Decision fusion for face authentication

J. Czyz[1], M. Sadeghi[2], J. Kittler[2] and L. Vandendorpe[1]

[1] Communications Laboratory

Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium

[2] Centre for Vision, Speech and Signal Processing

University of Surrey, Guildford, Surrey, GU2 5XH, UK

December 1, 2003

Abstract

In this paper we study two aspects of decision fusion for enhancing face authentication. First, sequential fusion of scores obtained on successive video frames of a user's face is used to reduce the error rate. Secondly, the opinions of several face authentication algorithms are combined so that the combined decision is more accurate than the best algorithm alone. The experiments performed on a realistic database demonstrate that the fully automatic multi-frame – multi-experts system proposed in this work allows a significant improvement over the static – single-expert system.

1 Introduction

In our electronically inter-connected society, reliable and user-friendly personal identity authentication is becoming more and more indispensable. Biometrics, which measures a physiological or behavioural characteristic of a person, such as voice, face, fingerprints, iris, etc., provides an effective and inherently reliable way to carry out personal identification [7]. Several factors influence the choice of a biometric trait or *modality* for a particular application. Among them, distinctiveness and user friendliness are certainly the most important. For distinctiveness, the biometric modality should be distributed with a large variance inside the target population. At the same time, it should ideally vary with a small variance for a given person. As for user friendliness, the sensors that capture the biometric modality should interfere with the user as little as possible. These two requirements are unavoidably contradictory, and make the authentication problem difficult.

The face modality is very important for real world applications because it is very well accepted by the users. In return, the acquired face images contain lots of variability. The pixel map of facial images varies drastically under variable illumination and 3D pose. Also the localisation and registration of the face sub-image is difficult when the background image is uncontrolled.

Robustness of face-based authentication can be improved by combining or fusing different sources of information related the identity to authentify. For example one could use several cameras oriented at different angles, or add other type of sensors like a microphone or a fingerprint sensor. In all cases strategies must be devised to combine the information coming from different sources.

In this paper we study two different aspects of decision fusion in the context of fully automatic face authentication. Firstly decision fusion is used combine the outputs of several face authentication algorithms. This type of fusion is referred to as intramodal fusion. Intramodal fusion has been recently studied for different biometric modalities [5, 9, 3]. Secondly we study sequential fusion, that is, the fusion of outputs of a single face authentication algorithm obtained on several video frames. During an access attempt the user is interacting with the authentication system over a certain period of time. Over this period many video frames are available for identity verification. For both fusion aspects, strategies for conciliating the different decisions are presented. Differences between intramodal fusion and sequential fusion are pointed out. The main contribution of the paper is a fusion architecture which takes into account the distinctive features of the intramodal and sequential fusion. Our experiments on a realistic face database show that the proposed architecture allows a significant improvement over a single frame – single expert approach.

The paper is organised as follows. In the next section we present biometric authentication. Also the two decision fusion aspects considered in this work are discussed. In Section 3, face authentication algorithms and the experimental setup is described. Experimental results are given and discussed afterwards. In the last section conclusions of our study are given.

2 Intramodal and Sequential Fusion of Face Authentication Experts

In this section, after presenting biometric authentication, we discuss intramodal and sequential fusion.

2.1 Biometric Authentication

Biometric identity authentication can be stated as follows. When performing verification, a biometric trait $\mathbf{x}$ of the person making the claim is recorded and compared to a reference trait, or template $\boldsymbol{\mu}_p$ that has been previously recorded. A score s reflecting the quality of the match between the template and the unknown biometric trait is compared to a threshold η to determine whether the claim is genuine (class ω_a) or false (class ω_b), i.e.

$$s(\mathbf{x}) \underset{\omega_b}{\overset{\omega_a}{\gtrless}} \eta \tag{1}$$

The level of performance of a biometric system is assessed through verification error rates. Two types of errors can be distinguished whether a genuine claim is rejected or an impostor claim is labelled as genuine. The former is referred to as False Rejection Rate (FRR) while the latter is referred to as False Acceptance Rate (FAR). Note that a data set disjoint from the training data is required to estimate the error rates and the threshold without bias.

2.2 Intramodal Fusion

In order to increase the verification performance, one may take advantage of multiple authentication algorithms, or experts, that provide their opinions on the same biometric data, and perform *intramodal fusion*. Various levels of combination are possible [10]: fusion at the feature level, fusion at the confidence level (also known as soft fusion) and fusion at the abstract level, where accept/reject decisions are combined (hard fusion). In this work we opt for confidence level fusion, that is, where the scores reported by the experts are combined. We believe that for authentication, confidence level fusion is a good compromise between dimensionality and information loss. Feature level fusion leads to high dimensional feature vectors for which the construction of a classifier may be problematic. At the decision level however, almost all information except for the class label has been thrown away: no confidence about the chosen class label can be deduced.

