ELSEVIER

Signal Processing 69 (1998) 59–79

SIGNAL PROCESSING

Image-based multimodal face authentication

Stéphane Pigeon*, Luc Vandendorpe

UCL – Laboratoire de Télécommunications et Télédétection, Place du Levant, 2 – B-1348 Louvain-La-Neuve, Belgium

Received 25 August 1997; received in revised form 15 September 1997

Abstract

Multimodality is a key word in order to increase the efficiency and the robustness of person authentication algorithms. Most of the multimodal authentication schemes currently developed, tend to combine speech and image-based features together and benefit from the high performance offered by the speech modality. Depending on the application, speech data is not always available or cannot be used. This paper takes these cases into account and investigates the best performance achievable through a system based on facial images only, using information extracted from both profile and frontal views. Starting from two different profile-related modalities, one based on the profile shape, the other on the grey level distribution along this shape, we will see how to issue a profile-based expert whose performance is improved compared to each profile modality taken separately. A second expert will use invariant parts of the frontal view in order to issue a frontal-based authentication. Different fusion schemes will be studied and the best approach will be applied in order to efficiently combine our two experts. This will result in a robust image-based person authentication scheme that offers a success rate of 96.5% on the M2VTS database.

Zusammenfassung

Multimodalität ist ein Schlüsselwort bei der Steigerung von Effizienz und Robustheit von Personen–Authentifizierungs Algorithmen. Die meisten Programme zur multimodalen Authentifizierung, die zur Zeit entwickelt werden, kombinieren Sprach- sowie Bild-basierte Eigenschaften und profitieren von der hohen Qualität der Sprachmodalität. In Abhängigkeit von der jeweiligen Anwendung sind Sprachdaten nicht zugänglich oder können nicht genutzt werden. In dieser Arbeit werden derartige Fälle betrachtet. Untersucht wird die best mögliche Güte, die mit einem System erzielt werden kann, daß nur auf Gesichtsbildern basiert, wobei Informationen aus der Profil- und Seitenansicht benutzt werden. Beginnt man von zwei unterschiedlichen Profil-abhängigen Modalitäten, bei denen die eine auf den Profilkonturen, und die andere auf der Graustufenskalierung entlang der Kontur basiert, läßt sich zeigen, wie eine Profil-basierter Experteneinheit mit höherer Güte im Vergleich zu jeder separat betrachteten Profilmodalität gewonnen wird. Eine zweite Experteneinheit benutzt invariante Anteile der Profilansicht, um eine Profil-basierte Authentifizierung zu erreichen. Unterschiedliche Verknüpfungen der Programme werden in dieser Arbeit untersucht und die Beste angewendet, um die beiden Expertensysteme effizient zu kombinieren. Als Resultat erhält man ein robustes auf Bildern basierendes Authentifizierungsprogramm für Personen, welches eine Erfolgsquote von 96.5% erzielt.

* Corresponding author. Tel.: + 32 10 478066; fax: + 32 10 472089; e-mail: pigeon@tele.ucl.ac.be.

0165-1684/98/$ – see front matter
PII: S0165-1684(98)00087-5

Résumé

La combinaison de multiples modalités biométriques permet d'améliorer sensiblement les performances et la robustesse des algorithmes d'authentification d'identité. Les techniques d'authentification multi-modales actuellement proposées font généralement usage d'un signal de parole et d'images du visage de la personne que l'on désire authentifier. Ces techniques bénéficient alors des excellentes performances offertes par la modalité parole. Cependant, selon l'application considérée, les données relatives à la parole ne sont pas toujours disponibles ou sont parfois inutilisables. Le présent article prend ces cas en considération et analyse le niveau de performance qu'il est possible d'atteindre en faisant usage d'un système basé sur des images faciales uniquement, à savoir une image du visage vu de profil et une image de face. A partir de deux modalités travaillant sur le visage de profil, la première étant basée sur le contour du profil et la seconde sur la distribution des niveaux de gris le long de celui-ci, nous verrons comment construire un expert profil dont les performances seront accrues par rapport à chaque modalité prise séparément. Un deuxième expert utilisera les parties invariantes de la vue du visage de face afin de déboucher sur une méthode d'authentification basée sur la vue frontale. Différentes techniques de fusion seront alors étudiées et la meilleure approche sera retenue afin de combiner efficacement nos deux experts. Ceci débouchera sur une méthode robuste d'authentification d'identité basée sur des images du visage et offrant un taux de succès de 96.5% sur la base de données M2VTS.

Keywords: Multimodal person authentication; Profile view; Fusion

1. Introduction

With the fast emergence of multimedia networks and their new services, the communication concept drastically evolved these last years. As the security issue has often been poorly addressed, the nature of the applications available over the network is mainly limited to non-commercial applications. Today, one has to reconsider the issue of secured access to local or centralized services taking the advantage of the latest breakthrough offered by the ever-growing multimedia environment. The main objective is to extend the scope of applications of network-based services by adding novel functionalities, enabled by multimodal verification strategies based on speech and face images. The aim of this paper is to develop an efficient face authentication algorithm, allowing to remotely verify the identity of a user using images taken from a distant camera. Before further describing the system developed here, let us first present some latest person recognition systems and their respective performance.

Basically, these systems can be divided into two main groups.

- Systems that need a close or physical contact with the user like fingerprint analysis, hand shape recognition or iris recognition. These systems are often not well accepted by the users due to the close contact constraint but offer a high performance. They can hardly be used in a multimedia environment where the only sensors are a low-cost camera and/or a microphone.
- Distant systems, dealing with an image captured from a distant camera or sound from a microphone. These systems are cheaper to implement (e.g., they can be software-based on a personal/multimedia computer), generally better accepted by the users for their convenience, but cannot compete with the performance offered by the first category. Therefore, multimodality (i.e., a combination of authentication algorithms which make use of non redundant feature sets) is a key point in order to increase the efficiency and robustness of these methods [1,3,4,11].

Having in mind the development of a system that runs on multimedia platforms, we will restrict this introduction to systems that belong to the second category only. These systems are mainly based on frontal and profile view images as well as speech recordings. Before giving an overview of the performance that can be expected from these modalities, we would like to stress how difficult it is to issue a comparison between the performance achieved by different algorithms, as referenced by their respective authors or in [6]. First, a wide range of databases is used. While some face databases well

represent the data available in real applications, others are more questionable and often make use of a test set that has been acquired during the same recording session as the one used to build the reference models. This results in staggering – but meaningless – performance measures. On the other hand, databases like the FERET database (static images) [15] or the M2VTS multimodal database (video sequences and sound) [16], offer a good material in order to test the expected performance of authentication algorithms in real life scenarios.

Apart from the database issue, another source of disparity between the performance measure of different algorithms consists in the way this performance has been evaluated. As a major example, the performance of a recognition algorithm can be expressed in terms of *identification* or *authentication* efficiency.

An identification system compares the biometric features of the person to identify with all the entries of a client database. The identity of the client who has the closest feature set is assigned to the candidate. Such a system needs an important computing time (all entries have to be checked) but performs well in terms of correct recognition ratio. Indeed, even if the features of one client may vary over time, the client can still be correctly identified as long as the difference between the current feature set and the reference set is smaller than the inter-user distances. Unfortunately, such a system does not deal with the problem of possible impostors. Any impostor will be able to enter the system under the identity of the client whose feature set is the closest to the impostor. Even if an acceptance threshold is defined (a threshold on the feature distance above which the identification is rejected), the risk that an impostor can gain access to the system increases with the size of the client database. The performance of identification algorithms described in the literature are often close to 100%, but unfortunately information about their behavior against impostor access is rarely commented. Brunelli and Poggio [5] report an identification rate of 90% using geometrical features extracted from the frontal view image of 47 people. This rate rises to 100% for a template matching running on the same database. The Moghaddam and Pentland approach [14] is based on eigenfaces and offers an identification accuracy of 99% using frontal views of 155 individuals. Yu et al. [19] report 100% correct identification over a database of 33 persons, by using fiducial marks extracted along the profile shape.

In an authentication scheme, the candidate has to claim his identity prior to accessing the system. This can be done by the introduction of a personal identification number or by a personal card that has to be read by the system. The features of the candidate are then compared with the entry associated with the claimed identity. Authentication succeeds if the distance between the candidate feature set and the claimed reference is below a given threshold which may depend on the claimed identity (individual thresholding). This system operates much faster compared to an identification scheme, since only one entry has to be checked. The performance of the system can be evaluated in terms of false rejection (FR) rate, i.e., the percentage of client accesses rejected by the system, and false acceptance (FA) rate, i.e., the percentage of impostor accesses accepted by the system. The success rate (SR) refers to $1 - \mathrm{FA} - \mathrm{FR}$. Goudail et al. [8] report an SR of 93.5% using local autocorrelation coefficients computed on the facial images of 116 persons. Konen and Schulze-Krüger [12] developed a system based on an extension of the Elastic Graph Matching that runs on frontal images and achieves an SR of 96% on a database of 87 persons. Beumier and Acheroy's profile-based authentication scheme offers an SR of 90% over 41 persons when the profile shape extends about a 500-line resolution [2].

