Scalability Analysis of Audio-Visual Person Identity Verification

Jacek Czyz[1], Samy Bengio[2], Christine Marcel[2], and Luc Vandendorpe[1]

[1] Communications Laboratory,
Université catholique de Louvain, B-1348 Belgium,
czyz@tele.ucl.ac.be,
[2] IDIAP, CH-1920 Martigny,
Switzerland
{Samy.Bengio,Christine.Marcel}@idiap.ch

Abstract. In this work, we present a multimodal identity verification system based on the fusion of the face image and the text independent speech data of a person. The system conciliates the monomodal face and speaker verification algorithms by fusing their respective scores. In order to assess the authentication system at different scales, the performance is evaluated at various sizes of the face and speech user template. The user template size is a key parameter when the storage space is limited like in a smart card. Our experimental results show that the multimodal fusion allows to reduce significantly the user template size while keeping a satisfactory level of performance. Experiments are performed on the newly recorded multimodal database BANCA.

1 Introduction

With the advent of digital communication and information society, reliable and user-friendly personal identity verification becomes more and more indispensable and critical. Biometrics, which measures a physiological or behavioural characteristic of a person, such as voice, face, fingerprints, iris, etc, provides an effective and inherently more reliable way to carry out personal identification [4]. Several factors influence the choice of a biometric trait for a particular application. Among them, distinctiveness and user friendliness are certainly the most important. For distinctiveness, the biometric trait should be distributed with a large variance inside the target population. At the same time, it should ideally remain constant for a given person, or vary with a small variance. As for user friendliness, the sensors that capture the biometric traits should interfere with the user as little as possible. Also the trait recordings should be done in an unconstrained and contactless manner. These two requirements are unavoidably contradictory. Therefore, it has been suggested to combine or *fuse* several easily accepted biometric traits, in order to achieve an acceptable level of distinctiveness and user friendliness at the same time. This technique is known as *multimodal biometrics*. The aptitude of multimodal biometrics for increasing correct verification rates over monomodal biometrics has been demonstrated in several previous studies, (see for example [3], [5], [8]).

A promising application consists in combining biometric efficiency and smart card (SC) security, by storing the user template on a SC [9]. SC's allow the template to be securely protected and avoid storing biometric data in a central server (central storage is not well accepted by users). However storage space of SC's and transmission speed between server and SC's limit the user template size. It is therefore important to evaluate performance as a function of the template size. In this work, we present an identity verification system based on the fusion of the face image and text independent speech data of a user. We analyse its scalability by evaluating the performance at different user template sizes. Although less accurate, text independent speaker verification allows more variability in the content uttered by the user. For this reason it is more user-friendly than text dependent speaker verification. The system presented is modular: each monomodal algorithm (face and speech) outputs a matching score reflecting its confidence in the presumed identity. The matching scores are then conciliated using a fusion algorithm which outputs the final authentication decision.

Our analysis of the experimental results on a realistic database shows that the fused system face-speech requires a much smaller number of parameters to represent a user than the best monomodal algorithm at the same performance level. Fusion can therefore help in reducing the storage space required for client data and thus improve the scalability of the verification system.

The paper is organised as follows. The monomodal algorithms and the fusion techniques employed are presented in section 2. In section 3, the scalability analysis is described. The database and the experimental protocol are presented in section 4. We discuss the results in section 5, and we draw conclusions in the last section.

2 Fusion of Face and Speaker Verification Algorithms

When the identity of a user has to be verified, speech and face are recorded and compared to previously created user template. A score reflecting the quality of the matching between the template and the data to verify is computed. The fusion of the two scores resulting from the speech and face algorithms leads to the final decision. Hereafter we describe briefly the speaker and face verification algorithms and the fusion techniques.

