Université catholique de Louvain
Faculté des Sciences Appliquées
LABORATOIRE DE TÉLÉCOMMUNICATIONS
ET
TÉLÉDÉTECTION

B - 1348 Louvain-la-Neuve Belgique

Decision Fusion in Identity Verification using Facial Images

Jacek Czyz

Thèse présentée en vue de l'obtention du grade de
Docteur en Sciences Appliquées

Examination Committee:

Luc VANDENDORPE (UCL/TELE) - *Supervisor*
Jean-Philippe THIRAN (EPFL, Switzeland)
Michel VERLEISEN (UCL/DICE)
Pierre DUPONT (UCL/INGI)
Josef KITTLER (University of Surrey, UK)
Jean-Didier LEGAT (UCL/DICE) - *President*

December 2003

Acknowledgements

I wish to express my most sincere gratitude to my supervisor Prof. Luc Vandendorpe for granting me the opportunity of carrying out this study. I thank him for his professionalism, encouragement and experienced guidance.

My deepest thanks to Prof. Josef Kittler from the Centre for Vision, Speech and Signal Processing at the University of Surrey (UK), for his valuable advice and fruitful discussions, and for giving me the opportunity of spending some time in his laboratory.

I am equally grateful to Mirek Hamouz and Alex Kostin from the University of Surrey for providing me with automatically determined eye coordinates. Also, I would like to thank Samy Bengio and Christine Marcel from the Institut Dalle Molle d'Intelligence Artificielle Perceptive (Switzerland) for providing speaker verification scores.

Many thanks to the examination committee for the many interesting comments that resulted in the considerable improvement of the manuscript.

I also would like to thank my fellow researchers and colleagues of the TELE lab for the pleasant working atmosphere.

Lastly and most importantly, the love, understanding and encouragement of Virginie were indispensable ingredients to achieve this work.

The research reported in this thesis has been funded by the European Union IST project BANCA.

Contents

List of abbreviations

EER	Equal Error Rate
FAR	False Acceptance Rate
FRR	False Rejection Rate
HTER	Half Total Error Rate
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
PM	Probabilistic Matching
GMM	Gaussian Mixture Model
SVM	Support Vector Machine
MLP	Multi-Layer Perceptron
NFL	Nearest Feature Line
EUCL	Euclidean Distance
NMM	Nearest Mean Metric
NOC	Normalised Correlation
GDM	Gradient Direction Metric
LFCC	Linear Frequency Cepstral Coefficients
XM2VTS	Extended M2VTS database
ROC	Receiver Operating Characteristic
HIST	Histogram Equalisation photo-metric normalisation
ZMUV	Zero Mean and Unit Variance photo-metric normalisation
SC	Smart Card
HAI	Hypotheses-driven Affine Invariant face localisation

Introduction

> *There is a general consensus in behavioural biology today that the main factor driving the evolution of large brains in primates was the computational load of social processing. We evolved large brains not because of the reproductive advantages of being able to become computer scientists, but rather because large brains could better support our various social conspiracies. [...]. Facial, postural, and gestural expression and identification were crucial elements in those social computations and our practice of them today is but a recent gloss on ancient biological foundations.*
>
> J. Daugman [24]

At every second, our brain analyses, performs segmentation, infers 3D information, classifies, recognises and interprets a huge amount of data captured by our eyes and ears about the scene we are looking at. When trying to automate this seemingly effortless process so that it could be performed by a machine, it appears that we hardly understand how this is done and even less how to program a computer to do it. Less complex animals such as flies or bees have developed, after billions of generations, a visual system that performs probably several times better than any machine vision algorithm available today.

Among the tasks of our brain, the recognition and verification of the identity of our peers is performed hundreds of times a day with very little error. This task is achieved using visual information, for example by looking at the person's face or at the gait, or auditory information for example by listening to a person's voice, or both. This shows that from the information flow that we perceive, our brain is able to extract relevant features which are (i) discriminative (ii) compact, so as to be memorised (iii) retrieved in a very fast way to be matched against new incoming data to recognise. If we could determine these features and the accompanying recognition process, the identity verification or recognition task could be automated, opening the doors to dozens of applications in data protection, security or entertainment. Unfortunately, although we know that identify-

ing a person using his/her face image is possible as we are doing it pretty well every day, the problem of *automatic* identity verification is not solved.