In the framework of the M2VTS project, a European ACTS project dealing with multimodal person authentication, several algorithms were tested using a common multimodal database of 37 persons [16]. Using this database, an HMM based speech authentication method offered the highest success rate, as far as single modalities are concerned, with an SR of 97.5% [10]. By combining speech information with labial features found in the associated image sequence, the SR then increased up to 99.4% [10]. A Dynamic Grid Matching carried on frontal grey level images of the same database achieved an SR of 89% [7]. By combining these frontal and speech features together, the SR then rose to 99.5% [7].

These excellent authentication rates of over 99% are boosted in fact by the high performance offered by the speech modality. Depending on the application, speech data is not always available (person authentication using static mug shots) or cannot be used (person authentication in noisy environments). This paper takes these cases into account and investigates the performance of a system based on facial images only, using both profile and frontal information extracted from the profile and frontal views.

Starting from two different profile-related modalities, one based on the profile shape, the other on the grey level distribution along this shape, we will see how to issue a profile-based expert whose performance is improved compared to each profile modality taken separately. A second expert will use the invariant parts of frontal view (eyes and nose area) in order to issue a frontal-based authentication. Then different fusion schemes will be studied and the best approach will be applied in order to efficiently combine our two experts. This will result in a robust image based authentication scheme that offers an SR of 96.5% under the same conditions as the above-mentioned M2VTS results.

This paper is organized as follows. In Section 2, we present the first profile-related modality which works on the *profile shape* and uses a chamfer matching technique to map the candidate profile to the reference profile. Section 3 introduces the second profile-related modality which is based on a grey level correlation between the candidate and the reference *profile images*, computed inside an area taken along the profile shape. Section 4 deals with the last modality, which also issues a grey level correlation measure but makes use of a rectangular window located inside the *frontal image*, covering major invariant features like the eyes and nose. Section 5 presents the M2VTS database and the test protocol that has been used during our various experiments. The performance of each modality is given in Section 6, when both global and individual thresholding schemes are used. Section 7 discusses the increase of performance that can be expected by performing a fusion between different modalities. Both hard and soft fusion strategies are studied under the assumption of independent modalities. In order to fulfill this last condition, the two profile-related modalities have been merged into one profile *expert*. This profile-related expert may then be considered as independent from the frontal grey level modality (referenced as the frontal expert). In Section 8, the best strategy is applied to the fusion of our profile- and frontal-based experts into a unique image-based person authentication algorithm. Finally, Section 9 concludes this work.

2. Profile shape matching

The first modality consists of the authentication of the profile outline and is inspired from [17]. The algorithm is based on a chamfer matching that directly works on the profile contour encoded as x–y coordinates. In this way, the performance of this modality mostly depends on the ability of the profile shape to dissociate different faces and not on the choice of a particular set of profile features and/or the quality of their extraction. Before introducing the profile matching method, we will first explain how to extract profile shapes from the color images taken from the M2VTS database.

2.1. Profile segmentation

All profiles used in this work are taken from the M2VTS database profile views [16]. The database offers a nearly constant lighting over the different shots and a uniform grey background. As the background luminance is very close to the luminance of the skin, we have to use color information in order to extract the profile outline from the image. This extraction is automatically performed in two stages: first, the head is segmented from the background, then the profile is extracted from the head.

The head segmentation is performed by means of color clustering according to the method proposed in [13]. This segmentation can be summarized into three steps.

1. The image is low-pass filtered in order to smooth the different color components and to reduce the effect of noise inside the different color areas.
2. A 2-D color histogram is computed using normalized red $R/(R+G+B)$ and green

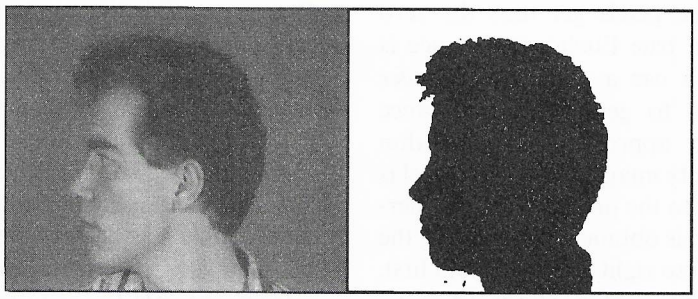

Fig. 1. Result of the background segmentation by means of color clustering.

$G/(R+G+B)$ components. As the uniform background represents the largest surface in the picture, its location inside the histogram is given by the highest peak. The background can be fully segmented by connecting the color components lying around the background peak.

3. By back projecting all pixels belonging to the background cluster into the original image, the background is extracted from the image (Fig. 1).

Once the head is segmented, the profile must be extracted from the head. In order to achieve good recognition results, we have to restrict this extraction to the invariant parts of the profile only, and,

- exclude the forehead when it can be affected by the hairstyle;
- exclude the area below the chin, as its contour highly depends on the tilt of the head which is likely to change from one shot to another;
- exclude the lower part of bearded faces.

For each user, it is also important to select the same part of the profile from one shot to the other in order to avoid the introduction of a bias in the residual chamfer distance once the best compensation parameters have been found (see Section 2.2). As mentioned above, this fixed part extracted from the profile outline, has to take into account the characteristics of the face, excluding features like hair, moustache or beard from the profile. Therefore, at user-definition time, the choice is given between different extraction modes: a first mode selects the full profile and assumes short hair and the absence facial hair (Fig. 2). A second mode only selects the lower part of the profile and has to be used when hair is suspected to cover the forehead. A third mode selects the upper area of the face for people wearing a moustache or a beard. The last mode is a combination of the two previous ones and selects the nose area only. All modes are normalized with respect to the nose height as illustrated in Fig. 2 (first mode). More information about these different extraction modes can be found in [17].

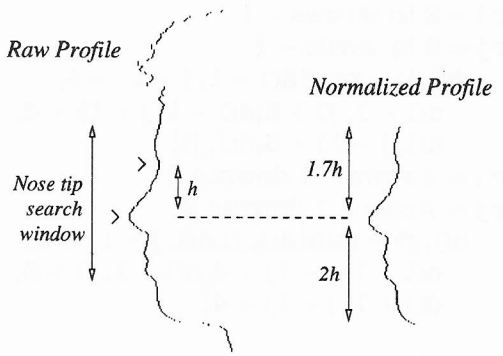

Fig. 2. Profile extraction (first mode).

2.2. Chamfer profile shape matching

The chamfer matching technique searches for the best match between two binary images. Geometric transformations are used to distort one image (here referred to as the *candidate image*) to another (the *reference image*) in order to minimize a given distance measure between them. These binary images are often derived from the image edges. Here, we make use of the shape of the normalized profile.

The first step of the algorithm is to generate a *distance map* from the reference profile. This distance map associates with each pixel of the reference profile picture, its distance from the closest

profile pixel (all profile pixels get thus the zero distance value). As the true Euclidian distance is costly to compute, we use a *sequential chamfer distance approximation* to generate the distance map [9]. The distance approximation algorithm starts from a zero/infinity image where each pixel is set to zero if it belongs to the profile, infinity otherwise. The distance map is obtained by applying the next formulas, from left to right/top to bottom first, and right to left/bottom to top afterwards (two passes are enough):

```
for i = 2 to #rows − 1
for j = 2 to #cols − 1
    d(i, j) = min{d(i − 1, j − 1) + 4,
      d(i − 1, j) + 3,d(i − 1, j + 1) + 4,
      d(i, j − 1) + 3,d(i, j)}
for i = #rows − 1 downto 2
for j = #cols − 1 downto 2
    d(i, j) = min{d(i, j),d(i, j + 1) + 3,
      d(i + 1, j − 1) + 4,d(i + 1, j) + 3,
      d(i + 1, j + 1) + 4}
```

By superposing the candidate image on this distance map and by summing up all distances found along the candidate profile, we get an estimate of the global distance that stands between them (Mean Squared criterion).