Face Verification Algorithm

The first step involves localisation and registration of the face part in the input image. In our implementation, we have skipped this step by manually locating the eye coordinates in the image. While often done in the literature, it biases optimistically the verification performance. After localisation, the face image is cropped and histogram equalisation is applied to reduce the effect of lighting variation. The Fisherface approach [1] is used to extract features from the gray level face image. This feature extraction technique is based on Principal Component Analysis (PCA) and on Linear Discriminant Analysis (LDA). LDA effectively projects the face vector into a subspace where within-class variations are minimised while between-class variations are maximised. Formally, given a set of face vectors x_i, each belonging to one of c classes $\{C_1, C_2, ..., C_c\}$, we compute the between-class scatter matrix S_b

$$S_b = \sum_{i=1}^{c} (\mu_i - \mu)(\mu_i - \mu)^T$$

and the within-class scatter matrix, S_w

$$S_w = \sum_{i=1}^{c} \sum_{x_k \in C_i} (x_k - \mu_i)(x_k - \mu_i)^T$$

where μ_i and μ are respectively the class conditional mean and the global mean. It is known that the projection matrix W which maximises the class separability criterion J

$$J = \frac{\|W^T S_b W\|}{\|W^T S_w W\|}$$

is a solution of the eigenproblem

$$S_w^{-1} S_b W = W\Lambda \tag{1}$$

where the diagonal matrix Λ contains the eigenvalues. In order to prevent S_w from being singular, an initial dimensionality reduction must be applied. This is achieved by taking the principal components of the face images. The face score s_f is computed by matching the newly acquired LDA face projection x to the user template x_t using normalised correlation $s_f = \frac{x^T x_t}{\|x\| \|x_t\|}$. Note that to compute the LDA basis, at least two images per person are required.

Speaker Verification Algorithm

The speaker verification algorithm used to compute the speech score is text independent and based on Gaussian Mixture Models (GMM) [7]. A parameterisation of the raw voice is performed, creating a vector of Linear Frequency Cepstral Coefficients (LFCC) for each section of 10ms of speech. On top of these coefficients, their first derivatives, as well as the log of the energy, are kept. Finally, a cepstral mean subtraction is performed in order to normalise the data. The user template is represented by a GMM that was adapted using a Maximum A Posteriori method from a general World Model GMM trained on a separate population of speakers. The speech score s_s is computed by estimating the log-likelihood ratio of a speech sequence of LFCC features $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ pronounced by the speaker i versus the world of all speakers Ω (world model)

$$s_s = \log p(\mathbf{X}|\, i) - \log p(\mathbf{X}|\Omega)$$

The densities $p(\mathbf{X}|\, i)$ and $p(\mathbf{X}|\Omega)$ given the ith speaker and world GMM models of N Gaussians can be computed as follows:

$$p(\mathbf{X}) = \prod_{t=1}^{T} p(\mathbf{x}_t) = \prod_{t=1}^{T} \sum_{n=1}^{N} w_n \cdot \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \tag{2}$$

where $\mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ is a Gaussian with mean $\boldsymbol{\mu}_n \in R^d$ where d is the number of features and with standard deviation $\boldsymbol{\Sigma}_n \in R^{d^2}$:

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) \tag{3}$$

Note that $\boldsymbol{\Sigma}$ is diagonal in the proposed implementation. The parameters that form the user template in this model are the means $\boldsymbol{\mu}_n$ of the N Gaussians, since the other parameters are fixed during the adaptation procedure and equal to those of the corresponding world model.

Fusion Algorithms

The fusion of the face and speech scores is performed using a second level classifier. That is, the two scores are considered as input features for a classifier which is trained on genuine and impostor score examples. In our experiments, we have opted for two fusion techniques. The first technique is based on a multi-layered perceptron (MLP) which can be viewed as a universal classifier. The MLP has two inputs where the two scores are fed in, and one output for the final fusion score $s = \sum_{i=1}^{m} tanh(w_{s,i} s_s + w_{f,i} s_f + b_i)$ where m is the number of hidden units and the parameters w and b are chosen to minimise the EER on the training set.

In the second fusion technique, a new score s is computed by averaging the weighted scores $s = w_s s_s + w_f s_f$. The fusion score s is then thresholded to obtain the final decision. The weights w_s and w_f and the threshold are found so as to minimise the EER on the training set.