Given a measurement $\mathbf{x}$, each expert i outputs a score $s^{(i)}(\mathbf{x})$ based on the same measurement $\mathbf{x}$. These scores can be concatenated into a score vector $\mathbf{s}$ and a second-level classifier can be trained to learn a decision boundary in the score space. The problem of intramodal fusion amounts then to choosing a suitable classifier for this task. In [9], a non-parametric Parzen estimation technique is used to estimate the joint score density for combining several fingerprint matchers. A weighted averaging of scores, which corresponds to a linear discriminant function in the score space is proposed in [8, 10]. Here we perform the fusion using the weighted averaging technique and using

a Support Vector Classifier.

Weighted averaging : in this method the decision is based on a new score s_w which is obtained by linear combination of the experts score, i.e.

$$s_w = \mathbf{w}^T \mathbf{s},$$

where the weights $\mathbf{w}$ are obtained by minimising the Equal Error Rate (EER) on a training set.

Support Vector Classifier (SVC): in this method, an SVC with a linear kernel is trained to separate genuine from impostor score vectors. Linear SVC's determine a linear discriminant function in the score space, but the SVC learning relies on the principle of maximising the margin (sum of distances to the closest positive and negative training examples) rather than minimising the training error. The reader is referred to [2] for an introduction to SVC's.

2.3 Sequential Fusion

When multiple video frames of the same user's face are available, a score s_j is obtained from each frame j. The problem of sequential fusion is similar to the intramodal case: find a function f so that the decision rule

$$f(s_1, s_2, ..., s_N) \underset{\omega_b}{\overset{\omega_a}{\lessgtr}} \eta$$

leads to higher verification performance. Let us emphasise the difference with the intramodal case. In the sequential frame combination, all scores are emitted by the same expert, so that they can be seen as multiple random outcomes of the same score distribution (the score distribution depends on the expert). For this reason, the combination should (i) give the same importance to all scores, i.e. average the scores ; or (ii) have a mechanism of for selecting the "best" frame or best score.

In case (i) the scores are averaged so that each s_i has the same importance in the decision. The scores s_i are drawn from a random variable S with score probability distributions $p(s|\omega_b)$ and $p(s|\omega_a)$ in case of impostors or clients. It is well known that the sample average $\bar{S} = 1/N \sum_{j=1}^{N} S_j$ of N samples S_j (considered here as random variables) drawn from a given distribution has the same mean than the distribution. Also, if the samples are drawn independently, the variance of $\bar{S}$ is σ^2/N where σ^2 is the variance of the score distribution S. Therefore $p(\bar{s}|\omega_c)$ $(c \in \{a, b\})$ has the same mean than $p(s|\omega_c)$ but a variance divided by N. Because the error rate depends directly on the overlap between the impostor and genuine sample mean densities, if the decision is taken using $\bar{s}$ rather than s, the error rate decreases as N increases. Note that correlation between the S_j's contributes to increase the variance of $\bar{S}$. The effectiveness of the average rule for reducing

the error is then significantly diminished.

In case (ii), a simple solution for choosing the best frame consists of using a template matching-based method: select the frame that gives the best match with the template in the sense of a distance measure. In the case of a dissimilarity (similarity) score, this results in taking the minimum (maximum) score for making the decision, i.e.

$$\min(s_1, s_2, ..., s_N) \underset{\omega_b}{\overset{\omega_a}{\lessgtr}} \eta.$$

This may favour both genuine accesses and impostor accesses. The merit of this combination rule depends essentially on the score probability function as demonstrated in the simulations below.

Suppose that the impostor and genuine score distribution are Gaussian with equal variance σ^2 but different means $p(s|\omega_c) \sim \mathcal{N}(s; \mu_c, \sigma^2)$ with $c \in \{a, b\}$. We draw independently N samples s_j from the genuine or the impostor distribution and base our decision on the average of s_j or the minimum s_j. Figure 1(left) shows the classification error versus the number of samples N for Gaussian distributed scores for sample average based and minimum based decision. Note that although the original distributions lead to an error of 16%, the average rule reaches almost zero error rate with $N = 10$. However, in practice such an improvement is unlikely because the samples are not drawn independently. A saturation is likely to occur when N increases. Both integration methods improve the decision over the one sample score case, but clearly the average rule outperforms the minimum rule.