Actually, we cannot directly compare the reference and the candidate profiles together. The candidate profile has first to be compensated for the possible geometric transformations that can affect it from one shot to the other, i.e. translation along the x and y axes (t_x, t_y), rotation in the x/y plane (θ_{xy}) and scale factor (z). Given a set of values for these transformation parameters, we build a *compensated profile* from the candidate profile. This compensated profile is superposed on the reference distance map and a global distance is computed. The best match between the candidate and the reference profiles is obtained by finding the set of transformation parameters minimizing this global distance. It thus reverts to minimize a cost function (distance) which, in our case, depends on t_x, t_y, θ_{xy} and z. This minimization is done through a classic

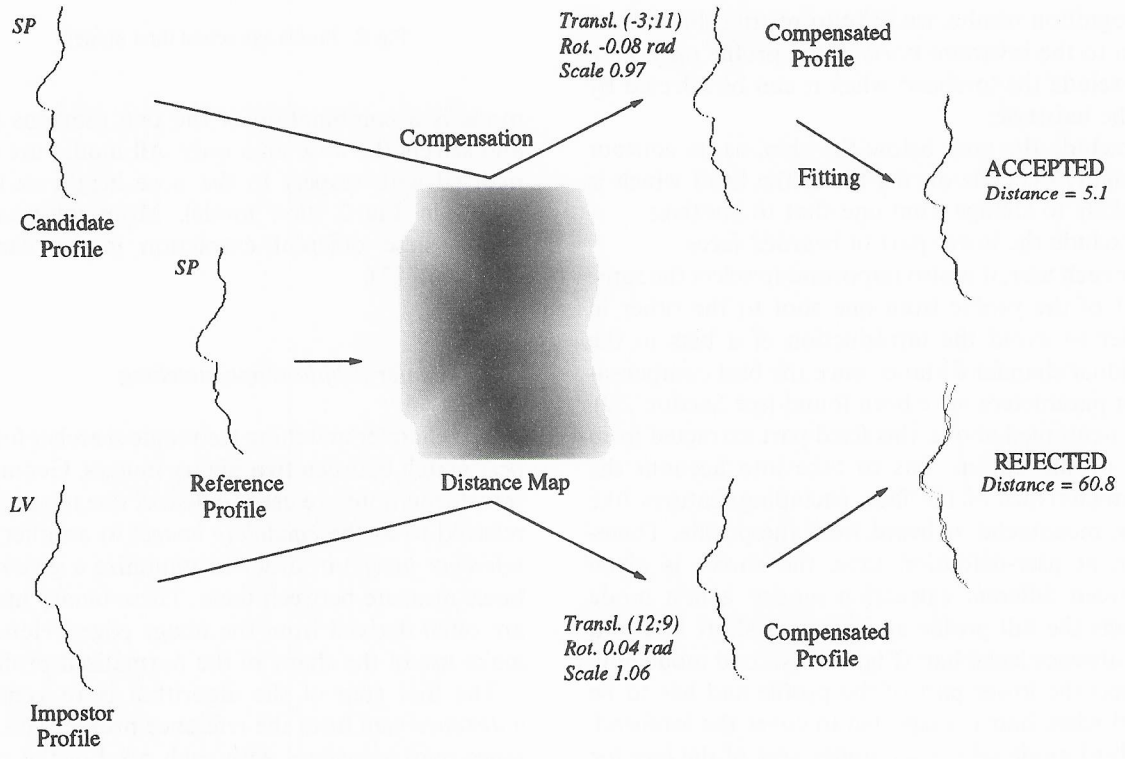

Fig. 3. The chamfer matching process.

multidimensional minimization method. We make use of the *Downhill simplex algorithm*, which requires only function evaluations and no derivatives [18]. The chamfer algorithm approximates the Euclidian distance with a maximum error of 6%. Hopefully, as the chamfer approximation is basically used as a cost function during the matching process, this error does not influence the quality of the profile shape mapping and let us avoid the computation of the real Euclidian distance.

The global matching process is illustrated in Fig. 3. First, the candidate profile is projected onto the reference distance map and a global distance is computed. By minimizing this distance, the optimum compensation parameters are found (t_x, t_y, θ_{xy} and z). Then, the residual distance between the best compensated and the reference profiles is used to decide whether the two profiles belong to the same person or not.

In order to avoid the simplex algorithm to converge towards a local minimum, attention must be paid to the initial parameters values used to initialize the algorithm. These values have to be as close as possible to the final solution for the algorithm to converge efficiently. t_x, t_y are estimated first by comparing the positions of the nose tip between the reference and the candidate profiles, z is found by comparing the two profile heights and θ_{xy} is arbitrarily set to zero. These values are used to initialize a first chamfer/simplex algorithm that runs into a low-resolution mode (the candidate profile and the reference distance map are down-sampled by a factor of 4 in both x/y directions). As output, we get refined values for t_x, t_y and z and a pretty good estimation for θ_{xy}. All these values are used to start the final full-resolution search. At the end of the chamfer matching process, the residual distance between the best compensated candidate and its associated reference is obtained and used as a matching score. This score will be further processed in Section 6.

3. Grey level profile matching

The second profile-related modality is based on grey level information along the shape of the profile and includes features like mouth width and height, nostrils, nose depth, eyes and eyebrows as accessed from the profile view. Once the best compensation parameters have been found during the chamfer matching process, the same parameters are used to compensate the candidate profile grey level image in order to issue a pixel-by-pixel comparison with the grey levels of the reference image. The Mean Squared Error (MSE) is used to express the distance between the reference profile and the compensated candidate.

Prior to the comparison, one has to normalize the grey level distribution between the two images to get rid of the illumination variability. Two kinds of normalization have been investigated:

- Dynamic normalization: the grey level distribution of the image is extended to its maximum dynamic (0..255). Unfortunately, this method does not work properly and a problem occurs when bright areas (e.g. when teeth are visible) are present in one image and not in the other (e.g. mouth closed). Assuming the same lighting conditions for the two images, one will already be close to the maximum dynamic (teeth luminance values are close to 255) while the other has to be normalized. This results in an overall brightness change in the second image but not in the first and causes the grey level comparison to fail.
- Mean normalization: the mean of the two images is set to half the maximum dynamic (127) by adding an offset to every pixel. This method has been used in our experiments.

Due to the presence of hair, the whole profile view cannot be used to carry out the comparison and only a small area taken along the profile shape has to be taken into account. The best results are given for an area width of 25 pixels and when the grey

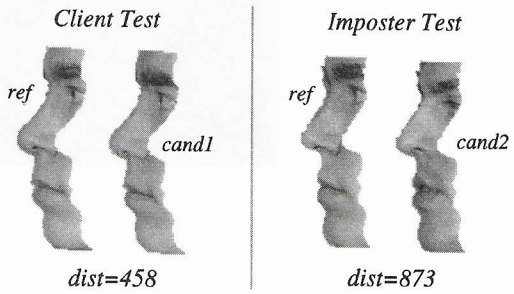

Fig. 4. Grey level matching.

level images are low-pass filtered prior to the matching. This low-pass filtering is meant to improve the quality of the matching in particularly noisy areas like the eyebrows. Some results of the grey level matching are shown in Fig. 4.

4. Grey level frontal face matching

Our frontal-based modality is similar to the grey level profile matching, in the sense that it also computes the grey level correlation between a candidate and a reference image, but using information from the frontal view instead of the profile. Prior to the matching itself, we make use of the same grey level normalization as described above. Images are low-pass filtered in both x/y directions in order to improve the quality of the matching. Again, the MSE criterion is used to compute the distance between the two grey level images.

In order to get rid of the face variability over the different shots, the grey level distance is computed inside a rectangular window that covers the most invariant features found inside the frontal view, namely the eyes/eyebrows and nose/nostrils features. This fixed window is automatically extracted from the input images – which are not low-pass filtered yet – using a technique that is similar to the technique proposed by [5]:

- from the frontal image, we compute the horizontal projection of horizontal gradient which offers a maximal peak at the eyes position and allows to precisely locate the vertical position of the eyes
- then, a horizontal 20 pixels-height strip is centered around the vertical eye coordinate and used to compute the vertical projection of the vertical gradient. This projection offers two distinct peaks at the horizontal locations of the two eyes.
- from these measures, we compute the coordinates of the point located in between the eyes. These coordinates are used to position a fixed-length matching window around the eyes and nose features (110×80 pixels).

This procedure is illustrated in Fig. 5.