3 Scalability Analysis

As stated in the introduction, we are interested in evaluating the performance of the monomodal and the multimodal algorithms at different user template sizes. While this template is relatively small for the face modality, it can be very large for the speech modality, mainly because of the large number of Gaussians that are necessary to represent faithfully the probability densities. For the face modality, the user template size is determined by the LDA subspace dimensionality. The LDA basis vectors, solution of (1), are ranked according to the magnitude of their corresponding eigenvalue. This magnitude is an indicator of the discriminatory power of the corresponding eigenvector. The performance of the face verification algorithm is then assessed at various numbers of LDA basis vectors, by gradually removing the less discriminative ones.

For the speech modality, the number of parameters of the user template depends on the number of Gaussians in the GMM and the feature vector size, i.e. the number of LFCC. The optimal number of Gaussians is normally determined by Maximum Likelihood on the world model. Here, we have trained the GMM on the world model with the number of Gaussians selected in the following set of values: 10, 25, 50, 100, 200 and 300. For each number of Gaussians, the performance of the speaker verification algorithm is assessed. We also studied the performance variation when k, the number of LFCC used to parameterise

Fig. 1. First frame of the 12 sessions of the BANCA database.

the raw voice varied between 4, 8, 12 and 16. Note that the derivatives of the LFCC and the signal energy are added to the feature vector. The feature vector size is therefore $2k + 1$.

The number of parameters of the multimodal template is simply the sum of the face and speech templates. We studied the performance variation of the multimodal system at different sizes of the speech and face templates.

4 Database and Experimental Protocol

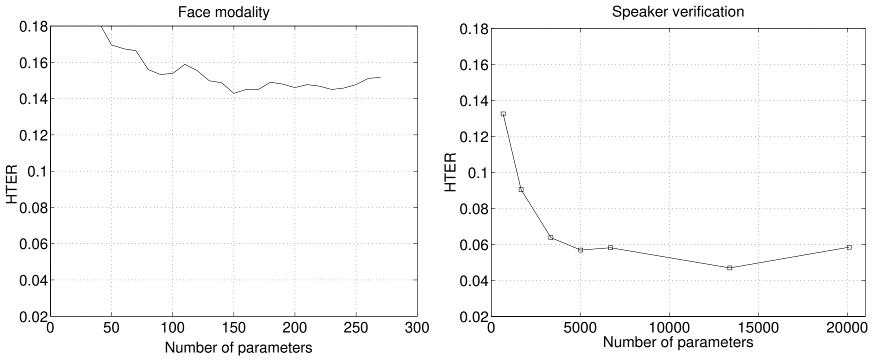
The experiments presented in the next section were performed on the English part of the BANCA database. This recently recorded database and the accompanying experimental protocol are described in detail in [2]. We give hereafter a short description. The data set contains voice and video recordings of 52 people in several environmental conditions. It is subdivided into two groups of 26 subjects (13 males and 13 females), denoted in the following by g1 and g2. Each subject recorded 12 sessions distributed over several months, each of these sessions containing 2 records: one true user access and one impostor attack. The impostor attacks are attempted only for subjects of the same sex, within the same group. The 12 sessions were separated into 3 different scenarios: controlled for sessions 1 to 4, degraded for sessions 5 to 8 and adverse for sessions 9 to 12. A low-cost camera has been used to record the sessions in the degraded scenario. For this scenario, the background noise for speech and video was unconstrained and the lighting uncontrolled, simulating a user authentificating himself in an office or at home using a home PC and a low cost web-cam. A more expensive camera was used for the controlled and adverse scenarios. The adverse scenario simulates a cash withdrawal machine, and was recorded outdoors. From one video session (about 30 seconds), five frames per person were randomly selected for face verification. At the same time, about 15 seconds of speech were recorded and used for speaker verification. During an impostor attack, the impostor utters the same text as the user that he is imposting. An additional set of 30 other subjects, 15 males and 15 females, recorded one session (audio and video) for each scenario. This set of data is used as world data. Figure 1 shows a subject of the BANCA database in the 12 sessions. The face images have already been located and registered. Notice how image quality varies accross the sessions.