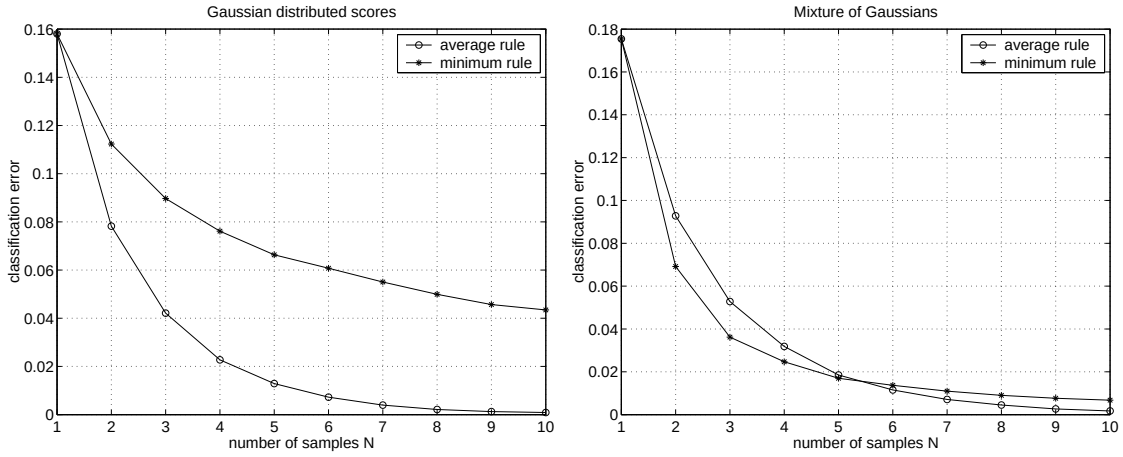

Figure 1: Classification error versus the number of samples N for Gaussian distributed scores (left) and Mixture of Gaussian score (right).

Gaussian hypothesis for scores may not be satisfied in practice. In particular the genuine score density has some properties that are common to all biometric system: they are usually asymmetric with a heavy tail or even bimodal [12]. The secondary mode is due to users who consistently

return large distance measures when samples are compared to stored templates (called "goats" in [4]). Another contributing factor to the heavy genuine density tail comes from failures during pre-processing. For example, faces poorely localised for face-based biometrics or failing silence removal for speech-based biometrics lead to large distance measures and false rejection. A more realistic choice is then to represent the genuine score density by a bimodal mixture of Gaussians

$$p(s|\omega_a) = \pi_1 G(s; \mu_{a1}, \sigma_{a1}^2) + \pi_2 G(s; \mu_{a2}, \sigma_{a2}^2),$$

where $\pi_1 + \pi_2 = 1$ and $\pi_1, \pi_2 \geq 0$ and $G(s, \mu, \sigma^2)$ is the Gaussian density function with parameters μ and σ. The impostor score density is kept Gaussian $p(s|\omega_b) = \mathcal{N}(s; \mu_b, \sigma_b^2)$. In this case, the error rates are obtained through simulations and results are presented in Figure 1(right). Genuine samples are randomly drawn from a mixture with parameters $\pi_1 = 0.8$, $\pi_2 = 0.2$, $\mu_{a1} = 0$, $\mu_{a2} = 3$ and $\sigma_{a1} = \sigma_{a2} = 1$. The impostor density has parameters $\mu_b = 3$ and $\sigma_b = 1$. Note that the "sample average" curve is quite similar to the pure Gaussian case. In contrast, the secondary mode changes drastically the "sample minimum" curve, which now outperforms the average rule for the first five frames that are combined.

From the simulations, it appears that the average rule is advantageous in the Gaussian case, while the minimum rule starts to outperform the average rule when the genuine density has a heavy tail.

2.4 Proposed Fusion Architecture

From the discussion above, we propose the following fusion architecture. For each expert i, the multiple scores $s_j^{(i)}$ $j = 1, 2, ..., N$ corresponding to multiple frames are first fused using either the average or the minimum rule (depending on the score distribution). The R resulting scores $s^{(i)}$ $i = 1, 2, ..., R$ are then fused using a second level classifier. The final decision is based on the output of the second level classifier. The experiments presented below show that this architecture allows a significant improvement over a single frame – single expert approach.

3 Experiments

3.1 Face Authentication experts

Once the face and eyes are located, the face is registered and histogram equalised. The normalised face image is then used to generate the accept/reject decision. In the results presented, we have used two different face verification algorithms, namely a Linear Discriminant Analysis (LDA) based

algorithm and an SVM based algorithm. Both methods are described in [11], we give hereafter a very short description.