Unlike the grey level profile modality that makes use of the compensation parameters issued from the chamfer shape matching, we don't know which parameters to apply in order to match the candidate's eye/nose window onto the reference image. Again, we will use the simplex algorithm in order to find the optimal frontal t_x, t_y, θ_{xy} and z parameters, and minimize the grey level distance between the two images. In order to speed-up the algorithm, approximative translation parameters (t_x, t_y) are provided to the first iteration of the simplex algorithm. These parameters are guessed by comparing the position of the point located in between

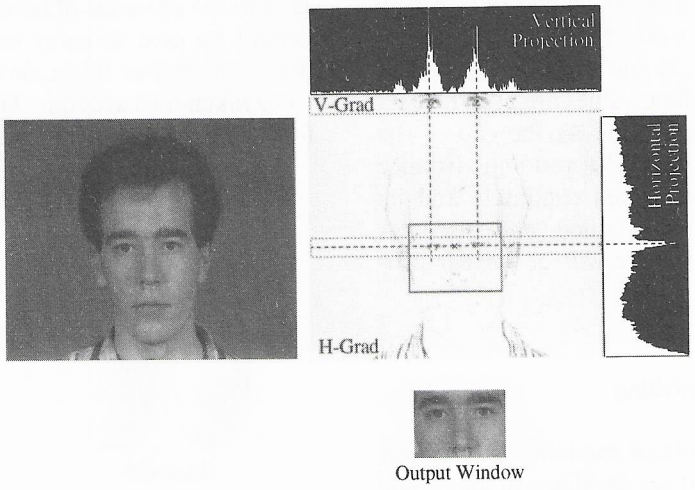

Fig. 5. Frontal features localization.

the eyes of the candidate and the reference images, according to the 3-step algorithm described above. As first estimates, $\theta_{xy} = 0$ and $z = 1$. Moreover, different scale factors are applied for the horizontal (z_y) and vertical (z_x) directions, allowing to compensate for a small rotation of the head around the vertical axis. However, this double scale factor also allows to better distort one impostor image onto the claimed reference image and is likely to degrade the performance of the system in terms of false acceptances. Therefore, only small variations between these two scale factors should be allowed. The cost function to be minimized by the simplex algorithm is modified accordingly:

$$\mathrm{MSE}(t_x, t_y, \theta, z_x, z_y) + \alpha|z_x - z_y|, \quad (1)$$

where α has been manually set to allow a maximal difference of 10% between the two scale factors.

5. Test setup

The M2VTS Multimodal Face Database [16] has been used to test the methods proposed here. This database is made of 37 faces, offers an overall resolution of 286×350 pixel and has been acquired under real conditions except for the nearly constant lighting and fixed background. The profile itself extends over a 100–150 pixel area. Profile views have been manually extracted from the *motion* sequences (see the M2VTS Database terminology) with a tolerance of about $90° \pm 15°$. The same tolerance applies for the manual selection of the frontal images. In our experiments, subjects do not wear glasses. Four different shots were used to perform our tests. These shots were taken at one week intervals. Different frontal views taken from the M2VTS database are shown in Fig. 6. Fig. 7 illustrates the different shots.

Our procedure for experimentation follows the M2VTS protocol [16]. One experiment uses a *training* and a *test* database. The *training* database is built of three shots (four are available) of 36 persons (37 available). The *test* database is built of the left-out shot of the left-out person and the left-out shot of the 36 persons present in the training database.

The training database is used for providing the *reference* models for each client but also to calibrate the different acceptance thresholds required during the test session. The performance of the identification algorithms is evaluated by matching the 37 candidate persons (36 clients and 1 impostor) from the test database with the 36 reference clients. Such an experiment provides 36 *authentic* and 36 *imposture* tests. An authentic test consists of candidate claims which are true. An imposture test consists of candidate claims which are false.

There are 4×37 possible experiments by leaving out one person and one shot, this means $4 \times 37 \times 36 = 5328$ client matches and 5328 impostor matches. All these 10656 matches were performed in order to evaluate the performance of our different algorithms and fusion schemes.

At last, authentication results are computed by matching the candidate (taken from test database) with each of its claimed references (the three shots taken from the training database) and by taking the best score (i.e. the lowest residual distance) as the final score.

6. Performance of the single modalities

A given match is accepted if its score (residual chamfer or grey level distance) is below a given threshold k. Otherwise, it is rejected. The decision threshold can be applied to the whole database (single decision threshold) or may depend on the claimed identity (individual thresholding). For each threshold – or set of thresholds in the case of individual thresholding – some clients are rejected (i.e. the false rejection FR(k)) while a number of impostors are able to enter the system (i.e. the false acceptance FA(k)). By varying the decision threshold k continuously, a receiver operating characteristics (ROC) curve can be drawn by plotting the different operating points (FA(k), FR(k)) in the same graphic.

ROC curves for the profile shape matching, the grey level profile matching and the grey level frontal matching are given in Figs. 8–10, respectively. Plain curves refer to the use of a single thresholding scheme, while dashed curves show the performance of a system where individual

Fig. 6. M2VTS database: some frontal views.

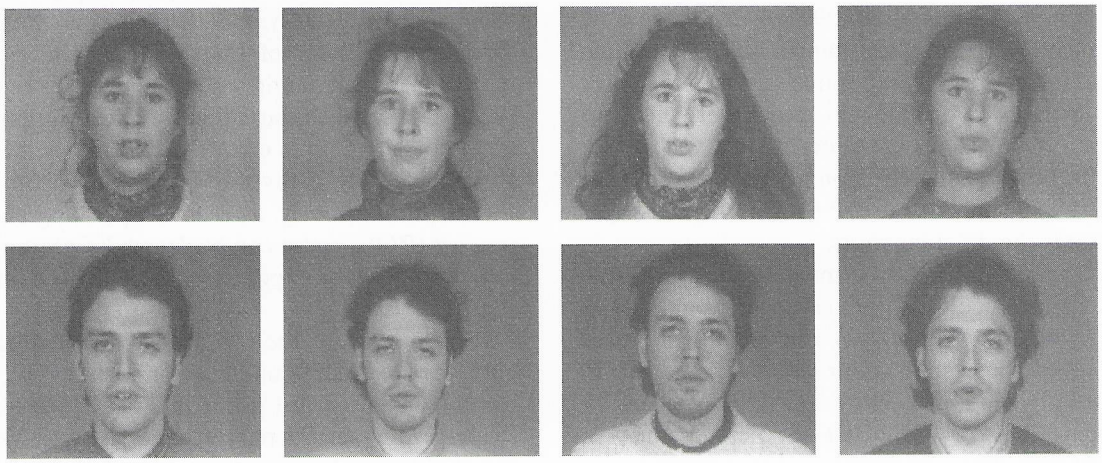

Fig. 7. M2VTS database: the different shots.

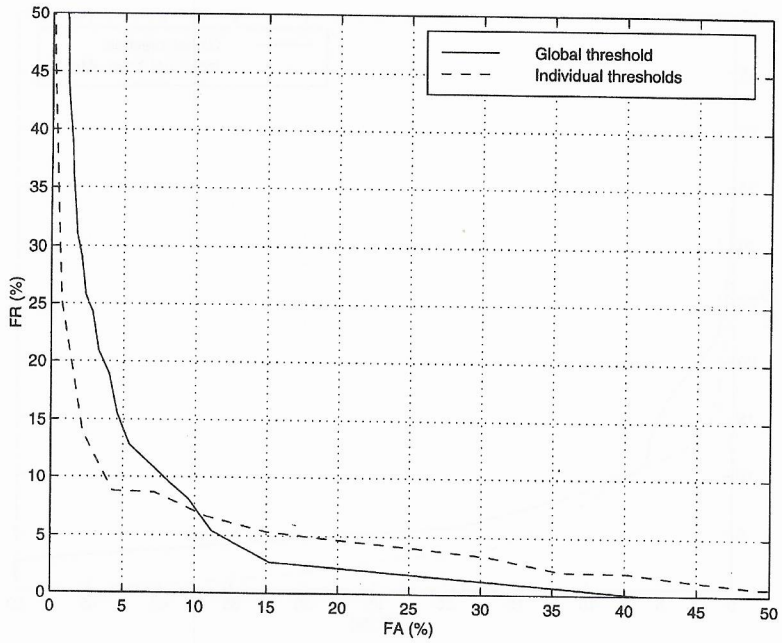

Fig. 8. ROC curves for the profile shape matching.

