PR 2041
pp: 1-11 (col.fig.: Nil)
PROD. TYPE: COM
ED: Prathiba
PAGN: Thanuja – SCAN: Santhosh

ARTICLE IN PRESS

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

PERGAMON

Pattern Recognition III (III) III–III

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Multiple classifier combination for face-based identity verification

Jacek Czyz[a,*], Josef Kittler[b], Luc Vandendorpe[a]

[a] *Communications Laboratory, Université catholique de Louvain, Place du Levant 2, Louvain-la-Neuve 1348, Belgium*
[b] *Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, Surrey, GU2 5XH, UK*

Received 23 May 2003; received in revised form 5 December 2003; accepted 1 January 2004

Abstract

When combining outputs from multiple classifiers, many combination rules are available. Although easy to implement, fixed combination rules are optimal only in restrictive conditions. We discuss and evaluate their performance when the optimality conditions are not fulfilled. Fixed combination rules are then compared with trainable combination rules on real data in the context of face-based identity verification. The face images are classified by combining the outputs of five different face verification experts. It is demonstrated that a reduction in the error rates of up to 50% over the best single expert is achieved on the XM2VTS database, using either fixed or trainable combination rules.

Keywords: Face authentication; Face verification; Identity verification; Multiple classifier systems; Classifier combination; A posteriori probability; Linear discriminant analysis

1. Introduction

When one deals with a complex pattern recognition problem, such as the verification of personal identity, several different approaches can be considered, leading to different classifier designs. Typically, after performance evaluation only the best classifier is retained, the others being simply disregarded. Instead of retaining only the best classifier, one could use jointly the information provided by all of them. It is now well established that the combination of a set of classifiers designed for a given pattern recognition problem may achieve higher classification rates than any of the classifiers taken individually [1–3]. A major factor behind any improvement is the diversity in the classifier opinions [4], and methods such as boosting and bagging have been introduced to create diversity [5]. Another factor is the combination rule itself, that is, the rule that will be applied in order to get a unified decision with a reduced error rate, which is of interest in this paper.

Among all possible combination rules, the sum, product, maximum, minimum, vote and median rules, which effectively combine a posteriori class probabilities given by each classifier, have received very much attention [1,6–9], because of their simplicity and because they do not require training. If the classifiers operate in the same measurement space, different classifiers provide different estimates of the same posterior probability. An example of such classifiers is an ensemble of neural networks with various architectures, training strategies or initialisation parameters. Tumer and Ghosh showed analytically that averaging these different posterior estimates reduces the estimation noise, and therefore improves the decision [9]. If the classifiers operate in different measurement spaces, that is they relate to different phenomena through different sensors, independence between the classifiers can be assumed. In this case the optimal combining rule is the product rule, in which the combined decision is based on the product of the posteriors coming from the classifiers [1]. In fact, for two-class problems in which the posterior estimation noise is small, Tax et al. showed analytically that the sum and product rule perform the same classification, independently of the fact that the classifiers operate in the same or different measurement spaces [8]. Ex-

* Corresponding author. Tel.: +32-10-478066; fax: +32-10-472089.
E-mail address: czyz@tele.ucl.ac.be (J. Czyz).

0031-3203/$30.00 © 2004 Published by Elsevier Ltd on behalf of Pattern Recognition Society.
doi:10.1016/j.patcog.2004.01.008

periments presented in Ref. [7] confirm this property: the two rules perform in the same way up to an estimation noise level. Above this level, the product degrades causing the sum to become the best combination rule.

In practice however, the outputs of the classifiers may not be posterior probabilities but discriminants or *scores*, reflecting the confidence of the decision. In order to utilise efficiently the scores in the combination rules above, they should be transformed into posterior probabilities. This transformation is not trivial in general.

On the other hand, the transformation of scores given by each classifier into posterior probabilities can be avoided if the scores are viewed as features; features that are fed into a second-level or combining classifier in order to obtain the final decision [10]. This approach appears to be especially efficient in the context of biometric identity verification. The reasons are that (i) identity verification is a two class problem where the classifier has to decide whether a signal is genuine or not. Therefore, the classifier outputs a one-dimensional score on which the decision is based. (ii) In general, the score cannot be interpreted as posterior probability. As a result, in many identity verification systems that rely on classifier combination or fusion, the scores are treated as features, and a second level classifier such as support vector classifiers, neural networks, Parzen classifiers, etc., is constructed over these scores. This approach has been experimented for both multimodal fusion [11–15], where scores coming from multiple modalities like face and voice are combined, and intramodal fusion [16–19], where scores coming from the same modality but different matchers are combined.

The second level classifier approach treats scores as arbitrary numerical features, discarding their confidence nature. Combination techniques taking the nature of scores into account, such as the probability rules cited above, may be more appropriate in some circumstances.

The contribution of this paper is twofold. Firstly, in the case of a two class problem in an identity verification scenario, we point out the link between combining posterior probabilities and combining directly the scores, when all classifiers operate in the same measurement space. Starting from the product combination rule, which is optimal under restrictive hypotheses, we relax the assumptions to study how it affects the combination efficiency. We compare it with the other posterior combination rules. The difference with existing work lies in the fact that we assume score distributions for each classifier instead of assuming estimation error distributions. This gives practical relevance to our study when choosing a particular fusion strategy. We present simulation results when scores have Gaussian distributions as well as results with real data in the case of combining multiple face verification classifiers.

Secondly, we propose an identity verification system based on an efficient and simple fusion strategy of multiple face verification algorithms. The same face image is used as input for several classifiers. The estimates of class a posteriori probability given by the classifiers are fused leading to the final decision. Although the individual classifiers are independently optimised, it is demonstrated that the fused system substantially decreases the error rates over the best individual face classifier.

The paper is organised as follows. In the next section, we introduce biometric identity verification and intramodal fusion. Probability-based and second level classifier-based combination rules are discussed in Section 3. The effect of correlation between classifiers on probability combination rules is emphasised. In Section 4, after describing the face verification algorithms, we present the combination results for face-based identity verification. Conclusions are given in the last section.

2. Intramodal fusion of biometrics experts

Biometric identity verification [20] 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.

$$\begin{aligned} s(\mathbf{x}) &\leqslant \eta \quad \text{genuine or } \omega_a, \\ s(\mathbf{x}) &> \eta \quad \text{impostor or } \omega_b. \end{aligned} \tag{1}$$

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.