Automatic verification of personal identity using facial images is the central topic of the thesis. This problem can be stated as follows. Given two face images, we must determine automatically whether they are two images of the same person or of different persons. The field of automatic personal identity verification and recognition using human specific characteristics is called *biometrics*. A biometric system is devoted to verify or recognise (see next section for clarification of these terms) the identity of a person. During the *enrolment phase*, images of the user's face are taken and used to create a *template*. During the *production phase*, images of user's face are recorded again when the user's identity has to be verified. A decision on the person's identity is taken on the basis of the comparison between the template and the new images.

Due to many factors the problem of verification using facial images is difficult: variability of facial appearance, sensitivity to noise, template aging, limited training data for the recognition algorithm, etc. In fact even the face localisation step, which consists of detecting and locating the person's face in an image, is not completely solved yet in the general case of completely unconstrained background.

However, one must keep in mind that the solution of a given vision problem such as automatic identity verification depends on the constraints imposed on the problem. Specialised hardware can be implemented to ease the vision algorithm task. For instance localisation can be solved perfectly if the background is uniform and the user is constrained to present his face straight to the camera. Many other specialised hardware setups are conceivable. However, they usually limit the field of application of the system (due to high cost or dimension problem) and/or constraint the user.

In this work we have tried to limit the constraints imposed on the user as much as possible, so that the resulting automatic verification algorithm is very general and does not rely on specific enrolment or production prerequisites. This means that we suppose the enrolment is done using a single ordinary camera in controlled conditions whereas production is done in unconstrained conditions.

Apart from the hardware approach to simplify the vision algorithm task, we can overcome some of the difficulties by *combining* different informative sources for the classification/recognition task. Intuitively, the more features, the better the recognition. This is true if the features ex-

tracted from the informative sources are independent. Unfortunately in most cases the features are correlated. Therefore strategies must be devised in order to take advantage in the most efficient way of the different sources of information.

In Pattern Recognition [30, 26, 37], high level logical statements (like class membership, or logical description, etc.) are inferred from the raw sensor measurements after some hardware and software processing. This process is depicted on Figure 1. To generate different sources of informa-

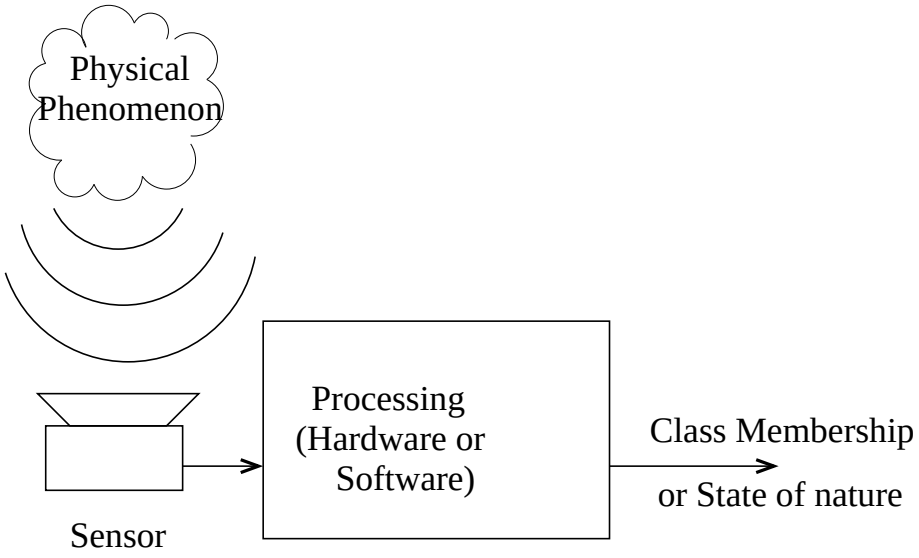


Figure 1: *General Pattern Recognition process. A sensor measures a physical phenomenon and outputs a possibly multidimensional signal. The signal is processed and high level information about the physical phenomenon, usually class membership, is extracted.*

tion, the general paradigm can be enriched as follows.

- Use different sensors for capturing different physical phenomena (but related to the same logical class / state of nature). E.g. multimodal biometrics, in the identity verification context, where the authentication is based on different biometric characteristics, such as the voice, fingerprints and face image of the subject.
- Use the same sensor which measures sequentially over time several instances of the same physical phenomenon.

- Use different sensors which measure the same phenomenon, often called multi-sensor fusion. E.g. effects of 3D Pose variability can be reduced if during enrolment several cameras, positioned at different orientations with respect to the user, record images of the user under different views. The test image, taken this time using only one camera, is then compared to all enrolment images of the subject under different poses. Another possibility is to use different types of sensors: e.g. an infra-red camera combined with a visible light camera.