In our testing protocol, session 1 only is used to enrol a new user, that is, to create its user template. In [2], this protocol is referred to as protocol P.

This demanding feature of the testing protocol was introduced because having to record several enrolment sessions may be tedious for the users in realistic applications. The remaining sessions are used to simulate genuine and impostors accesses. The testing protocol specifies a validation set, used to set the speech and face algorithm parameters as well as to train the fusion algorithm. A second set, the evaluation set, is used to assess the global system. Group g1 (group g2) is successively validation (evaluation) set and evaluation (validation) set. As in cross-validation, results from the two configurations are averaged.

Since only one session is available to create the user template, the LDA basis has to be computed with another face data set comprising several images per person. We chose the XM2VTS face database [6] for availability reasons. As this database contains 295 persons, the user template size is limited to 294 numbers.

As in any biometric system, two types of errors are possible: *false acceptance* when an impostor claim is accepted and *false rejection* when a genuine claim is rejected. These two errors depend on the biometric system threshold. To assess the performance, we adopted the following methodology. The threshold corresponding to the equal error (EER), that is, when the false acceptance rate (FAR) and the false rejection rate (FRR) are equal, is adjusted on the validation set. With this threshold, the system is tested on the evaluation set which leads to a false acceptance rate (FAR) and a false rejection rate (FRR) . From this two errors, we compute the half total error rate (HTER) which reflects the global performance of the verification algorithm HTER = (FAR+FRR)/2.

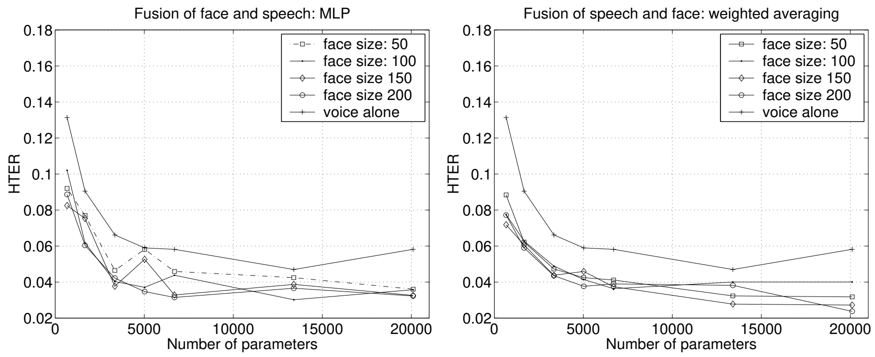


(a) (b)

Fig. 2. (a) HTER versus user template size for face modality. (b) HTER versus number of Gaussians needed to represent user template for speech modality.

5 Experimental Results and Discussion

According to the protocol described in the previous section, we studied the variation of the HTER as a function of the user template size. Figure 2(a) shows the variation of the HTER versus the user template size (expressed in number of parameters needed to store it) for the face verification algorithm. From the figure, the HTER decreases significantly with the first 100 basis vectors. The minimum HTER is reached at 150 vectors, and increases above 150 vectors. This means that the last features extracted (from 150 to 294) slightly degrades the classification and should not be included in the user templates. The minimum HTER obtained is 14.3%. This high value can be partially explained by the fact that only one enrolment session is available for creating the user template.



(a) (b)

Fig. 3. (a) HTER for MLP fusion vs. number of Gaussians and face template size. (b) HTER for weighted averaging fusion vs. number of Gaussians and face template size.

Scalability results for the speech modality are shown on figure 2(b). The curve on this figure corresponds to the variation of the HTER with the number of Gaussians and 16 LFCC coefficients. Since the other LFCC parameterisations offer higher error rates at equal number of parameters of the user template, we only present results with 16 LFCC coefficients. The best HTER for the speech modality is equal to 4.7% and is obtained with 200 Gaussians. It requires 13400 parameters to be stored.