In order to increase the verification performance, one may take advantage of multiple classifiers, or experts, that provide their opinions on the same biometric data, and perform *intramodal fusion*. Given a measurement $\mathbf{x}$, each expert i outputs an estimate of the posterior probability $f_i^c(\mathbf{x})$, where $c \in \{a, b\}$ based on $\mathbf{x}$. These estimates can therefore be seen as different versions of the true (unknown) posterior corrupted by estimation noise [9,21]. This suggests that a better estimate can be obtained by averaging the different estimates.

Except in some special cases such as k-NN classifiers or neural networks, experts usually do not output posterior probabilities but scores $s_i(\mathbf{x})$. If one intends to combine classifiers by the averaging or product rule, scores have to be mapped to posterior probabilities. Depending on the classifier, several mappings are available, for example the logistic function [22]. Here we propose to use the score a posteriori

probability as the estimate $f_i^c(\mathbf{x})$, that is

$$f_i^c(\mathbf{x}) = P(\omega_c|s_i(\mathbf{x})) = \frac{p(s_i(\mathbf{x})|\omega_c)P(\omega_c)}{p(s_i(\mathbf{x}))}.$$

Clearly this mapping from s_i to f_i^c is non-linear, hence averaging the s_i and the f_i^c does not have in general the same effect on the combined decision. To compute $P(\omega_c|s_i(\mathbf{x}))$, the last equation shows that it is necessary to estimate the one-dimensional probability distribution function $p(s_i|\omega_c)$. This operation requires some training data similar to the data used for estimating the error rates and the threshold as suggested in the preceding section.

3. Combination rules

In this section, we present the different score combination rules that use posterior probabilities. Then we compare the combination accuracy in the case of Gaussian distributed scores with optimal combination for different values of the distribution parameters. We then discuss second level classifiers employed for the problem of identity verification.

3.1. Combining posterior probabilities

Consider a pattern $\mathbf{x}$ that needs to be classified as either ω_a or ω_b. We have at our disposal R experts, each of them outputs a score $s_i(\mathbf{x})$ which reflects its opinion about pattern $\mathbf{x}$. For clarity, we drop the dependency on $\mathbf{x}$ in the following. The optimal decision rule, is the Bayes rule which assigns to the pattern the class with maximum posterior probability, i.e. choose ω_a if

$$P(\omega_a|s_1, s_2, \ldots, s_R) > P(\omega_b|s_1, s_2, \ldots, s_R). \quad (2)$$

This is the best combination that can be made using the scores as it minimises the probability of error. Under the assumption that the scores s_i are class conditionally independent for both classes and equal priors, the product rule can be derived from (2), that is, choose class ω_a if

$$\prod_{i=1}^{R} P(\omega_a|s_i) > \prod_{i=1}^{R} (1 - P(\omega_a|s_i)) \quad (3)$$

otherwise choose ω_b. When the scores are not conditionally independent, this rule is no more optimal in general. In fact, for sufficiently accurate experts, scores are likely to be positively correlated because the experts will agree and classify correctly the majority of patterns. In addition to the product, several other combination rules have been introduced and may be more appropriate when score independence is not satisfied:

- Sum : $\sum_k P(\omega_a|s_k) > \sum_k P(\omega_b|s_k)$.
- Max: $\max_k P(\omega_a|s_k) > \max_k P(\omega_b|s_k)$.
- Min: $\min_k P(\omega_a|s_k) > \min_k P(\omega_b|s_k)$.

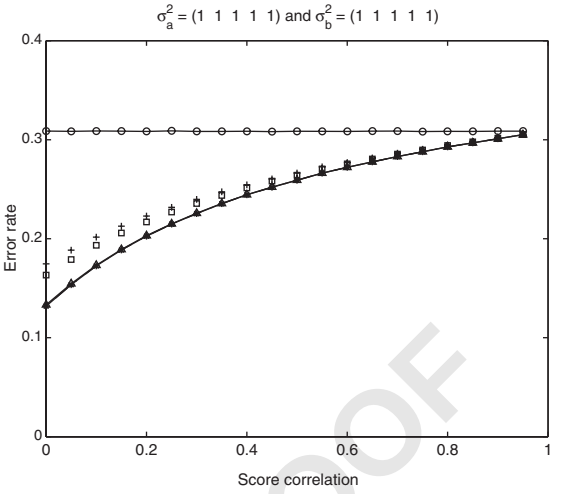

Fig. 1. Combination error versus expert correlation ρ_c in the case $\sigma_{ai} = \sigma_{bi} = 1$ for $i = 1, 2, \ldots, 5$. Key: ○ best single expert; □ max/min;+vote/median; △ sum; * product; × Bayes error.

- Vote: $\sum_k u(P(\omega_a|s_k) - 0.5) > \sum_k u(P(\omega_b|s_k) - 0.5)$, where $u(.)$ is the unit step function.
- Median $\mathrm{med}_k P(\omega_a|s_k) > \mathrm{med}_k P(\omega_b|s_k)$.

For two-class problems, the min and max rules perform exactly the same classification, the same applies for the median and vote rules when combining an odd number of experts [6]. Notice that none of the rules above make use of the correlation between the experts. Only the statistics of the individual score distribution s_k are needed to apply the rules.

3.2. Simulation: Gaussian score combination

The score densities $p(s_i|\omega_c)$ are in general unknown. To study the effect of dependence between the experts, we must assume a certain score distribution. Suppose we have R experts that output Gaussian scores. Each score distribution is characterised by the variances σ_{ci}^2 and the means μ_{ci} $(c \in \{a, b\})$. The class conditional joint score densities $p(s_1, s_2, \ldots, s_R|\omega_c)$ are also Gaussian with covariance matrices Σ_c, and means $\boldsymbol{\mu}_c$ where the variance σ_{ci}^2 is the ith diagonal element of Σ_c and μ_{ci} the ith element of $\boldsymbol{\mu}_c$ $(c \in \{a, b\})$. The properties of the experts are determined by the covariance matrices Σ_c and means $\boldsymbol{\mu}_c$ as (i) the correlation between the experts is reflected in the off-diagonal elements of Σ_c, (ii) the accuracy of the ith expert depends on σ_{ci}^2 and $\mu_{ai} - \mu_{bi}$. In this particular case, the Bayes error can be computed and compared to the error given by a posteriori probability combination rules, when scores are not independent.