- Use the same sensor which measures the same physical phenomenon but the signal is processed using different processing algorithms. Hence the outputs of the different algorithms may not be the same. Strictly speaking, this does not bring a new source of information, but we may consider that the processing algorithm performed on the raw signal of figure 1 is removing information to generate the class membership (it is a mapping from a high dimensional space to a finite set). Therefore the outputs of several different processing algorithms working on the same signal may contain different information.

In all cases, the information gathered from different sources must be conciliated or combined or fused at some stage of the pattern recognition process presented in figure 1. Fusion can be achieved at various stages among which we may distinguish:

1. Fusion at the *measurement level*: although conceivable, this fusion level is seldom used in practice.

2. Fusion at the *feature level*: extracted features from the various information sources can be concatenated and used as input of an overall classifier.

3. Fusion at the *confidence level* also known as soft fusion: the classifier outputs a number reflecting its confidence in the decision.

4. Fusion at the decision level also known as hard fusion, where logical information e.g. class membership, is combined.

The central idea of the thesis is to propose strategies on how to combine the different information sources in order to improve the verification accuracy. We will study principally the fusion at the confidence level and

analyse the improvement it can bring for applications in identity verification. We believe that confidence level fusion is a good compromise between dimensionality and information loss. Feature level fusion leads to high dimensional feature vectors for which the construction of a classifier may be problematic. At the decision level however, almost all information except for the class label has been thrown away: no confidence about the chosen class label can be deduced in this case.

The specific application area has been defined within the framework of the European Union project IST BANCA (Biometric Access Control for Networking and e-Commerce Applications). The objectives of this project were to develop and implement a complete secured system with enhanced authentication and access control schemes for applications over the Internet, such as tele-working and Web-banking services. This required the development of robust multimodal biometric verification schemes based on speech and facial images. Our efforts are concentrated on the face verification problem, i.e. verify the identity of a person using images of his/her face. We studied how information fusion, in the various instances presented above, could improve the verification performance.

More specifically, we have implemented and thoroughly optimised a number of face verification algorithms. We studied their individual properties (such as how their accuracy depends on algorithm parameters, image size, or sensitivity to mis-registrations). We have also studied how to combine the outputs of the different algorithms in order to reduce the verification error rates. As all algorithms operate on the same modality (face modality in this case), this fusion type is referred to as *intramodal fusion*. This is an example of fusion of different processing algorithms working on the same signal.

Another fusion aspect considered in this thesis is the fusion of confidences obtained sequentially on several video frames of the same person's face. Instead of making the desicion based on a single image, several images are taken of the person whose identity is to be verified.

The *Multimodal fusion* of face and speaker verification algorithms illustrates the case of different sensors capturing different physical phenomena related to the same state of nature. In this case, the speech and face of the same subject are recorded and processed by different algorithms which output separate opinions. These two opinions are then conciliated at the fusion stage. Note that the speaker verification algorithm has been developed at IDIAP, (Institut Dalle Molle d'Intelligence Artificielle Perceptive, Martigny, Switzerland) within the framework of the BANCA project.

Special care has been taken to test the proposed face verification and fusion algorithms. To avoid optimistic bias in the performance evaluation, realistic face image databases reflecting real application scenarios have been used. Also, concerning methodology, the experiments were performed according to rigorous testing protocols.

Although the terminology used in this work is specific to the identity verification application, we believe that many of the results can be extended to any two-class pattern recognition problem. In fact, the fusion problem is in line with a growing area of research known as *multi-classifier systems* in pattern recognition [63, 104] and *ensemble methods* in machine learning [27]. In this respect, the thesis tackles two different points of view: the practical problem of identity and the general problem of classifier combination.

Verification, identification, recognition and authentication

To avoid misunderstandings, hereafter we highlight the differences between verification, which is of interest in this work and the other related terms which are identification, recognition and authentication.

In the biometrics context, identification or recognition is the task in which, given a biometric signal, the identity of the person related to the signal is retrieved from a database of N subjects. Identification can be done in a closed world scenario, where we are sure that the biometric signal under inspection is related to a subject of the database. Identification in this case is a one-to-N matching process. In the open world scenario, the biometric signal under inspection is not necessarily related to one of the N subjects. The identification algorithm must in this case output a *reject* option. The open scenario is obviously much more difficult as the class of the "rest of the world" must be somehow modelled. In this context, recognition is a synonym of identification.

Regarding verification or authentication, a biometric signal and a hypothetical identity are given. The task is to verify or ensure that the signal is related to given identity. The verification case, which is mainly devoted to security applications, is always in a open world scenario. In fact, in the verification task the role of impostors, i.e. people unknown to the system, is essential since the verification algorithm is supposed to reject them. In verification or authentication, the subject is very often supposed to be cooperative as he/she must also introduce a hypothetical identity.