The LDA approach is used to extract features from the gray level face image. LDA effectively projects the face vector into a subspace where within-class variations are minimised while between-class variations are maximised. The LDA score $s^{(1)}$ is computed by matching the newly acquired LDA face projection $\mathbf{y}$ to the user template $\mathbf{y}_t$ using normalised correlation

$$s^{(1)} = -\frac{\mathbf{y}^T \mathbf{y}_t}{\|\mathbf{y}\| \|\mathbf{y}_t\|}.$$

In the SVM-based method, to label the face vector $\mathbf{x}$ as genuine or impostor, the classifier evaluates the quantity

$$s^{(2)} = \sum_{i=1}^{l} y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b,$$

where $\mathbf{x}_i$ is the input vector of the ith training example, l is the number of training examples, the α_i and b are the parameters of the model, and $K(\mathbf{x}, \mathbf{x}_i)$ is the kernel function. In our case, the kernel is linear.

To locate automatically the face in the image, two different face localisation methods have been used. In the first method, the whole image is exhaustively scanned at different scales using a small window. The content of each window is classified by a SVM classifier into face or non-face classes. See [11] for more details. The expert using LDA face verification and the SVM-based face localisation is referred to as **LDA1**. The expert called **SVM** is using the SVM-based verification and localisation.

In the second face localisation method, Gabor filters are used to detect facial features such as corners of eyes, nostrils, etc. in the input image. Feature configurations that correspond possibly to a face sub-image are transformed into a normalised face space. In the face space, natural and geometric variability of faces is highly reduced because faces are registered. Full affine invariance in thus achieved before classification in face/non-face classes. More details can be found in [6]. The expert using LDA face verification and the face localisation method just described is referred to as **LDA2**.

3.2 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 [1]. 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 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 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.

In the experiments presented, two protocols are considered. The first protocol, referred to as protocol G in [1], uses the first session of the 3 scenarios to enrol a new user, that is, to create its user template. The second protocol (protocol P) uses session 1 only to enrol a new user. 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.

3.3 Experimental Results

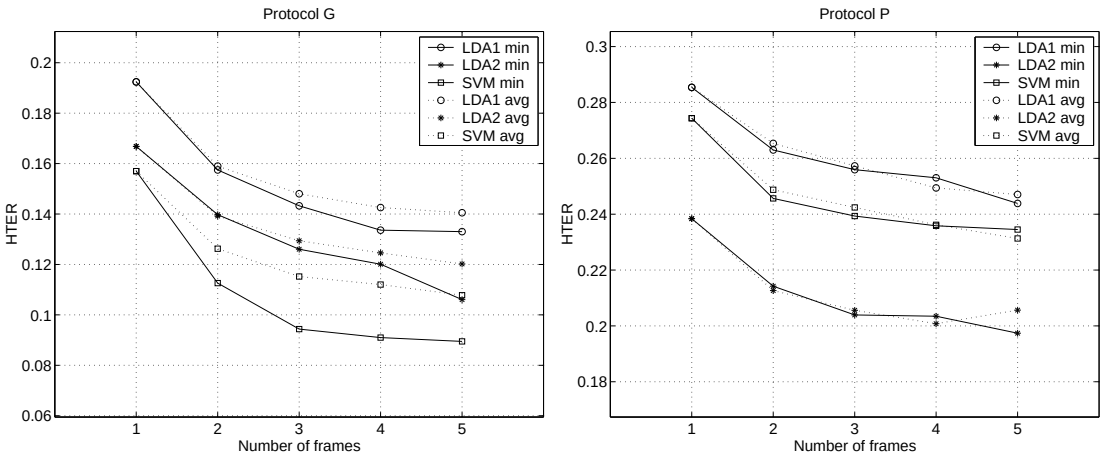

Figure 2: Sequential fusion results using minimum (min) and average (avg) rules obtained on the BANCA database for protocol G (left) and protocol P(right).

Figure 2 shows the average HTER obtained on the BANCA database using the minimum and the average rules above for protocol G (left) and protocol P (right). For the three experts considered, we evaluate the Half Total Error Rate HTER = (FAR + FRR)/2 as function of the number of frames fused. Starting with one frame, we add successively a frame to the set and base

the decision on the set of frames. It appears that, in the case of protocol G, the minimum rule gives the best improvement, up to several percents, over the single frame technique. The average rule gives a slightly weaker improvement. In the P protocol case (Figure 2(right)), the HTER is much higher than for the G protocol because only one session is available for training. Again, with 5 frames, the HTER is significantly smaller with respect to a single frame based decision. Note that this time the two rules perform approximatively the same. As expected, when the number of test frame increases, the HTER decreases, but the improvement seems to saturate quickly. It is likely that no further improvement could be obtained with a larger number of frames. Table 1 summarises the HTER obtained in this multi-frame – single expert case.