In the simulation results presented below the case $R = 5$ is studied and all pairs of experts share the same correlation ρ_c. Class ω_a is centred at the origin and class ω_b is centred at $(1, 1, 1, 1, 1)^T$. Figs. 1–3 show the classification er-

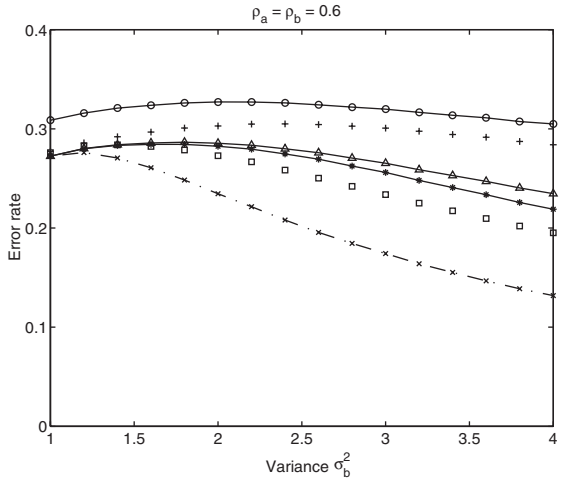

Fig. 2. Combination error versus σ_{bi}^2 in the case $\sigma_{ai} = 1\ \forall i$ and $\rho_c = 0.6$. Key: ○ best single expert; □ max/min;+vote/median; △ sum; * product; × Bayes error.

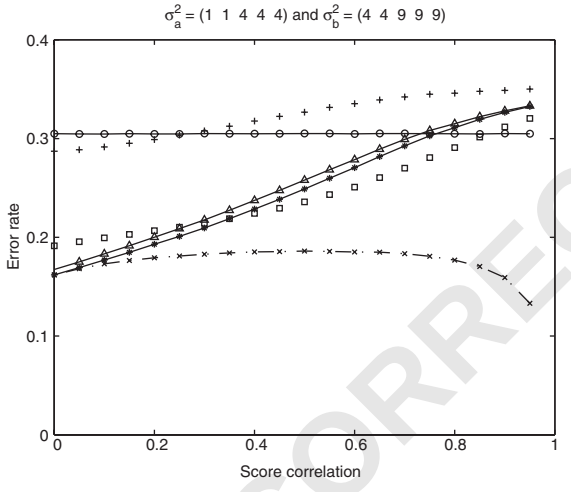

Fig. 3. Combination error versus expert correlation ρ_c in the case $\sigma_{ai} = 1$ and $\sigma_{bi} = 4$ when $i \in \{1, 2\}$, and $\sigma_{ai} = 4$ and $\sigma_{bi} = 9$ when $i \in \{3, 4, 5\}$. Key: ○ best single expert; □ max/min;+vote/median; △ sum; * product; × Bayes error.

ror obtained by the probability combination rules versus the Gaussian parameters for several representative situations. In Fig. 1 the combination error versus the score correlation is shown in the case $\sigma_{ai}^2 = \sigma_{bi}^2 = 1$ for $i = 1, 2, \ldots, 5$ and $\rho_a = \rho_b$. It can be seen that for low correlation values, the combination brings a substantial improvement over the single expert classification. The improvement decreases as correlation increases since the experts provide gradually more similar opinions and become redundant. In this particular Gaussian case, the product rule achieves the same error rate as the optimal combination, independently of the correlation value. In fact, product is optimal if $\sigma_{ai} = \sigma_{bi} = \sigma_i\ \forall i$ and the ratio

$$r = \frac{\sigma_i(\mu_{aj} - \mu_{bj})}{\sigma_j(\mu_{ai} - \mu_{bi})} \tag{4}$$

is equal to 1 (equal accuracy condition) and the correlation is the same for all expert pairs i and j (see Appendix A). The sum performs very closely to the product as predicted by the theory [8] while the max/min and vote are above the optimal rule for low correlation values. This suggests that when combining several experts that are very similar (equal accuracy and equal correlation between expert pairs) and the score variances are equal for both classes, the product rule is close to the optimum. Fig. 2 shows the error rate in the case $\rho_a = \rho_b$=0.6 and $\sigma_{ai}^2 = 1$ for $i = 1, 2, \ldots, 5$ versus σ_{bi}^2. When $\sigma_{bi}^2 = 1\ \forall i$ all the rules perform close to optimal. As the σ_{bi}^2 increase the rules depart from optimality quickly, they stay however below the single expert error rate. Interestingly, the max/min rule performs the best when the difference between the two score variances is large. Again, sum and product show similar performance while vote/median is the worst. When combining experts showing different accuracies (see Fig. 3), the product rule provides error rates close to optimal up to a correlation of 0.35. Above this limit, the best rules become the max/min rules. When correlation is further increased, the combination rules may degrade the performance versus the best single expert. The vote rule provides very little improvement even for low values of correlation.

To summarise, the product rule is the optimal in two cases: (a) experts are independent or (b) experts are Gaussian with equal accuracy and correlation, and equal variances for the two classes. In case (a), Fig. 3 shows that for moderate values of correlation (below 0.3), product may be close to optimality. In case (b), Fig. 2 shows that, with high correlation between experts, product rule is close to optimality only if σ_{ai}^2 and σ_{bi}^2 are not very different. If they are, max/min rule could be used instead.

3.3. Direct combination of scores

As mentioned in the introduction, the use of posterior probabilities can be avoided if a second level classifier is trained to combine the scores. The classifier is therefore implicitly learning the dependencies between the different experts. In Ref. [16], a non-parametric Parzen estimation technique is used to estimate the joint score density and rule (2) is used directly for combining several fingerprint matchers. A support vector classifier and a neural network fusion are compared in Ref. [15] for audio–visual authentication. A weighted averaging of scores, which corresponds to a linear discriminant function in the score space is proposed in Refs. [12,13,19]. In Refs. [17,14], a logistic regression based combination is used. In identity verification, the number of experts that are combined is usually small, which results in low dimensionality for the score space. For this reason, the dimensionality to number of sample ratio is often favourable.

Note that such a fusion system is not modular since the combiner must be re-trained when a new expert is added.

4. Face verification expert combination

In this section we start by presenting our face verification experts. Then we describe the face database and the experimental protocol. Finally, combination results using both a posteriori probability combination rules and trainable combiners are given.

4.1. Face experts

Two of our five experts are based on linear discriminant analysis (LDA), two on probabilistic matching (PM) and one on colour histogram comparison. For all experts, the first step involves localisation and registration of the face part in the input image. In order to keep focus on the classification task, we have skipped this step by manually locating the eye coordinates in the image. The face image is cropped and photometric normalisation is applied to reduce the effect of lighting variation. Note that all the experts use the same eye coordinates, however they operate on different image sizes and croppings.