Authentication is a synonym of verification.

In this thesis we are looking at the verification problem exclusively. This means that the decision problem that we are confronted with is a binary hypothesis test. Indeed, it is to be verified whether the claimed identity of the subject under test be confirmed or rejected.

Structure of the thesis

The thesis is organised as follows. Chapter 1 introduces the reader to the problem of biometric identity verification. The specific terminology and concepts of biometric identity verification are embodied in the firmly founded decision theory. Optimal decision rules under several criteria are derived. In particular, the minimax decision rule is identified as the most frequently used in practice as it does not depend on class prior probabilities.

Chapter 2 materialises the general principles introduced in the preceding chapter to the identity verification using face images. A brief overview of the related literature is given. Furthermore, a unifying framework for subspace face verification is presented and discussed.

Chapter 3 introduces the two face databases on which experiments have been performed in this thesis. Experimental setup and testing protocol are described in detail. Limitations of the databases are given and motivations for choosing these datasets are explained.

Chapter 4 presents the experimental results obtained using the algorithms described in Chapter 2. The influence of the algorithm parameters, as well as image size and matching schemes are studied in detail. The results are analysed and allow to point out current limitations of the verification algorithms.

In Chapter 5, the problem of decision fusion is introduced in the framework of identity verification. Furhtermore, two techniques of classifier combination are discussed, namely the *a posteriori* probability combination and the trainable combination in the score space. The *a posteriori* probability combination rules which follow from the classifier combination theory, are studied using simulated and real data. The aim is to determine the influence of classifier correlation on the accuracy of these rules. The chapter also presents a face verification system that combines several verification algorithms. The combined system is shown to be more accurate than any of the single verification algorithms.

Chapter 6 shows how the use of an independent source of informa-

tion can reduce the error rates and/or the size of the feature vector. The property is illustrated on the multimodal fusion of a face and speaker verification algorithm. A verification system combining the face and speech of a person is presented an studied. It is shown that the multimodal fusion reduces significantly the error rates and the user template size.

Chapter 7 studies the fusion of multiple instances of the physical phenomenon. Instead of using only one face image, we show how simple techniques improve the decision accuracy when several video frames acquired during a user access are available. The problem is first tackled analytically under simple hypotheses. Afterwards we test our simple combination rules on real face data and show that, provided the rule is chosen carefully, the multiple instances of the biometric data help to reduce the number of errors committed by the verification algorithm.

Overall conclusions of the thesis are given in the last chapter.

Original contributions

The thesis is written with the aim of creating unity and coherence, thus original contributions are not easily distinguishable. For this reason, we explicit the original contributions of this work in the following:

1. A unifying framework for subspace face recognition and verification is presented (Chapter 1 and 2).

2. A multi-resolution study of a leading subspace method for face verification is presented [23] (see Chapter 2).

3. The link between the multi-classifier combination theory and what is done in practice in the case of fusion in biometrics has been established [22] (see Chapter 5).

4. A simple method for combining *a posteriori* probabilities given by several classifiers has been developed and tested [22] (see Chapter 5).

5. A face verification algorithm that is based on the combination of three different face experts has been developed [21]. It performed honourably at the recent face verification contest organised at the Audio- and Video-based Biometric Person Authentication conference (AVBPA 2003). (see Chapter 5).

6. A multimodal identity verification system based on speech and face images is proposed [20]. Also it is shown that multimodality helps in reducing template size for applications with limited storage space (see Chapter 6).

7. The sequential integration of expert outputs or scores has been formalised and tested for face verification. The merits of different integration rules versus the score probability distribution have emerged (see Chapter 7).

Chapter 1

Biometric identity verification

In our electronically inter-connected society, reliable and user-friendly personal identity verification or recognition is indispensable in many sectors of our life. To verify the identity of a person can be done in several different ways [51]:

- using person's possession. If the person possesses a physical object (like keys, cards, etc.), identity can be verified (in fact assumed) based on this object possession.
- using person's knowledge. Identity can be verified by checking if the person knows a correct password or a PIN (Personal Identification Number) code.
- using person's physiological or behavioural characteristics. Some traits particular to the person like fingerprints, hand geometry and face, or some behavioural characteristics like the voice or the signature. These identity verification methods are known as biometrics.

The two first identity verification techniques, although easy to implement have some drawbacks: keys can be lost or worse, stolen. Passwords can be forgotten, guessed by unauthorised people. Many people write down the PIN code of their ATM bank cards on the card itself [91]. This of course removes the usefulness of the passwords. Also, in our daily life, the number of passwords that people have to remember in order to take advantage of all services is getting larger (using the same password for all services is very insecure as when the password is disclosed, the adversary have access to all passwords).