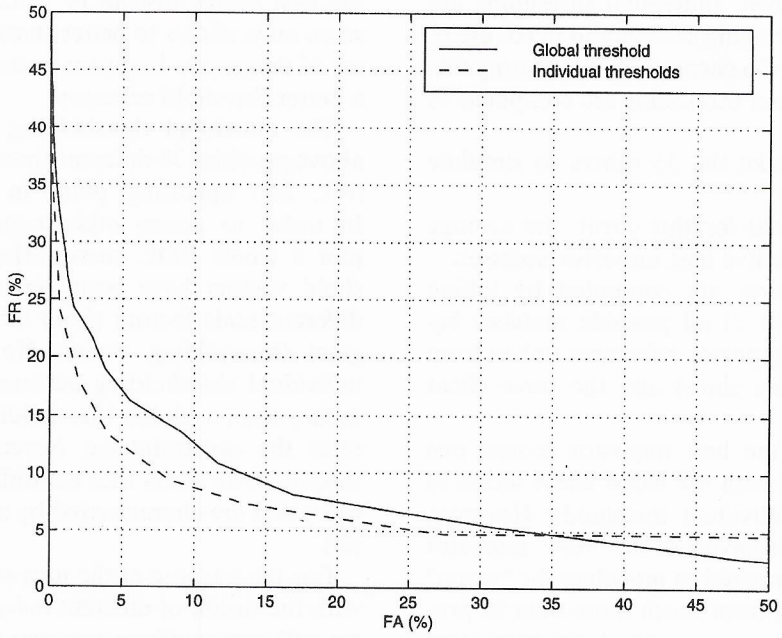

Fig. 9. ROC curves for the grey level profile matching.

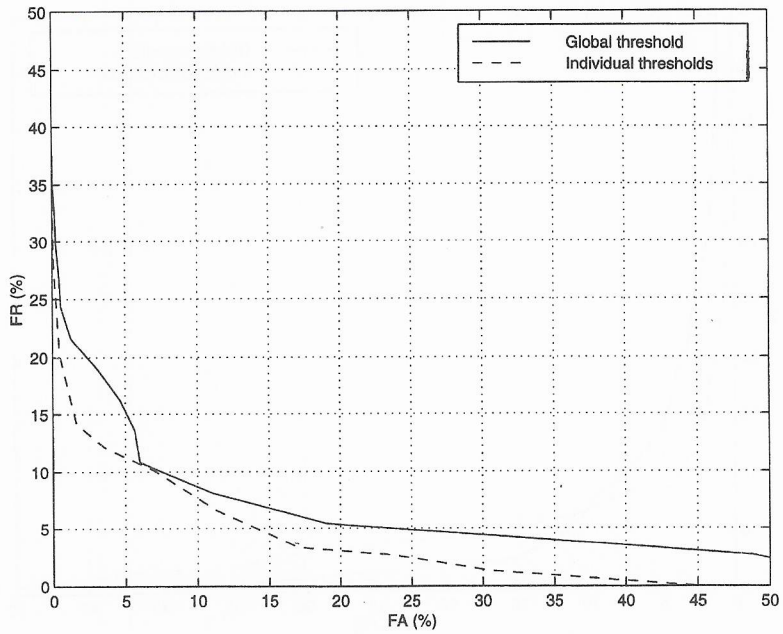

Fig. 10. ROC curves for the grey level frontal matching (frontal expert).

thresholds are applied. Individual thresholds are acquired during a training session and make use of the training dataset (36 clients/3 shots). During this session, the individual thresholds are computed as follows:

- for each client, take the 35 others to simulate impostor accesses;
- take as a threshold for this client, the average score between the five best impostor accesses;
- the impostor scores are computed by taking the best score out of all possible matches between the three impostor references (taken from the three available shots) and the three client references.

Instead of taking the best impostor scores, one could have worked with the worst client scores in order to fix the individual thresholds. However, characterizing the behavior of the "best" impostor is more reliable compared to modeling the "worst" client, as we benefit from much more data to process. Therefore, impostor scores have been used here above. Moreover, computing an average over the best scores instead of working with the best score only, allows to better characterize the behavior of dangerous impostor accesses, and results in a better threshold selection.

The individual thresholding scheme described above provides 36 different thresholds and a single (FA, FR) operating point in the ROC curve. In order to access other operating points and plot a whole ROC curve, other individual threshold vectors have been generated by applying different scale factors (from 0.5 to 5) to the original thresholding vector. Not surprisingly, an individual thresholding achieves a better performance than a global thresholding, as it allows to relax the constraint (i.e. increase the acceptance threshold) for users that are difficult to impost (i.e. users that are characterized by a distinctive feature set).

For the purpose of the next section, which deals with the fusion of different *independent* modalities, we will now combine our two profile-related modalities into one unique *expert*. This profile expert

will then be considered as being independent from the frontal grey level modality – hereafter referred to as the *frontal expert* – as it gives access to the depth information of the face, i.e. information which cannot be accessed from a frontal view and thus cannot be part of the extracted frontal features. Thanks to this independence, the fusion of these experts will result in a drastic improvement of the performance of our authentication scheme, as later shown in Section 8.

The shape and grey level profile modalities are combined into the profile expert by summing up their respective scores. In order to equally weight these two modalities, their scores are normalized with respect to their average client score over the training set. The ROC curve of the profile expert is shown in Fig. 11.

Table 1 summarizes the performance of the different modalities and experts by giving some particular operating points. The equal error rate (EER) stands for the point where FA = FR and the success rate (SR) refers to the operating point where 1 − FA − FR reaches a maximum. The FR is also given for an FA of 1%.

7. Fusion of independent experts

The fusion of different experts can be performed in two different ways: fusion at *decision* level or at *score* level. In the first case, different decisions (acceptance or rejection) are issued from each expert independently and then combined together according to simple rules like AND/OR operators or majority vote. On the other hand, one can also work by first combining the score *values* and then issue the final decision by thresholding the aggregated score.

A fusion scheme based on the scores (soft fusion) is likely to outperform a fusion that is based on individual decisions (hard fusion), since some information is lost after a thresholding is performed. We thus should work in the score value domain as long as possible during the fusion, then proceed to the final decision as the last step only. However, as long as simple soft fusion techniques are concerned, like a linear combination of the different scores for example, hard fusion may sometimes offer a better performance depending on the nature of the data to be merged, as shown in Section 7.2 and during our different tests.

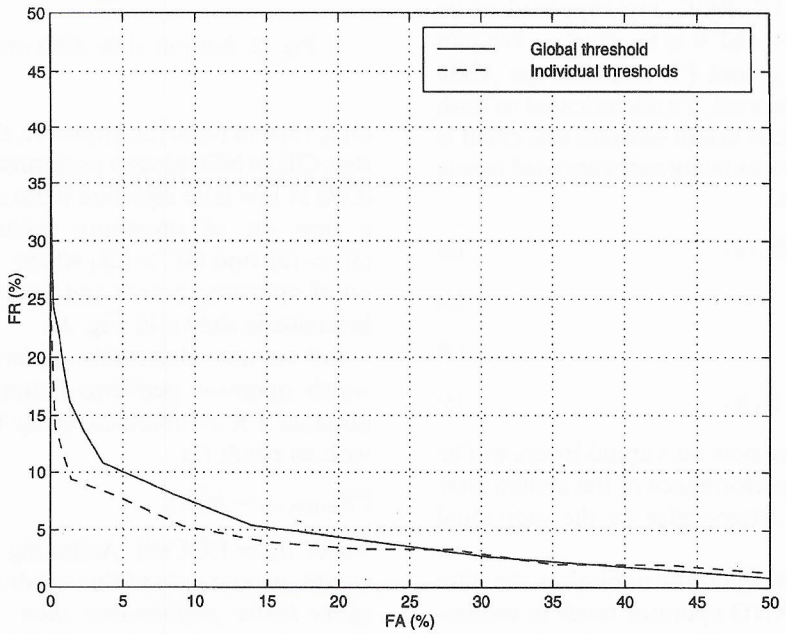

Fig. 11. ROC curves for the profile expert.

Table 1
Performance of the different modalities and experts

	Thresholding	EER (%)	SR (%)	$FR_{FA=1\%}$ (%)
Profile shape	Global	9	83.5	50
	Individual	8	87	23
Profile grey level	Global	11	78.5	29
	Individual	9	82.5	22.5
Profile expert	Global	8	85.5	18.5
	Individual	7	89	11
Frontal expert	Global	9	83	23
	Individual	8.5	84.5	17.5

Both approaches are studied in this section. The fusion at decision level is based on AND and OR operators, while the fusion at score level performs a thresholding on the best linear combination of the two modalities scores.

7.1. *Hard fusion: AND and OR operators*

Let us denote (fa_1, fr_1) and (fa_2, fr_2), two operating points of two independent experts with $fr_i = ROC_i(fa_i)$, $i = 1,2$. ROC_i are supposed to be decreasing functions and $0 \leqslant fa_i, fr_i \leqslant 1$. We can easily express the FA and FR rates for the AND (the client is authenticated, if authenticated in both modalities) and the OR fusion schemes (the client is authenticated as soon as being authenticated in one modality) as follows:

$$FA_{or} = fa_1 + fa_2 - fa_1 fa_2, \tag{2}$$

$$FR_{or} = fr_1 fr_2, \tag{3}$$

$$FA_{and} = fa_1 fa_2, \tag{4}$$

$$FR_{and} = fr_1 + fr_2 - fr_1 fr_2. \tag{5}$$

Please note that, from now on, capital letters will be used to refer to the performance of the system after fusion, while small letters refer to the individual experts.