4.1.1. LDA-based experts

After cropping, the image is transformed into gray levels and histogram equalised. The Fisherface approach [23] is used to extract features from the face image. This feature extraction technique is based on Principal Component Analysis (PCA) and 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 $\mathbf{x}_i$, each belonging to one of N classes or persons in identity verification $\{C_1, C_2, \ldots, C_N\}$, we compute the between-class scatter matrix S_b

$$S_b = \sum_{i=1}^{N} (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^{\mathrm{T}}$$

and the within-class scatter matrix, S_w

$$S_w = \sum_{i=1}^{N} \sum_{\mathbf{x}_k \in C_i} (\mathbf{x}_k - \boldsymbol{\mu}_i)(\mathbf{x}_k - \boldsymbol{\mu}_i)^{\mathrm{T}},$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}$ 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

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

is a solution of the eigenproblem [24]

$$S_b W = S_w W \Lambda, \tag{5}$$

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 first expert referred to as **LNC** uses normalised correlation to compute the score given by

$$s_1 = \frac{\mathbf{y}^{\mathrm{T}} \boldsymbol{m}_p}{\|\mathbf{y}\| \, \|\boldsymbol{m}_p\|},$$

where $\mathbf{y}$ and $\mathbf{m}_p$ are, respectively, the test and template images expressed in the LDA subspace.

The second expert referred to as **GDM** uses the gradient direction metric [25] for the matching process. In contrast to normalised correlation, this technique uses the other classes to define the score function. The idea is to compute the distance between the probe $\mathbf{y}$ and the reference $\boldsymbol{m}_p$ in the LDA subspace along the most discriminative direction. This direction is given by the gradient of the a posteriori probability $P(C_p|\,\mathbf{y})$ of person C_p given $\mathbf{y}$. The distribution $p(\mathbf{y}|C_j)$ is assumed to be Gaussian with mean $\mathbf{m}_j = W^{\mathrm{T}}\boldsymbol{\mu}_j$ and identity covariance matrix. Under this assumption the gradient of the a posteriori probability of person C_p given $\mathbf{y}$ can be written

$$\nabla P(C_p|\,\mathbf{y}) = a \sum_{\substack{j=1 \\ j \neq p}}^{N} p(\mathbf{y}|C_j)(\boldsymbol{m}_j - \boldsymbol{m}_p),$$

where a is a scalar that does not influence the gradient direction. See [25] for further details. The GDM score for person C_p is therefore given by

$$s_2 = \frac{|(\mathbf{y} - \boldsymbol{m}_p)^{\mathrm{T}} \nabla P(C_p|\mathbf{y})|}{\|\nabla P(C_p|\mathbf{y})\|}.$$

4.1.2. Probabilistic matching experts

In the Probabilistic Matching method [26] proposed by Moghaddam et al. the algorithm classifies the pixel intensity difference $\boldsymbol{\Delta} = \mathbf{x} - \mathbf{x}_r$ between the probe and the reference images as intraclass variations Ω_I or interclass variations Ω_E. The similarity between the probe and reference is based on the probability of image variation $P(\Omega_I|\boldsymbol{\Delta})$. This probability can be written using the probability densities $p(\boldsymbol{\Delta}|\Omega_I)$ for intraclass and $p(\boldsymbol{\Delta}|\Omega_E)$ for interclass variations thanks to Bayes formula. These high-dimensional probability densities are assumed to be Gaussian and are obtained from training data using an eigenspace density estimation technique. The method relies on principal component analysis to form a low-dimensional estimate of the complete density which can be evaluated using only the first principal components of $\boldsymbol{\Delta}$.

Probabilistic Matching associated with two different photometric normalisation techniques is used to obtain the experts **PM1** and **PM2**. For PM1, the image pixels are transformed to have a zero mean and a unit variance and for PM2 the images are normalised by histogram equalisation.

4.1.3. Colour histogram based expert

In contrast to the preceding experts, the fifth expert **HST** uses only the colour information contained in the face to make the decision. To limit the effect of lighting variation, image pixels are first rescaled to have zero mean and unit variance. Images are represented in the HSV colour space and the histogram of the Hue component (H) is computed for each image. The score is computed by taking the L_1 norm between the H component of images $\mathbf{x}$ and $\boldsymbol{\mu}_p$, i.e.

$$s_5 = \sum_i |h_i(\mathbf{x}) - h_i(\boldsymbol{\mu}_p)|,$$

where $h_i(.)$ represents the ith histogram bin value. The choice of the colour component H and the L_1 norm results from optimisation on the validation data set (see next section).

Note that histograms do not contain any information about the image topology, the face morphology is completely disregarded. For this reason and due to variability of colour, the individual performance of expert HST are low. However as expert HST operates on colour only, it provides information little correlated to the other experts.

4.2. Database and experimental protocol

Our experiments were performed on frontal face images from the extended M2VTS database (XM2VTS) [27]. XM2VTS is a publicly available multimodal database recorded specifically for assessing the performance of biometric approaches to identity verification. It contains face and speech recordings of 295 persons. The subjects were recorded in four separate sessions uniformly distributed over a period of 5 months. One session consists of two recordings. A detailed description of the standard experimental protocol can be found in Ref. [28]. The protocol divides the database into 200 clients or system users (used for testing false rejection rate) and 95 impostors (used for testing false acceptance rate). For each client/impostor one frame per recording per session is selected thus giving 8 images per subject in total. The total number of images is thus 2360 images. The protocol specifies a partitioning of the database into disjoint sets for training, validation and testing. For the clients, three images are used for training, three for validation and two, corresponding to the last session, for testing. For the impostors, 25 impostors are used for validation and the remaining 70 for testing. Therefore, the validation set leads to 600 genuine or client matchings (3 images $\times$200 clients) and 40 000 impostor matchings (8 images $\times$25 impostors $\times$200 clients) because each impostor image is matched against the 200 clients. As for the test set, it generates 400 genuine or client matchings (3 images $\times$200 clients) and 112 000 impostor matchings (8 images $\times$70 impostors $\times$200 clients). The training set serves to compute the LDA subspace for experts GDM and LNC, and to estimate the intra-personal and inter-personal densities for experts PM1 and PM2. In the single expert scenario, the validation data scores serve to set the threshold η_{EER} at the equal error rate (EER). The individual expert performance is then assessed on the test data by measuring the FAR(η_{EER}) and FRR(η_{EER}). In the multiple expert scenario, the validation set scores are used to train the combiner or estimate the score distributions $p(s_i|\omega_a)$ in the a posteriori probability combination case. The multiple expert system is then assessed on test data.