The third identity verification technique, which is of interest in this thesis, concerns a tangible component of identity: a physiological or behavioural characteristic. The biometric characteristics cannot be misplaced or forgotten. Although biometrics do not solve all security problems (see for example [43]) and is not appropriate for all applications, they represent a powerful mean of identity verification in some applications when correctly implemented. Moreover, when combined with the other verification techniques (keys and passwords) a certain level of security can be reached. Our main problem is to devise algorithms that can *automatically* verify identity by using biometrics.

1.1 Authentication in a decision theory framework

In this section we state the problem of identity in the framework of decision theory. Several decision rules, depending on the error criterion chosen, are presented. We discuss conditions to obtain reliable verification error rate estimation.

The general problem of identity verification or authentication can be stated as follows. An identity claim is accompanied by a measurement of physiological or behavioral features emanating from the person making the claim. Given the measurement or signal and the associated identity, the task is to automatically decide that the signal is genuine (then access is granted, the claim is *accepted*) or the identity claim is false in which case the claim is rejected. The signal or features are delivered by the biometric sensor and we suppose that it is in digital form, i.e. it has already been converted from analogic to digital by sampling and quantising the analogic signal. In this work we look at the raw signal as a n-dimensional real vector $\mathbf{x} \in \mathbb{R}^n$. Because the signal can be an image, or a audio or video recording, the dimensionality of the signal space n is usually very large, which is fundamentally the reason that makes the problem difficult as we will see in this work. An instance of signal $\mathbf{x}$ is usually called *sample*.

This is a typical pattern recognition or classification problem with two classes [37, 30, 26]: $\mathbf{x}$ must be classified as belonging to class ω_a in the genuine case or to class ω_b in the impostor case. An error occurs when the signal is wrongly classified. Therefore problem amounts to finding a mapping from $\mathbb{R}^n$ to $\{\omega_a, \omega_b\}$ or a *decision rule* that minimises an error criterion. Depending on the criterion chosen, different forms of decision rule can be derived as shown in the following.

1.1.1 Minimum error decision rule

If we want to minimise the probability of error $P(error)$, the Bayes decision rule [30] described in this paragraph should be used. Having observed a certain biometric signal or sample $\mathbf{x}$, if we decide ω_a the probability of making the wrong decision, i.e. the probability of error $P(error|\mathbf{x})$ is simply the probability that $\mathbf{x}$ belongs in fact to class ω_b, that is $P(\omega_b|\mathbf{x})$. Conversely, if we decide ω_b the probability of error is $P(\omega_a|\mathbf{x})$. Clearly, we can minimise the probability of error for a given $\mathbf{x}$ by deciding ω_a if $P(\omega_a|\mathbf{x}) > P(\omega_b|\mathbf{x})$ and ω_b otherwise. The probability of error can be written

$$P(error) = \int P(error|\mathbf{x})p(\mathbf{x})d\mathbf{x},$$

where $p(\mathbf{x})$ is the probability density of $\mathbf{x}$. Because our decision rule lead to $P(error|\mathbf{x})$ minimum for every $\mathbf{x}$, the integral is also minimum. The rule which assigns the class that has maximum *a posteriori* probability, i.e.

$$P(\omega_a|\mathbf{x}) \underset{\omega_a}{\overset{\omega_b}{\lessgtr}} P(\omega_b|\mathbf{x}),$$

is the rule minimising the probability of error. It is called the Bayes decision rule. Using Bayes formula, *a posteriori* probabilities can be expressed using the conditional probability densities $p(\mathbf{x}|\omega_c)$

$$P(\omega_c|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_c)P(\omega_c)}{p(\mathbf{x})},$$

where $c \in \{a, b\}$, the decision rule can be re-written

$$\frac{p(\mathbf{x}|\omega_a)}{p(\mathbf{x}|\omega_b)} \lessgtr \frac{P(\omega_b)}{P(\omega_a)}. \tag{1.1}$$

The right side of these inequalities does not depend on $\mathbf{x}$ and reflect our prior knowledge about the occurrence of class ω_a and class ω_b. The ratio of the two densities (left side of the inequalities) is called the likelihood ratio $l(\mathbf{x})$ [116]. The decision rule can therefore be written

$$l(\mathbf{x}) \lessgtr \eta,$$

where η is a threshold. These last inequalities define a decision boundary that separates the space of $\mathbf{x}$ into a genuine domain Ω_a for class ω_a and an impostor domain Ω_b for class ω_b. When a particular $\mathbf{x}$ is measured, it is labelled as genuine or impostor whether it falls in Ω_a or Ω_b.