Intuitively, the OR operator minimizes the false rejection while the AND operator tends to minimize the false acceptance. This can also be shown from the above formulas, as we always have

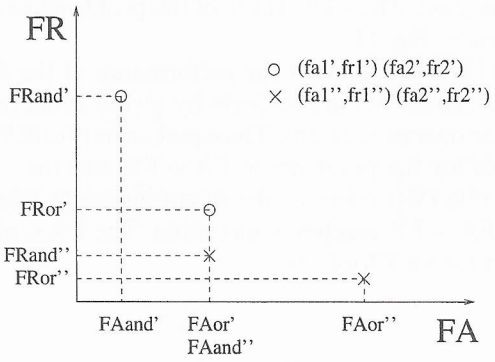

Fig. 12. Behavior of the AND and OR operators.

$(3) \leqslant (5)$ and $(4) \leqslant (2)$. However, this does not mean that OR (AND) always performs better than AND (OR) at low false rejection (false acceptance), as for a new set of operating points we can have: $(3)'' > (5)'$ and $(4)'' > (2)'$, where $'$ refers to the first set of operating points and $''$ to the new one. This behavior is shown in Fig. 12.

Let us now determine under which condition which operator performs better. At $FA = 0$, the minimal FR is obtained using the OR operator, with as result (3)

$$FR_{or|FA=0} = fr_1^0 fr_2^0, \tag{6}$$

where $fr_i^0 = ROC_i(0)$. Assuming continuous ROC curves, we can extend this result and state that *OR offers better performance than AND at low FA*. Depending on the values of fr_i^0 the FA range where this better performance can be observed, may vary:

for fr_i^0 close to 1, this range is too small for the OR to be really useful. However this case is not common in practice, as it refers to experts or modalities that could not distinguish clients from impostor reliably. For usual (low) fr_i^0, the OR operator offers improved performance over a wide range of low FA, as further shown when we will address Fig. 15.

From the knowledge of the two individual ROC curves, we can easily draw the ROC curve of the OR operator under the hypothesis of a small FA. By considering the linear approximation of ROC_i around zero, we can approximate Eqs. (2) and (3) by

$$FA_{or} \simeq fa_1 + fa_2, \tag{7}$$

$$FR_{or} \simeq (fr_1^0 - \alpha_1 fa_1)(fr_2^0 - \alpha_2 fa_2)$$

$$\simeq fr_1^0 fr_2^0 - \alpha_1 fr_2^0 fa_1 - \alpha_2 fr_1^0 fa_2, \tag{8}$$

where α_i denotes the absolute value of the ROC_i slope around 0. These expressions are valid for small values of false acceptances and false rejections only, i.e. the domain where useful operating points are located. For a given FA_{or} (which fixes the sum $fa_1 + fa_2$), the minimum of Eq. (8) is achieved when $fa_j = FA_{or}$, where j corresponds to the modality that offers the highest coefficient $\alpha_j fr_{(1-j)}^0$, and $(1-j)$ to the other modality. We then can express Eq. (8) the following way:

$$FR_{or} \simeq fr_1^0 fr_2^0 - \alpha_j fr_{(1-j)}^0 FA = fr_{(1-j)}^0 (fr_j - \alpha_j FA)$$

$$= fr_{(1-j)}^0 ROC_j(FA). \tag{9}$$

Fig. 13 shows two arbitrary ROC curves (interrupted lines), representing the behavior of two hypothetical modalities, and their fusion according to the OR rule. All possible combinations of operating points (fa_i, fr_i) were used to characterize the OR performance (dots). The approximation (9) (plain curve) reliably provides the best OR operating points, as long as low FA rates are concerned.

Thanks to the symmetry of Eqs. (2) and (5), the same kind of results remains valid for the FA rate of the AND operator, under the hypothesis of a low FR rate:

$$FA_{and} \simeq fa_{(1-j)}^0 ROC_j^{-1}(FR). \tag{10}$$

This behavior is shown in Fig. 14.

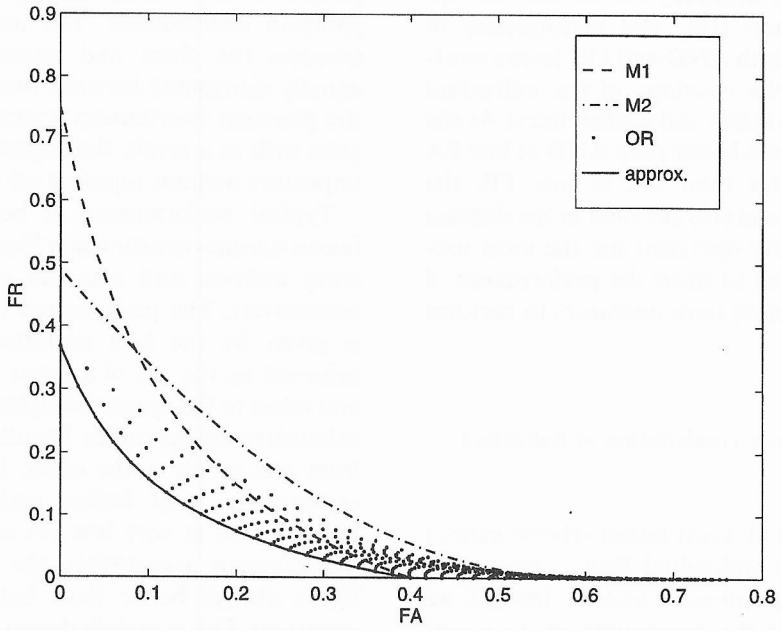

Fig. 13. OR fusion scheme: two hypothetical modalities (M1 and M2), the OR operating points and the results of approximation (9).

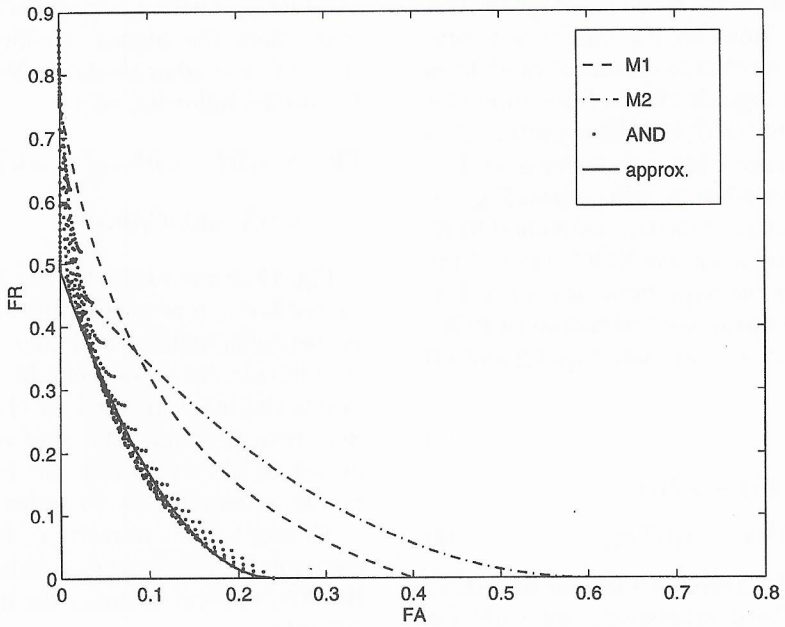

Fig. 14. AND fusion scheme: two hypothetical modalities (M1 and M2), the AND operating points and the results of approximation (10).

Fig. 15 compares the OR and the AND fusion strategies, based on arbitrary curves and the approximations (9) and (10). The performance of a system based on both AND and OR fusion strategies, is given by the envelope of the individual performance curves (plain and dashed lines). As the OR operator performs better than AND at low FA and the AND better than OR at low FR, the approximations (9) and (10) are valid in the domain where their respective operators are the most useful, and may be used to draw the performance of a system that combines both operators to perform a hard fusion.

7.2. Soft fusion: linear combination of individual scores

The performance of a soft fusion scheme cannot be drawn from the individual ROC curves anymore, as done in the previous section. Instead, we have to go down to the distribution of the scores inside each modality. Two kind of distributions have been considered here for the particular properties of their sum, namely the uniform and gaussian distributions. The uniform distribution assumes the client and impostor scores to be equally distributed between two boundaries, while the gaussian distribution assumes infinite boundaries with as a result, the impossibility to reject all impostors without rejecting all the clients.