The results of the fusion experiments are presented on figure 3(a) and 3(b) for MLP and weighted averaging fusion respectively. From these figures, it appears that the multimodal fusion always outperforms the best single modality (speech in our case). The lowest fusion HTER obtained is equal to 2.38%. The improvement thus reaches almost 50% in spite of the weakness of the face algorithm. Furthermore, the MLP fusion achieves an HTER of 3.77%, i.e. better than the speech modality alone, with only 50 Gaussians instead of 300. In this case, the number of parameters to be stored for the user template is 3500 (3350 for speech and 150 for face), which is almost 4 times less than what is needed for the system using the speech modality only. The speaker verification algorithm with 50 Gaussians for user template achieves an HTER of 6.62%. This result may be of practical interest when storage space is of concern, for example in a biometric system coupled with smart cards [9]. The limited storage and transmission speed of a smart card require user templates as small as possible. The fusion of modalities is therefore a way of improving the performance and reducing the number of parameters needed to be stored and transmitted, without decreasing performance.

6 Conclusions

A multimodal identity verification system using the speech and the face image of a user is presented. The experiments were conducted on a realistic database and according to a test protocol that allows only one enrolment session. The results show that the text independent speaker verification algorithm is robust and provides good results in spite of the uncontrolled nature of the data. In comparison, the face verification algorithm appears to be weak. A substantial improvement is gained when the outputs of the two monomodal algorithms are fused using simple techniques, the performance getting close to real world application requirements. An empirical analysis of the algorithm scalability with respect to the user template size is presented. It shows that fusion may help in reducing the number of parameters needed to be stored while keeping a satisfactory level of performance. Future work will be devoted to the design of a fully automatic audio-visual authentication system with automatic face location and registration.

Acknowlegments

This work was carried out within the framework of the European Project IST BANCA. We thank the CVSSP laboratory at University of Surrey (UK) for providing the eye coordinates for the BANCA database.

References

1. P. Belhumeur, J. Hespanha and D. Kriegman, “Face recognition: Eigenfaces vs. Fisherfaces: Recognition using class specific projection”, IEEE Trans. Pattern Analysis and Machine Intelligence, 19(7), 1997.
2. S. Bengio, F. Bimbot, J. Mariethoz, V. Popovici, F. Porée, E. Bailly-Balliere, G. Matas and B. Ruiz “Experimental protocol on the BANCA database” Technical Report IDIAP-RR 02-05, IDIAP, 2002.
3. B. Duc, E. S. Bigun, J. Bigun, G. Maitre, and S. Fischer. “Fusion of audio and video information for multi modal person authentication” Pattern Recognition Letters, 18:835-843, 1997.
4. A. Jain, R. Bolle and S. Pankanti “Biometrics: personal identification in a networked society”, Kluwer Academic Publishers, 1999.
5. J. Kittler, M. Hatef, R. P. W. Duin and J. Matas “On combining classifiers” IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 20, No. 3, pp. 226-239, 1998.
6. K. Messer, J. Matas, J. Kittler, J. Luettin and G. Maitre “XM2VTSDB: The extended M2VTS database” in Proc. of Int. Conf. on Audio and Video based Biometric Person Authentication, Washington, USA, 1999.
7. D. A. Reynolds and R. C. Rose “Robust Text-Independent Speaker identification using Gaussian mixture speaker models” in IEEE Trans. on Speech and Audio Processing, vol. 3, no. 1, pp. 72-83, Jan. 1995.
8. A. Ross, A. Jain and J.-Z. Qian “Information fusion in Biometrics” in Proc. of Int. Conf. on Audio and Video based Biometric Person Authentication, Halmstad, Sweden, 2001.
9. R. Sanchez-Reillo “Including Biometric Authentication in a smart card operating system”, Int. Conf. on Audio- and Video-based Person Authentication, Halmstad, Sweden, 2001.