We can at this stage distinguish two broad categories of models that can be invoked to solve the classification problem.

- Generative Models: in this category, one intends to estimate the conditional probability densities $p(\mathbf{x}|\omega_c)$ ($c \in \{a, b\}$), and perform classification using Equation (1.1).
- Discriminative Models: the boundary function or equivalently the likelihood ratio $l(\mathbf{x})$ is estimated directly.

Although optimal in theory, in the sense that it minimises the probability of error, it is very difficult to implement Equation (1.1) in real world problems. The reason is that the probability laws generating the samples $\mathbf{x}$ are almost always unknown. Moreover, even when probability laws are assumed, for example to be Gaussian, due to the high dimension of the space, a large number of samples is required to obtain meaningful estimates of first and second order moments. This is why discriminative techniques such as Neural Network or Support Vector Machines (SVM) have become very popular in classification tasks. However, generative techniques that are successful exist. For example, Gaussian Mixture Modelling (GMM) is a generative technique used in speech recognition and speaker recognition/authentication. See Chapter 6 for a description of speaker authentication using GMM.

In practice we thus look for a mapping $s(\mathbf{x})$ from $\mathbb{R}^n$ onto $\mathbb{R}$, that approximates as well as possible the unknown likelihood ratio $l(\mathbf{x})$. The function $s(\mathbf{x})$ is called the *score*. See next section for more details on the score function. The decision rule becomes

$$s(\mathbf{x}) \underset{\omega_a}{\overset{\omega_b}{\lessgtr}} \eta. \tag{1.2}$$

1.1.2 Error estimation

In order to assess the performance of a identity verification system, two types of error have to be analysed.

- An identity claim is rejected although it is genuine. This type of error is usually referred to as *false rejection*.
- An identity claim is accepted although it is false. This type of error is usually referred to as *false acceptance*.

The number of false rejections and false acceptances should be as low as possible, ideally zero. These two error rates have to be distinguished because the cost incurred by false acceptances may be very different from

the false rejection cost. In the space $\mathbf{x}$ the probability of false acceptance error e_A and false rejection error e_R can be written

$$e_R = \int_{\Omega_b} p(\mathbf{x}|\omega_a)d\mathbf{x}, \quad \text{for the probability of false rejection and} \tag{1.3}$$

$$e_A = \int_{\Omega_a} p(\mathbf{x}|\omega_b)d\mathbf{x}, \quad \text{for the probability of false acceptance.} \tag{1.4}$$

This amounts to calculate the probability of observing impostors in the genuine domain Ω_a and genuine samples in the impostor domain Ω_b. It is much easier to express these errors in the one-dimensional score space

$$e_R(\eta) = \int_{\eta}^{\infty} p(s(\mathbf{x})|\omega_a)ds, \qquad e_A(\eta) = \int_{-\infty}^{\eta} p(s(\mathbf{x})|\omega_b)ds, \tag{1.5}$$

and the total error probability which is minimised by Bayes rule becomes

$$P(error) = P(\omega_a)e_A + P(\omega_b)e_R. \tag{1.6}$$

Of course we do not have access in practice to the conditional score probability densities $p(s(\mathbf{x})|\omega_c)$ ($c \in \{a, b\}$) so that e_A and e_R must be estimated from data. To do so, a set of N_a genuine and N_b impostor scores must be gathered using an independent data set, usually by matching biometric features x_i coming from the same person and matching feature coming from different persons. The last integrals are then simply estimated by the False Acceptance Rate (FAR) R_A and False Rejection Rate (FRR) R_R

$$R_A(\eta) = \frac{1}{N_b}\sum_{i=1}^{N_b} u(\eta - s(\mathbf{x}_i)) \qquad x_i \in \omega_b, \tag{1.7}$$

$$R_R(\eta) = \frac{1}{N_a}\sum_{i=1}^{N_a} u(s(\mathbf{x}_i) - \eta) \qquad x_i \in \omega_a, , \tag{1.8}$$

where $u(.)$ is the unit step function. It must be emphasised that there is a trade-off between the R_A and R_R: by varying the threshold η one can make one of the rates arbitrarily small while inevitably increasing the other. Imagine that we set the threshold at a number larger than any score. In this case rule (1.2) is satisfied for any genuine access but also for any impostor access, i.e. R_R = 0% and R_A = 100%. For this reason the performance characterisation of an identity verification system is usually achieved by reporting the so-called Receiver Operating Characteristic