Typical performances of both hard and soft fusion schemes are shown in Figs. 16 and 17 considering uniform and gaussian score distributions, respectively. The performance of the linear fusion is given by the best performance that can be achieved by the use of a linear score combination and refers to the optimal weighting as a result of an exhaustive weight search. Results are quite different from one graph to the other. Under the uniform assumption, hard fusion performs better than linear fusion at very low FA or FR, while under the gaussian assumption, the linear fusion performs always better than both AND and OR operators. This is mainly due to the fact that, under the gaussian assumption, it is impossible to have

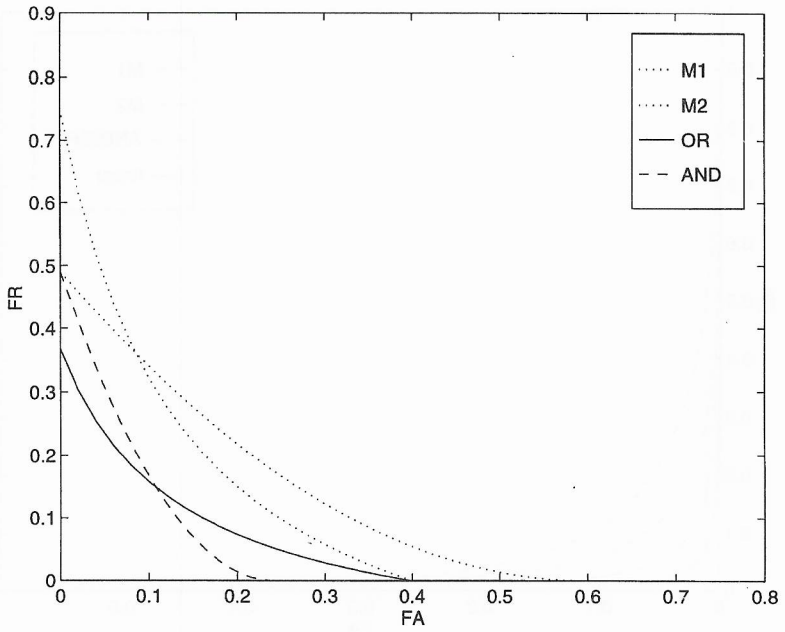

Fig. 15. Comparison between the OR and the AND fusion schemes.

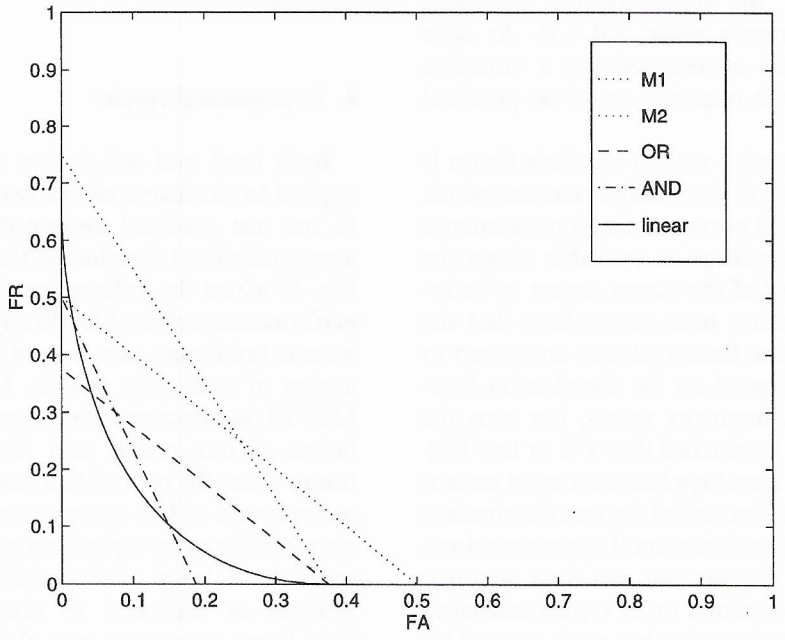

Fig. 16. Behavior of the hard and soft fusion schemes under the assumption of uniform score distributions.

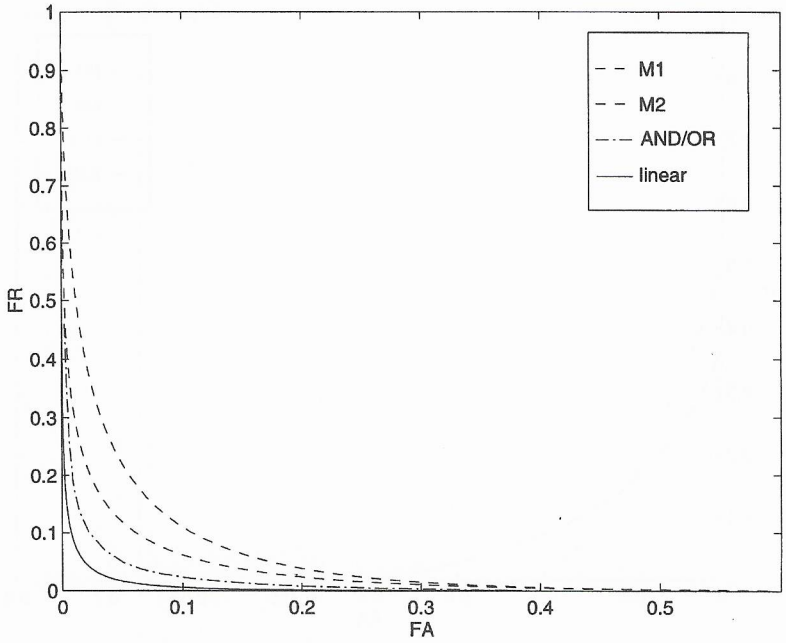

Fig. 17. Behavior of the hard and soft fusion schemes under the assumption of Gaussian score distributions.

less than 100% false rejection when FA = 0 as mentioned above, or, symmetrically, less than 100% false acceptance when FR = 0. As seen in Eqs. (2)–(5), this corresponds to a situation where OR and AND operator are of no practical use.

From all these results, we can conclude that it is almost impossible to theoretically determine which fusion scheme should provide the best performance as long as no information is available about the statistical properties of the expert scores to be integrated. Indeed, it has been shown here that the relative merits of one fusion scheme compared to the other, highly depend on the distribution function of client and impostor scores, but also the operating point to be reached (low FA or low FR). As the behavior of a real application expert cannot be represented by a theoretical statistic distribution without making some detrimental approximations, it seems better to directly use real data acquired during a training session in order to test and select the optimal fusion scheme for that application. This will be done in the next section, where the fusion of our experts will be tested using the M2VTS database.

8. Experimental results

Both hard and soft fusion schemes have been applied to the fusion of our two experts according to the test protocol described in Section 5 and using individual thresholds. Results are shown in Fig. 18 where the different curves refer to the best performance achievable through a given fusion scheme (minimum envelope of all possible combination of operating points). Making use of the M2VTS database and for the particular case of the fusion of our profile and frontal experts, hard fusion offers the overall best results. Moreover, the performance of OR operator supersedes the AND operator for every operating point. This behavior could have been predicted from the theoretical analysis of Section 7 as illustrated in Fig. 19. This figure represents two ROC curves that are roughly similar to our two experts as well as the

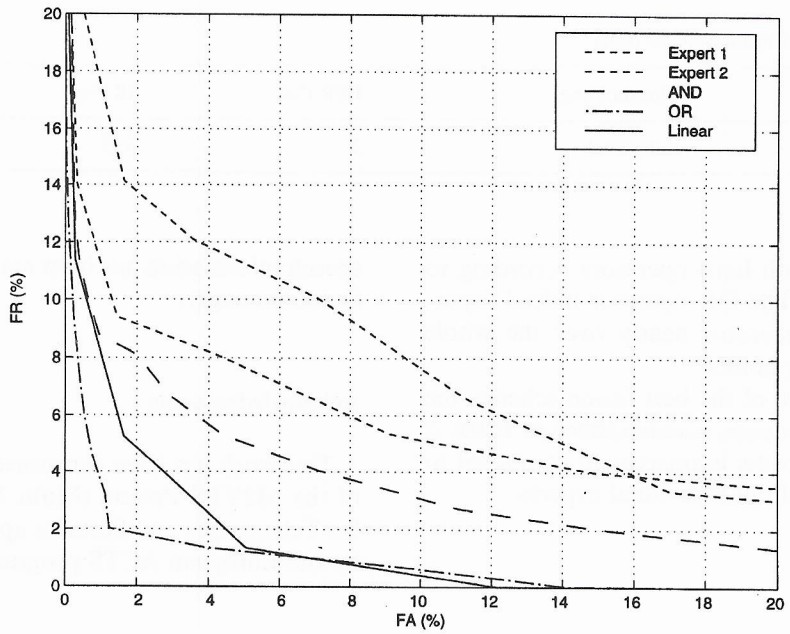

Fig. 18. Performance of the hard and soft fusion schemes on the M2VTS test database.