	Experts		
Protocol	LDA1	LDA2	SVM
G	13.30	10.68	8.95
P	24.39	19.74	23.45

Table 1: Multi-frame - single expert performance

Following the fusion architecture described in Section 2.4, the scores obtained after sequential fusion can be combined using the second level classifier. These intramodal fusion results are reported in Table 2. Note that the minimum rule has been used to fuse the sequential scores. From the table it appears that in all cases, the intramodal fusion further decreases the HTER over the multi-frame – single expert results. In partical the fusion of experts SVM and LDA2 leads to an error rate of 5.58% in protocol G and 17.65% in protocol P, using the SVC-based fusion. For comparison the best result in the single frame – single expert case is 15.70% in protocol G and 23.84% in protocol P. Note that the two fusion techniques allow approximatively the same improvement, with a slight advantage for SVC. Interestingly, the fusion of experts LDA1 and LDA2, improves the HTER although they differ only by the face localisation procedure.

Protocol	Fusion	Combined Experts		
	techn.	LDA1 & LDA2	LDA2 & SVM	LDA1 & SVM
G	w.avrg	9.53	6.27	8.44
G	SVC	9.34	5.58	7.32
P	w.avrg	19.60	18.38	20.83
P	SVC	18.43	17.65	20.06

Table 2: HTER obtained on the BANCA database using intramodal fusion for protocol P and G

4 Conclusion

We discussed how decision fusion can be used to improve the performance of automatic face authentication. Intramodal and sequential fusion are used at two different stages in the authentication process. For the sequential fusion, two combination rules are presented and it is shown that the minimum rule is advantageous over that average rule when the genuine score density has a heavy tail. Both rules allow a significant improvement over the single frame system. For the intramodal fusion, the very simple weighted averaging and the more complex Support Vector Classifier have shown to perform similarly on the BANCA face database. Experiments show that the error rates are substantially improved thanks to the multi-frame – multi-expert architecture, which gives practical relevance to the proposed approach. Recently Zhou and Chellappa proposed to use recursive Bayesian filtering for face authentication or recognition in video [13]. A performance comparison between this approach and the sequential fusion presented in this paper could be an interesting future study.

References

[1] E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran. The BANCA database and evaluation protocol. In *4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA*. Springer-Verlag, 2003.

[2] C. J. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[3] J. Czyz, J. Kittler, and L. Vandendorpe. Combining face verification experts. In *Proc. of Int. Conf. on Pattern Recognition, Quebec, Canada*, August 2002.

[4] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. 'sheep, goats, lambs and wolves': a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In *Proceedings of the International Conference on Spoken Language Processing*, 1998.

[5] D. Genoud, G. Gravier, F. Bimbot, and G. Chollet. Combining methods to improve the phone based speaker verification decision. In *Proc. of Int. Conf. on Spoken Language processing*, 1996.

[6] M. Hamouz, J. Kittler, J. Kamarainen, and H. Kalviainen. Hypotheses-driven affine invariant localisation of faces in verification systems. In *Proceedings of the International Conference on Audio- and Video-based Biometric Person Authentication*, 2003.

[7] A. K. Jain, R. Bolle, and R. Pankanti. Introduction to biometrics. In A. K. Jain, R. Bolle, and R. Pankanti, editors, *Biometrics: personal identification in a networked society*, pages 1–43. Kluwer academic publisher, 1999.

[8] A. K. Jain, S. Prabhakar, and S. Chen. Combining multiple matchers for a high security fingerprint verification system. *Pattern Recognition Letters*, 20, 1999.

[9] S. Prabhakar and A. K. Jain. Decision-level fusion in fingerprint verification. *Pattern Recognition*, 35:861–874, 2002.

[10] A. Ross, A. K. Jain, and J.-Z. Qian. Information fusion in Biometrics. In *Proc. Int. Conf. on Audio- and Video-based Person Authentication*, pages 355–359, 2001.

[11] M. Sadeghi, J. Kittler, A. Kostin, and K. Messer. A comparative study of automatic face verification algorithms on the BANCA database. In *Proceedings of the International Conference on Audio- and Video-based Biometric Person Authentication*, 2003.

[12] J. Wayman. Technical testing and evaluation of biometric identification devices. In A. K. Jain, R. Bolle, and R. Pankanti, editors, *Biometrics: personal identification in a networked society*, pages 345–368. Kluwer academic publisher, 1999.

[13] S. Zhou and R. Chellappa. Probabilistic human recognition from video. In *Proc. of Int. Conf. on Automatic Face and Gesture Recognition*, 2002.