(ROC) curve (term originating from the field of detection [116]). The ROC curve is a plot of 1-FRR versus the FAR parametrised by the threshold η. A given threshold defines an operating point on the ROC curve. This point depends on the target application, whether the cost incurred by false acceptances or false rejections is higher, in other words whether low FAR or low FRR is more critical. The rate at the operating point leading to equal FAR and FRR is called equal error rate (EER) and its associated threshold η_{EER}. Because it is a single number for a given identity verification system and dataset, the EER allows a coarse assessment of the identity verification system. Note that EER is determined *a posteriori*, i.e. the threshold is set after the N_a and N_b matches are generated. In a real world system, the threshold must of course be set *before* the system is in use. It is therefore more realistic to set the threshold *a priori* and determine the FAR and FRR at this threshold. This results in rates that may occur when using the identity verification system in a real world application.

Using Equation (1.5), the score densities $p(\mathbf{x}|\omega_i)$ with $i \in \{a, b\}$, can be obtained by deriving e_R and e_A with respect to η. Therefore, we can use

$$\hat{p}(s(\mathbf{x})|\omega_a) = -\frac{d}{d\eta} R_R(\eta), \tag{1.9}$$

and

$$\hat{p}(s(\mathbf{x})|\omega_b) = \frac{d}{d\eta} R_A(\eta) \tag{1.10}$$

as estimates of the score probability densities.

1.1.3 Decision rule for minimising the risk

The loss or cost incurred by making the wrong decision may be different whether an impostor is accepted or a genuine sample is rejected. For this reason we introduce a *risk* associated with a wrong decision. We choose the risk to be proportional to the probability of error. Given $\mathbf{x}$, the risk $r_a(x)$ associated with making decision ω_a is

$$r_a(\mathbf{x}) = c_{ab} P(\omega_b|\mathbf{x}), \tag{1.11}$$

where c_{ab} is proportionality constant or the cost incurred by accepting an impostor. Likewise the risk $r_b(x)$ associated with making decision ω_b is

$$r_b(\mathbf{x}) = c_{ba} P(\omega_a|\mathbf{x}), \tag{1.12}$$

where c_{ba} is the cost incurred by rejecting a genuine match. The decision for minimum risk becomes, choose the class that has minimum risk, which can be written

$$l(\mathbf{x}) \underset{\omega_a}{\overset{\omega_b}{\lessgtr}} \frac{c_{ab}}{c_{ba}} \frac{P(\omega_b)}{P(\omega_a)}. \tag{1.13}$$

As we see in this last equation, the introduction of cost changes effectively only the value of the threshold but not the functional form of the likelihood ratio $l(\mathbf{x})$. This means that, changing prior knowledge or decision costs c_{ab} and c_{ab} has an effect only on the threshold; the score function remains the same.

1.1.4 Minimax rule

The rules (1.1) and (1.13) involve explicitely the prior probabilities $P(\omega_a)$ and $P(\omega_b)$, which reflect our prior knowledge of impostor or genuine sample occurance. Unfortunately, these probabilities are unknown. It is very difficult to evaluate the prior probability of observing an impostor. To solve this problem, an approach consists of designing the decision rule so that the worst overall risk for any value of the prior is as small as possible.

The overall risk R is the risk cumulated for all $\mathbf{x}$, which can be written

$$R = \int_{\Omega_a} r_a(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} + \int_{\Omega_b} r_b(\mathbf{x}) p(\mathbf{x}) d\mathbf{x},$$

because Ω_a is the domain where decision ω_a is taken and therefore risk r_a is taken. The same applies for domain Ω_b. Using the definition of the risks in (1.11) and (1.12) and the fact that $P(\omega_b) = 1 - P(\omega_a)$ we have

$$R = c_{ab} \int_{\Omega_a} p(\mathbf{x}|\omega_b) d\mathbf{x} + P(\omega_a) \left(c_{ba} \int_{\Omega_b} p(\mathbf{x}|\omega_b) d\mathbf{x} - c_{ab} \int_{\Omega_a} p(\mathbf{x}|\omega_b) d\mathbf{x} \right).$$

From this equation it appears that the risk depends linearly on the prior probability $P(\omega_a)$ once the decision regions Ω_a and Ω_b are fixed. The idea is to choose the decision region so that the factor multiplying $P(\omega_a)$ vanishes. In this case the risk does not depend on the priors. The vanishing condition is

$$c_{ba} \int_{\Omega_b} p(\mathbf{x}|\omega_a) d\mathbf{x} = c_{ba} \int_{\Omega_a} p(\mathbf{x}|\omega_b) d\mathbf{x}, \tag{1.14}$$

which is simply

$$c_{ba} e_R = c_{ab} e_A,$$

or equal error rate (EER) condition if $c_{ab} = c_{ba}$. The decision rule becomes

$$l(\mathbf{x}) \underset{\omega_a}{\overset{\omega_b}{\lessgtr}} \eta,$$

but this time η is chosen so that condition (1.14) is satisfied. Note that the minimax rule does not minimise the error as rule (1.1) neither the risk as rule (1.13) but we are in a situation where the risk is constant for any value of the prior probabilities $P(\omega_a)$ and $P(\omega_b)$. This releases us from a particular choice of prior probabilities.