Table 1
Individual verification results on the XM2VTS database

	Valid	Test		
Expert	EER	FAR	FRR	HTER
GDM	2.17	3.00	2.69	2.84
LNC	2.95	2.75	3.83	3.29
PM1	4.67	3.75	5.57	4.66
PM2	3.45	3.00	4.73	3.86
HST	16.18	18.25	16.46	17.36

The verification result is measured with the equal error rate on the validation set (EER) and the false acceptance (FAR), false rejection (FRR) and half total error rate (HTER) on the test set.

4.3. Results

In this section, we present the individual expert performance and combination results obtained on the XM2VTS face database.

4.3.1. Individual expert performance

Table 1 shows the verification results of the experts presented in the previous section taken individually. The first column of the table shows the EER obtained on the validation set. The three last columns show the false acceptance and false rejection rates and the half total error rate HTER = (FAR + FRR)/2 obtained on the independent test set at the same threshold.

From Table 1, it can be seen that expert GDM offers the best verification rates (half total error rate of 2.84%). The other experts show error rates slightly higher except for expert PM1 (probabilistic matching based with zero mean and unit variance pixel value normalisation) and HST (colour histogram) which have an half total error rate of 4.66% and 17.36%, respectively. The expert GDM, LNC and PM2 have approximatively the same accuracy.

Table 2 shows the second-order statistical properties of the experts. From the table it can be noticed that the expert HST has a low correlation with the other gray level-based experts. The ratio r is defined by Eq. (4).

4.3.2. Trainable combiners

To compare the probability combination rules to trainable combiners, the scores were combined with the following classifiers: weighted averaging, Bayes quadratic classifier

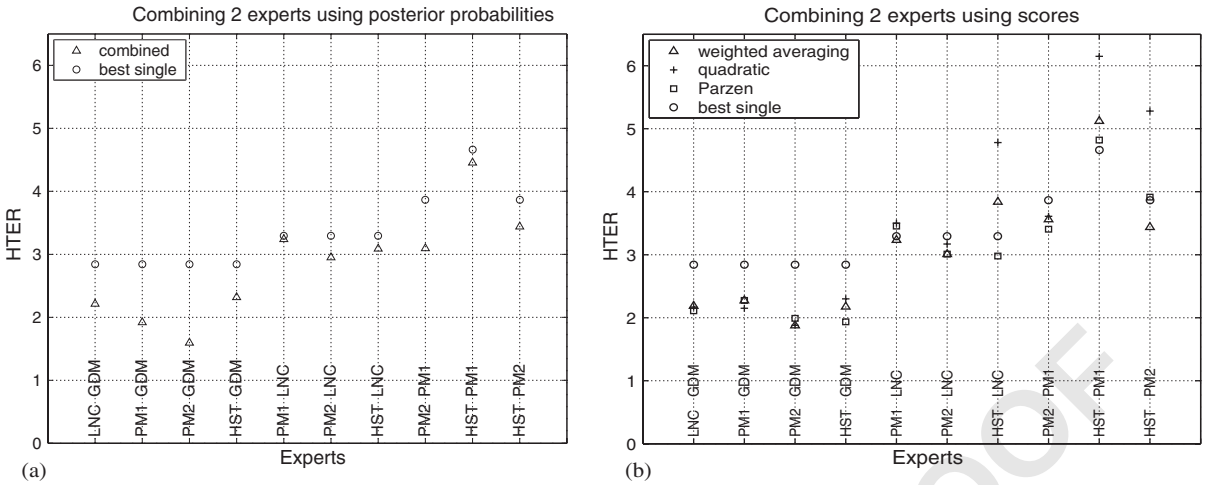

Fig. 4. Combination of two face verification experts. Left a posteriori probability combination rules. Right: trainable combiners.

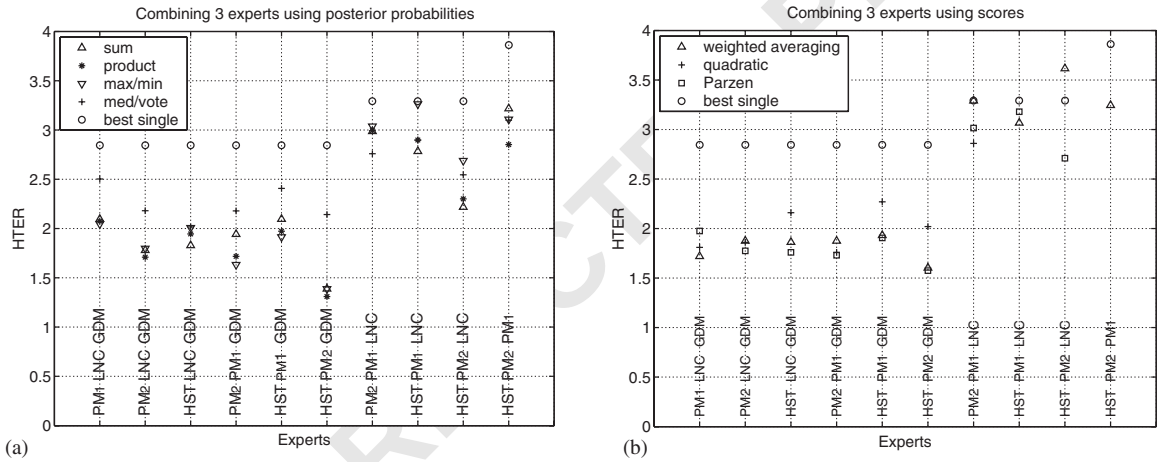

Fig. 5. Combination of three face verification experts. Left: a posteriori probability combination rules. Right: trainable combiners.

and Parzen classifier. The weighted averaging computes a new score by summing with weights the expert scores. The weights are found so as to minimise the EER on the validation dataset. The Bayes quadratic classifier assumes Gaussian distributed scores. For the Parzen classifier the class conditional score densities are estimated over the validation set using a Gaussian kernel. The joint densities are used to compute the a posteriori probabilities and rule (2) leads directly to the decision. The combiners are trained on the 600 genuine matching scores and the 40 000 impostor matching tests forming the validation set.

