristics curve (ROC), which is a plot of FRR versus FAR. The point on the ROC defined by FAR=FRR is the Equal Error Rate point (EER). The crossover accuracy is measured as *1/EER*, which can be interpreted as how many user the system can distinguish correctly before an error is committed. It should be noted that FAR, FRR and EER are data-dependent measurement and often does not reflect the real statistics.

3.3 Experiment Results

From the first database, 5 samplings of ROC are examined and their median is then plotted in Fig. 3(a). It can be observed that the voice MLP performs better than the face MLP because the ROC of the voice MLP lays nearer to the origin. However, their EERs are about the same, i.e., 0.10. By analyzing the density of FAR, out of 30 of the combined MLPs for 30 persons, 66.7% of them achieved FRR=0, 16.0% of them achieved FRR=0.25 (1 false rejection out of the combined 2 face vectors × 2 voice vectors) and 17.3% of them achieved FRR=0.50. As for the FAR, 98.7% of them achieved FAR=0 and 1.3% of them achieved FAR=0.009.

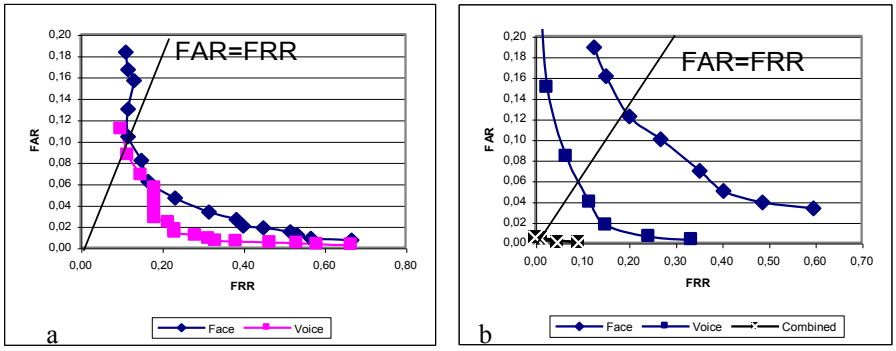

Fig. 3(a) Median of 5 ROC based on 30 persons (database I) **(b)** Median of 5 ROC based on 4 persons (database II)

In Fig. 3(b), the EER for the overall face MLP is about 0.15 while the EER for the overall voice MLP is around 0.07. The weak recognition rate of the voice MLP may be caused by the significant lost of information during the sampling of wavelet coefficients. Fig. 3(b) shows that there is a significant gain of performance when the two features are combined even though the EER is not measurable.

4 Conclusion

The prototype of hybrid person authentication system using a vector abstraction scheme and learning-based classifiers is a promising technique. From both the design and research points of view it is a valuable tool because it greatly facilitates the search of new extractors and classifiers. The extractors are made up of a sequence of operators which themselves are made up of signal processing and image processing algorithms. From a biometric data, an extractor extracts discriminative information and represents them in the form of a vector. A vector is then classified using its proper classifier that is made up of a set of learning-based matching algorithms.

We have chosen the frontal face and text-dependent voice biometrics as a prototype using the proposed scheme. The classifiers for each of the biometric data are Multi-Layer Perceptrons with a logical decision fusion scheme. Experiments using real biometric data show that a multi-modal approach is better than any single modalities. We found that the realized prototype depends on the performance of the extractors, i.e., how discriminative they are in extracting user-dependent information and the state of classifiers, i.e., how they are adequately trained and configured (the threshold value) before putting to use.

Our future directions will be to test the quality of vectors using vector quantization or k-means network that can measure inter-class and intra-class distance in order to search for the best discriminative extractor for a given application. By using learning-based classifiers, a large amount of biometric data is required not only to train the system but also to test the system independently.

References

1. Brunelli, R. and Falavigna, D.: "Personal identification using multiple cues," IEEE Trans. on *Pattern Analysis and Machine Intelligence*, Vol. 17, No. 10, pp. 955-966, 1995.
2. Dieckmann, U., Plankensteiner, P., and Wagner, T.: “SESAM: A biometric person identification system using sensor fusion,” In *Pattern Recognition Letters*, Vol. 18, No. 9, pp. 827-833, 1997.
3. Gonzalez, R., and Woods, R.: "Digital Image Processing", 2nd edition, Addison-Wesley, 1993.
4. Jain, A., Bolle, R., and Pankanti, S.: “BIOMETRICS: Personal identification in networked society,” 2nd Printing, Kluwer Academic Publishers, 1999.
5. Kittler, J., Li, Y., Matas, J. and Sanchez, M. U.: “Combining evidence in multi-modal personal identity recognition systems,” In *Proc. 1st Int. Conf. On Audio Video-Based Personal Authentication*, pp. 327-344, Crans-Montana, 1997.
6. Maes S. and Beigi, H.: "Open sesame! Speech, password or key to secure your door?", In *Proc. 3rd Asian Conference on Computer Vision*, pp. 531-541, Hong Kong, 1998.
7. Masters, T.: “Signal and image processing with neural networks: A C++ Sourcebook”, Academic Press, 1994.
8. Poh, N. and Korczak, J.: “Biometric Authentication System”, Res. Rep. LSIIT, ULP, 2001.