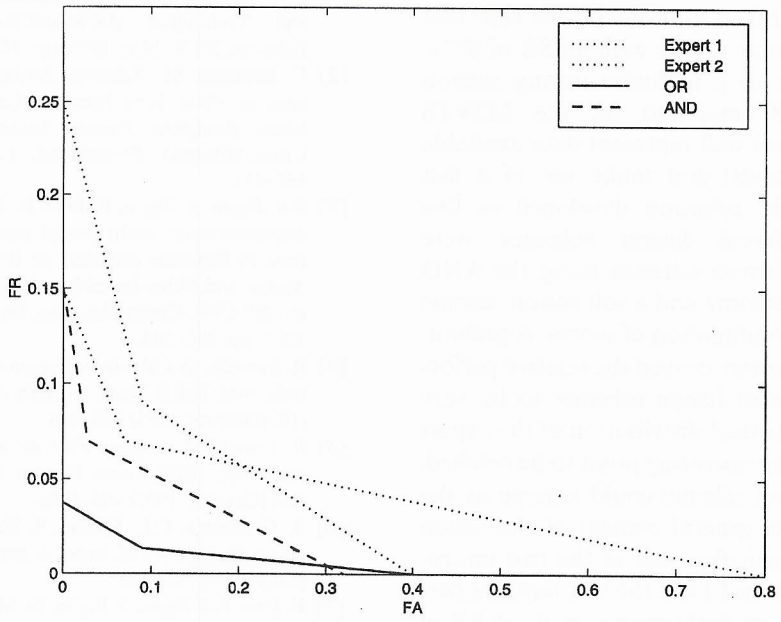

Fig. 19. Performance of both hard fusion operators on theoretical ROC curves that are roughly similar to the experts tested on the M2VTS database.

Table 2
Performance of the best fusion scheme (OR)

	Thresholding	EER (%)	SR (%)	$FR_{FA=1\%}$ (%)
OR fusion scheme	Individual	2	96.5	3

performance of both hard operators according to Eqs. (9) and (10). The OR operator indeed supersedes the AND operator nearly over the whole range of operating points.

The performance of the best fusion scheme, the OR operator in our case, is summarized in Table 2, and results in a drastic improvement compared to the performance of our individual experts.

9. Conclusion

This paper presented a novel multimodal person authentication approach based on static images taken from the frontal and profile facial views. Two reliable independent authentication experts were developed. One is based on frontal features only and offers a Success Rate of 84.5%, the other makes use of information taken from the profile view and achieves slightly better results with an SR of 89%. These figures refer to a intensive testing session (more than 10 000 matches) on the M2VTS database (which does well represent data available from real applications) and make use of a fast individual threshold selection developed in this contribution. Different fusion schemes were studied: two hard fusion schemes using the AND and OR logical operators and a soft fusion scheme based on a linear combination of scores. A preliminary theoretical analysis proved the relative performance of the different fusion schemes to be very sensitive to the statistical distribution of the expert scores but also to the operating point to be reached. No particular fusion scheme could emerge as the one to apply in the general context of the fusion problem. In the particular case of the two image-based experts developed here, the OR logical operator achieved the best performance, with an SR of 96.5%. This result rates high considering the database that has been used and the fact that no speech information has been made available to our fusion manager.

Acknowledgements

This work has been performed in the framework of the M2VTS Project (Multi Modal Verification for Teleservices and Security applications) granted by the European ACTS programme.

References

[1] M. Acheroy, C. Beumier, J. Bigün, G. Chollet, B. Duc, S. Fischer, D. Genoud, P. Lockwood, G. Maitre, S. Pigeon, I. Pitas, K. Sobottka, L. Vandendorpe, Multi-modal person verification tools using speech and images, in: Proc. European Conference on Multimedia Applications, Services and Techniques (ECMAST'96), Louvain-La-Neuve, Belgium, 28–30 May 1996, pp. 747–761.

[2] C. Beumier, M. Acheroy, Automatic profile identification, in: Proc. First Internat. Conf. Audio- and Video-based Biometric Person Authentication (AVBPA'97), Crans-Montana, Switzerland, 12–14 March 1997, pp. 145–152.

[3] E.S. Bigün, J. Bigün, B. Duc, H. Bigün, S. Fisher, Expert conciliation for multi modal person authentication systems by Bayesian statistics, in: Proc. First Internat. Conf. Audio- and Video-based Biometric Person Authentication (AVBPA'97), Crans-Montana, Switzerland, 12–14 March 1997, pp. 291–300.

[4] R. Brunelli, D. Falavigna, Person identification using multiple cues, IEEE Trans. Pattern Anal. Machine Intell. 17 (10) (October 1995) 955–966.

[5] R. Brunelli, T. Poggio, Face recognition: Features versus templates, IEEE Trans. Pattern Anal. Machine Intell. 15 (10) (October 1993) 104–1052.

[6] R. Chellappa, C.L. Wilson, S. Sirohey, Human and machine recognition of faces: A survey, Proc. IEEE 83 (5) (May 1995) 705–740.

[7] B. Duc, E.S. Bigün, J. Bigün, G. Maître, S. Fischer, Fusion of audio and video information for multi modal person authentication, Pattern Recognition Letters 18 (9) (1997) 835–843.

[8] F. Goudail, E. Lange, T. Iwamoto, K. Kyuma, N. Otsu, Face recognition system using local autocorrelations and multiscale integration, IEEE Trans. Pattern Anal. Machine Intell. 18 (10) (October 1996) 1024–1028.

[9] G. Borgefors, Hierarchical chamfer matching: A parametric edge matching algorithm, IEEE Trans. Pattern Anal. Machine Intell. 10 (6) (November 1988) 849–865.

[10] P. Jourlin, J. Luettin, D. Genoud, H. Wassner, Acoustic-labial speaker verification, Pattern Recognition Letters 18 (9) (1997) 853–858.

[11] J. Kittler, J. Matas, K. Jonsson, R. Sanchez, Combining evidence in personal identity verification systems, Pattern Recognition Letters 18 (9) (1997) 845–852.

[12] W. Konen, E. Schulze-Krüger, ZN-face: A system for access control using automated face recognition, in: Proc. Internat. Workshop on Automatic Face- and Gesture-Recognition, Zurich, Switzerland, 26–28 June 1995, pp. 18–23.

[13] J. Matas, J. Kittler, Spatial and feature space clustering: Applications in image analysis, in: Proc. 6th Internat. Conf. Comp. Anal. and Patterns, Prague, Czechia, 6–8 September 1995.

[14] B. Moghaddam, A. Pentland, Face recognition using view-based and modular eigenspaces, Automatic Systems for the Identification and Inspection of Humans, in: Proc. SPIE, Vol. 2277, July 1994.

[15] P.J. Phillips, H. Moon, P. Rauss, S.A. Rizvi, The FERET September 1996 database and evaluation procedure, in: Proc. First Internat. Conf. Audio- and Video-based Biometric Person Authentication (AVBPA'97), Crans-Montana, Switzerland, 12–14 March 1997, pp. 395–402.

[16] S. Pigeon, L. Vandendorpe, The M2VTS multimodal face database (Release 1.00), in: Proc. First Internat. Conf. Audio- and Video-based Biometric Person Authentication (AVBPA'97), Crans-Montana, Switzerland, 12–14 March 1997, pp. 403–409. See also http://www.tele.ucl.ac.be/M2VTS/.

[17] S. Pigeon, L. Vandendorpe, Profile authentication using a Chamfer matching algorithm, in: Proc. First Internat. Conf. on Audio- and Video-based Biometric Person Authentication (AVBPA'97), Crans-Montana, Switzerland, 12–14 March 1997, pp. 185–192.

[18] W.H. Press, S.A. Teukolsky, W.T. Vertterling, Numerical Recipes in C: The Art of Scientific Computing, Cambridge University Press, Cambridge, 1988.

[19] K. Yu, X. Jiang, H. Bunke, Face recognition by facial profile analysis, Proc. International Workshop on Automatic Face- and Gesture-Recognition, Zurich, Switzerland, 26–28 June 1995, pp. 208–213.