4.3.3. Combination results and discussion

The combination results are shown in Figs. 4, 5, 6 for, respectively, two, three and four experts for both probability combination rules and for trainable combiners. Each point in these figures corresponds to an HTER (reported on the y-axis) obtained with the subset of experts reported on the x-axis. The HTER of the best single expert from the subset is also reported in the figures and represented with a circle.

It can be easily shown that for two classes and for two experts the sum, product, maximum, minimum and median rules amount to the same rule. This is why only one combination technique is shown in Fig. 4(left). Remarkably from the figure, all probability rules lead to improved performance. In contrast the trainable combiner may actually decrease the performance in some cases. This happens mainly when the very weak expert HST is in the combination. The probability rules do not seem to be affected by this problem. An ap-

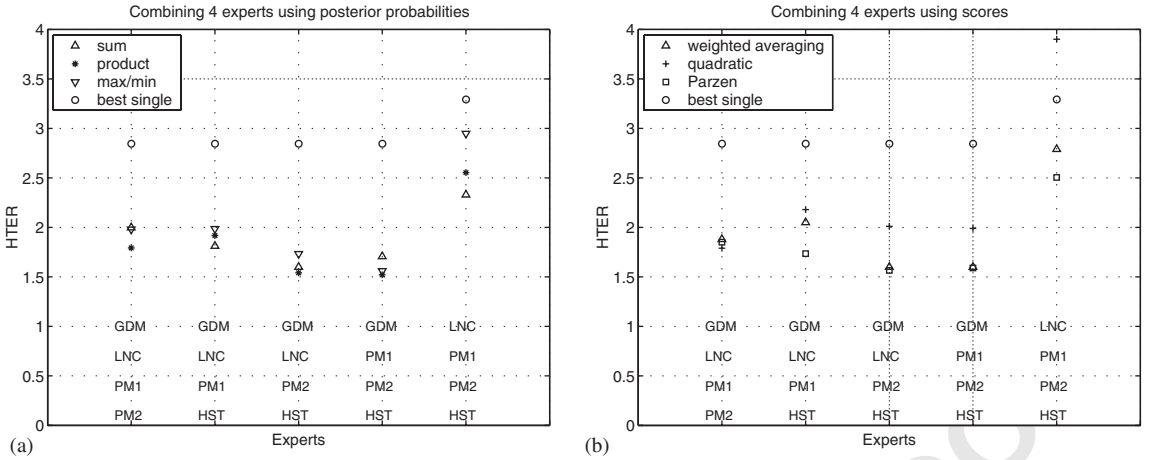

Fig. 6. Combination of four face verification experts. Left: a posteriori probability combination rules. Right: trainable combiners.

Table 2
Expert score statistics

Expert	Ratio r	σ_{a1}	σ_{a2}	σ_{b1}	σ_{b2}	ρ_a	ρ_b
GDM LNC	1.10	3.72	1.23	1.00	1.00	0.77	0.44
GDM PM1	1.19	3.72	0.89	1.00	1.00	0.57	0.32
GDM PM2	1.10	3.72	0.90	1.00	1.00	0.57	0.36
GDM HST	1.95	3.72	0.98	1.00	1.00	0.26	0.04
LNC PM1	1.09	1.23	0.89	1.00	1.00	0.73	0.72
LNC PM2	1.00	1.23	0.90	1.00	1.00	0.82	0.81
LNC HST	1.77	1.23	0.98	1.00	1.00	0.24	0.09
PM1 PM2	0.92	0.89	0.90	1.00	1.00	0.76	0.75
PM1 HST	1.63	0.89	0.98	1.00	1.00	0.14	0.13
PM2 HST	1.77	0.90	0.98	1.00	1.00	0.19	0.10

See text for expert description.

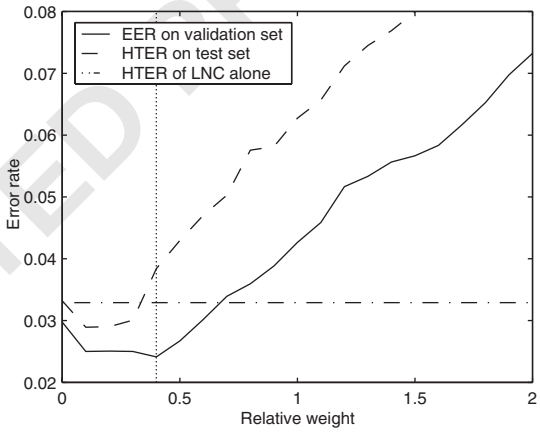

Fig. 7. Validation data EER and test data HTER versus relative weight when combining HST and LNC. The HTER of single expert LNC is also drawn. The vertical line shows the optimal weight found on the validation set. Obviously it is not optimal on the test set.

preciable improvement is brought when PM1 and PM2 are combined although they differ only by the pre-processing technique (cropping and photo-normalisation). When LNC and PM1 are combined, almost no improvement is observed. Note that the relative combination improvement cannot be predicted from the expert correlations shown in Table 2. For example, PM1 and PM2 have approximatively the same correlation as PM1 and LNC. Also LNC and PM2 have similar error rates. However, the first pair of expert leads to an improvement while the second not.

When a third expert is added (see Fig. 5) differences start to appear between the probability rules. Note that we have qualitatively the same performance arrangement as for the Gaussian score simulation: product, max and sum are roughly performing equally well (with a slight superiority exhibited by product and max) and vote/median being slightly worse. As predicted by the theory [8], sum and product perform the same most of the time.

When a fourth expert is added no additional improvement is observed. In fact, a slight decrease may happen in some cases. This suggests that a "classifier selection" step, with reference to feature selection as pointed out in Ref. [16], is required to gain the maximum from the fusion. The vote combination is not shown on Fig. 6 since it is not well defined for an even number of experts. Except for the combined subset involving LNC, PM1, PM2 and HST, all the rules (probability-based and trainable combiners) give similar results for all subsets of experts.

In most cases, the probability and the trainable combinations allow approximatively the same improvement. However, in contrast to trainable combination we never observe a performance degradation when combining probabilities.

One reason could be that estimation in one-dimensional spaces is very effective with respect to training data amount since no dependency between variables has to be learned. Moreover, the probability rules really use the confidence nature of the scores: a weak expert like HST outputs a large fraction of a posteriori probabilities around 0.5. Most of the time, HST does not influence the decision except in the cases when it outputs a high confidence (a probability close to 0 or 1). This acts in favour of the combined decision.

Let us analyse more precisely the reasons for which combination decreases performance. Consider the weighted averaging combiner: the weights are tuned so as to minimise the EER on the validation set. Since an expert gets a weight equal to zero when it increases the EER, the EER of the multiple expert system cannot be higher than the EER of the best expert. A degradation is observed when the combiner training data, i.e. the expert scores on the validation set, do not represent faithfully the scores for new patterns of the test set. This happens when the individual experts are overtrained, even slightly. Fig. 7 shows the validation data EER (solid curve) and test data HTER variation (dashed curve) versus the relative weight when combining experts HST and LNC. It can be seen that the two curves are quite different, which means that the distributions are different for validation and test data. The weight leading to the EER minimum is unfavourable for the HTER on the test data. In fact this difference comes mainly from the HST expert which is overtrained: its false rejection rate on the test data 2% higher than on the validation set. This example confirms the ideas of Duin [2]: the use of one single data set for training the combiner and the experts should be avoided unless special precautions are taken against expert overtraining. The validation set should be divided in two parts, one for training the experts, and one for training the combiner. In our case the experts have been designed without the purpose of future combining, they have been optimised using the whole validation set. Although the posterior probabilities are also affected by overtraining (the estimated probabilities are not correct), our results suggest that the probability combination rules are more robust with respect to overtraining.

We end the discussion by pointing out the minimum HTER obtained on this dataset: 1.30% by combining experts GDM (LDA with gradient direction metric), PM2 (probabilistic matching) and HST (colour histogram-based). This is more than 50% of improvement over GDM, the best single expert. This is however an a posteriori choice of the optimal sub-set of experts as it is determined by taking the minimum HTER on the test set.

5. Conclusions

When combining classifiers, the product rule is optimal in some special cases, for example when classifiers are independent. However, this rule and the other probability combination rules can be used with success in practical situations even if the optimality conditions are not fulfilled. Our goal was to evaluate the effect of correlation between the classifiers on the probability combination efficiency in a two-class problem scenario. With Gaussian distributed scores, it appears that the product, sum and maximum (equivalent to minimum in two-class problems) are relatively robust to correlation. To evaluate the merits of these rules on real data, we studied the problem of identity verification with facial images. Five experts, which output an opinion about the same image and that have been individually optimised, were developed and combined (intramodal combination). The a posteriori probability estimates or confidences given by each expert can be obtained easily from the score given by the expert. The advantage is that this combination technique is modular: no re-training is needed if an expert is added to the multiple expert system. The probability rules perform well on these data, although the experts are very correlated and one them exhibits weak performance. A performance improvement over the best single expert is observed in all cases. For comparison, the experts are combined using trainable combiners. In addition to the fact that they require training, the combiners show less stable results: while they allow an improvement comparable to the probability rules, they sometimes degrade the performance especially when the weak and overtrained expert is in the combining subset.

If the subset of face verification experts to be combined is chosen carefully, and this is still an open problem, the improvement brought by the combination is encouraging: up 50% over the best single expert using the XM2VTS database. This shows multiple intramodal expert combination can increase the robustness of identity verification based on facial images.

Acknowledgements

This work has been carried out within the European project IST BANCA. We thank François Damien for providing the scores for expert HST.

Appendix A.

In this appendix, we show that for experts with Gaussian distributed scores with equal accuracy, equal correlation between all expert pairs for both classes, and equal variance for both classes, i.e. $\rho_a = \rho_b = \rho$, $\sigma_{ai} = \sigma_{bi} = \sigma_i \ \forall i$ and $\Sigma_a = \Sigma_b = \Sigma$, the product combination rule is optimal in a Bayesian sense. We follow the notation from Section 3.2. The classification error for expert i is $e_i = P(\omega_a)P(s_i > \eta|\omega_a) + P(\omega_b)P(s_i < \eta|\omega_b)$. Assuming equal priors and setting the threshold η to $(\mu_{ai} - \mu_{bi})/2$, the error can be written $e_i = \phi((\mu_{ai} - \mu_{bi})/2\sigma_i)$, where $\phi(.)$ is the cumulative distribution

of a Gaussian variable Z with zero mean and unit variance. Since the error is equal for all experts (equal accuracy) we have

$$\frac{\mu_{ai} - \mu_{bi}}{\sigma_i} = K, \tag{A.1}$$

where K is a constant independent of i. We rescale the score s_i so that $\sigma_i = 1\ \forall i$. The optimal fusion strategy of the scores s_i amounts to implementing Eq. (2), which can be written in this case: choose class ω_a if

$$\mathbf{s}^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b) > \frac{1}{2}(\boldsymbol{\mu}_a^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_a - \boldsymbol{\mu}_b^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_b),$$

where $\mathbf{s} = (s_1, s_2, \ldots, s_R)^{\mathrm{T}}$. Because of Eq. (A.1), we have $\boldsymbol{\mu}_a = \boldsymbol{\mu}_b + K\mathbf{v}$ with $\mathbf{v} = (1, 1, \ldots, 1)^{\mathrm{T}}$ and the optimal rule becomes $K\mathbf{s}^{\mathrm{T}}\Sigma^{-1}\mathbf{v} > K(\boldsymbol{\mu}_a + \boldsymbol{\mu}_b)^{\mathrm{T}}\Sigma^{-1}\mathbf{v}$. The covariance Σ has all its diagonal elements equal to 1 and all off-diagonal elements equal to ρ, because all experts share the same correlation. It is easy to see that $\mathbf{v}$ is an eigenvector of this particular matrix, with eigenvalue $C(\rho) = 1 + (R-1)\rho$ where R the number of experts. The vector $\mathbf{v}$ is also an eigenvector of matrix Σ^{-1} with eigenvalue $C(\rho)^{-1}$. Therefore, the optimal rule becomes simply $\mathbf{s}^{\mathrm{T}}\mathbf{v} > (\boldsymbol{\mu}_a + \boldsymbol{\mu}_b)^{\mathrm{T}}\mathbf{v}$. In this particular case, the optimal decision rule does not depend on the correlation between the experts. The product rule is therefore also optimal for any value of correlation.

References

[1] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Recognition Mach. Intell. 20 (3) (1998) 226–239.

[2] R.P.W. Duin, The combining classifier: to train or not to train, in: Proceedings of the International Conference on Pattern Recognition, Quebec, Canada, 2002.

[3] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (1) (2000) 4–37.

[4] K. Ali, M. Pazzani, On the link between error correlation and error reduction in decision trees ensembles, Technical Report 95-38, Department of Information and Computer Science, University of California, Irvine, 1995.

[5] L. Breiman, Bagging predictors, Mach. Learning 24 (1996) 123–140.

[6] L. Kuncheva, A theoretical study on six fusion strategies, IEEE Trans. Pattern Recognition Mach. Intell. 24 (2) (2002) 281–286.

[7] F. Alkoot, J. Kittler, Experimental evaluation of expert fusion strategies, Pattern Recognition Lett. 20 (1999) 1361–1369.

[8] D. Tax, M. van Breukelen, R. Duin, J. Kittler, Combining multiple classifiers by averaging or by multiplying? Pattern Recognition 33 (2000) 1475–1485.

[9] K. Tumer, J. Ghosh, Error correlation and error reduction in ensemble classifiers, Connect. Sci. 8 (3/4) (1996) 385–404.

[10] L. Kuncheva, J.C. Bezdek, R.P.W. Duin, Decision templates for multiple classifier fusion: an experimental comparison, Pattern Recognition 34 (2) (2001) 299–314.

[11] S. Pigeon, L. Vandendorpe, Image-based multimodal face authentication, Signal Process. 69 (1) (1998) 59–79.

[12] B. Duc, E. Bigun, J. Bigun, G. Maitre, S. Fischer, Fusion of audio and video information for multimodal person authentication, Pattern Recognition Lett. 18 (9) (1997) 835–843.

[13] A. Ross, A.K. Jain, J.-Z. Qian, Information fusion in biometrics, in: Proceedings of the International Conference on Audio- and Video-based Person Authentication, 2001, pp. 355–359.

[14] P. Verlinde, G. Chollet, M. Acheroy, Multi-modal identity verification using expert fusion, J. Inform. Fusion 1 (1) (2000) 17–33.

[15] S. Bengio, C. Marcel, S. Marcel, J. Mariéthoz, Confidence measures for multimodal identity verification, Inform. Fusion 3 (4) (2002) 267–276.

[16] S. Prabhakar, A.K. Jain, Decision-level fusion in fingerprint verification, Pattern Recognition 35 (2002) 861–874.

[17] A.K. Jain, S. Prabhakar, S. Chen, Combining multiple matchers for a high security fingerprint verification system, Pattern Recognition Lett. 20 (1999) 1371–1379.

[18] D. Genoud, G. Gravier, F. Bimbot, G. Chollet, Combining methods to improve the phone based speaker verification decision, in: Proceedings of the International Conference on Spoken Language Processing, 1996, pp. 1757–1760.

[19] J. Czyz, J. Kittler, L. Vandendorpe, Combining face verification experts, in: Proceedings of the International Conference on Pattern Recognition, Quebec, Canada, 2002.

[20] A.K. Jain, R. Bolle, R. Pankanti, Biometrics: personal identification in a networked society, Kluwer Academic Publishers, Dordrecht, 1999.

[21] J. Kittler, J. Matas, K. Jonsson, M.R. Sanchez, Fusion of audio and video information for multimodal person authentication, Pattern Recognition Lett. 18 (9) (1997) 845–852.

[22] R. Duin, D. Tax, Classifier conditional posterior probabilities, in: Advances in Pattern Recognition, Lecture Notes in Computer Sciences, vol. 1451, Springer, Berlin, 1998, pp. 661–619.

[23] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Recognition Mach. Intell. 19 (7) (1997) 711–720.

[24] R. Duda, P. Hart, D. Stork, Pattern Classification, Wiley, New York, 2000.

[25] J. Kittler, Y.P. Li, J. Matas, On matching scores for LDA-based face verification, in: Proceedings of British Machine Vision Conference, 2000.

[26] B. Moghaddam, W. Wahid, A. Pentland, Beyond eigenfaces: probabilistic matching for face recognition, in: Proceedings of the International Conference on Automatic Face and Gesture Recognition, 1998.

[27] K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre, XM2VTSDB: the extended M2VTS database, in: Proceedings of the International Conference on Audio- and Video-based Person Authentication, 1999, pp. 72–77.

[28] J. Luettin, G. Maitre, Evaluation protocol for the extended M2VTS database, IDIAP, available at http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/face-avbpa2001/protocol.ps (1998).

About the Author—JACEK CZYZ received the Degree in Physical Engineering from the Université Libre de Bruxelles, Belgium in July 1997 and the Ph.D. degree from the Université catholique de Louvain (UCL), Belgium, in December 2003. Before joining the UCL Communications Laboratory in March 2000, he worked at Creo Products, a manufacturer of pre-press equipment with headquarters in Vancouver, Canada. His research interests include Pattern Recognition, Computer Vision and Image Processing.

About the Author—JOSEF KITTLER graduated from the University of Cambridge in Electrical Engineering in l971 where he also obtained his Ph.D. in Pattern Recognition in 1974 and the Sc.D. degree in 1991. He joined the Department of Electronic and Electrical Engineering of Surrey University in 1986 where he was a Professor, in charge of the Centre for Vision, Speech and Signal Processing. He has worked on various theoretical aspects of Pattern Recognition and on many applications including automatic inspection, ECG diagnosis, remote sensing, robotics, speech recognition, and document processing. His current research interests include Pattern Recognition, Image Processing and Computer Vision. He has co-authored a book with the title "Pattern Recognition: a statistical approach" published by Prentice-Hall. He has published more than 300 papers. He is a member of the Editorial Boards of Pattern Recognition Journal, Image and Vision Computing, Pattern Recognition Letters, Pattern Recognition and Artificial Intelligence, and.Machine Vision and Applications.

About the Author—LUC VANDENDORPE received the Electrical Engineering degree (summa cum laude) and the Ph.D degree from the Université catholique de Louvain (UCL) Louvain-la-Neuve, Belgium in 1985 and 1991 respectively. Since 1985, L. Van-dendorpe is with the Communications and Remote Sensing Laboratory of UCL where he first worked in the field of bit-rate reduction techniques for video coding. From October 1992 to August 1997, IT Vandendorpe was Senior Research Associate of the Belgian NSF at UCL, and invited assistant professor. Presently he is Professor. He is mainly interested in filtering, source/channel coding, multirate digital signal processing, digital communication systems. He is associate editor of the IEEE Trans. on Wireless Communications and a member of the Signal Processing Committee for Communications. He was an editor of the IEEE Trans. on Communications for Synchronisation and Equalization between 2000 and 2002.