1.1.5 Neyman-Pearson rule

Yet another formulation of the decision genuine-impostor is possible. As two types of errors can be committed e_A and e_R (false acceptance and false rejection), the Neyman-Pearson rule is the one which minimises one of the error probabilities, while the second error is kept equal to a constant e_0. For example we can fix the impostor acceptance to a maximum allowed by the application and minimise the genuine rejection. This constrained minimisation is solved using Lagrange multipliers. The problem becomes: choose the decision domains Ω_a and Ω_b so as to minimise the quantity L

$$L = e_R + \lambda(e_A - e_0),$$

where λ is a Lagrange multiplier. Using the definition of e_R and e_A and the fact that

$$\int_{\Omega_b} p(\mathbf{x}|\omega_a)d\mathbf{x} = 1 - \int_{\Omega_a} p(\mathbf{x}|\omega_a)d\mathbf{x},$$

we have

$$L = (1 - \lambda e_0) + \int_{\Omega_a} (\lambda p(\mathbf{x}|\omega_b) - p(\mathbf{x}|\omega_a))d\mathbf{x}.$$

In order to minimise L, Ω_a must hold all $\mathbf{x}$'s for which the integrand is negative, as those $\mathbf{x}$'s decrease the value of L. This amounts to require that $\mathbf{x} \in \omega_a$ if $\lambda p(\mathbf{x}|\omega_b) - p(\mathbf{x}|\omega_a) < 0$ and $\mathbf{x} \in \omega_b$ otherwise. This can be written as

$$\frac{p(\mathbf{x}|\omega_a)}{p(\mathbf{x}|\omega_b)} \underset{\omega_a}{\overset{\omega_b}{\lessgtr}} \lambda. \tag{1.15}$$

The resulting decision rule is identical to the Bayes rule (1.1) minimising the error probability except that the threshold is now given by λ. However, the preceding discussion shows that the likelihood ratio test not only

minimises the probability of error but also the error for one class while maintaining the error for the other class constant.

Note that Ω_a and Ω_b are almost completely determined by the likelihood ratio $l(\mathbf{x})$, that is the ratio of the class probability densities. Only one parameter (the threshold) has to be set to completely determine the decision domains. In the Neyman-Person rule, this parameter is λ and it is determined by the constraint

$$e_0 = \int_{\Omega_a} p(\mathbf{x}|\omega_b) d\mathbf{x}.$$

1.2 Verification: a two class problem?

In the preceding sections, we have presented verification as a two class problem, the impostor and genuine classes. Two class modelling is very convenient because of the simplicity of the decision rule: a function of $\mathbf{x}$ to be compared to a threshold. In fact, because each person has an identity, identification is inherently a multiclass problem. Consider the case of faces: a class C_p corresponds to the set of all possible face images of a person p under any lighting conditions, any expression, etc. The face space is therefore populated with clusters corresponding to different identities C_i as depicted in Figure 1.1. A natural way to model the problem is to assign a class to each person, which is clearly a multi-class approach. The problem of recognition is, having measured $\mathbf{x}$, determine the corresponding identity, that is the corresponding class. In the verification problem (see Figure 1.2), a presumed identity is given, say C_p, and we must decide if $\mathbf{x}$ belongs to C_p or not. In this case the genuine class ω_a is C_p while the impostor case ω_b are all the other persons $\{C_1, C_2, ...\}$. Clearly verification and recognition are related and one could extend one case to the other easily.

An important point is that, in the verification case a different decision rule (hence a different score function $s(\mathbf{x})$) must be applied when a different person's identity has to be verified. An approach where the score depends on the complete probability density of the person $p(\mathbf{x}|\mathbf{x} \in C_p)$ could be considered. Persons that have their faces more variable (due to make up, more expression changes, etc.) would have a more relaxed decision boundary. Unfortunately, this approach requires large amounts of training data of the same person to model faithfully the probability density. In high dimensional spaces, to estimate the second order statistics (variances and correlations forming the covariance matrix) requires in general more

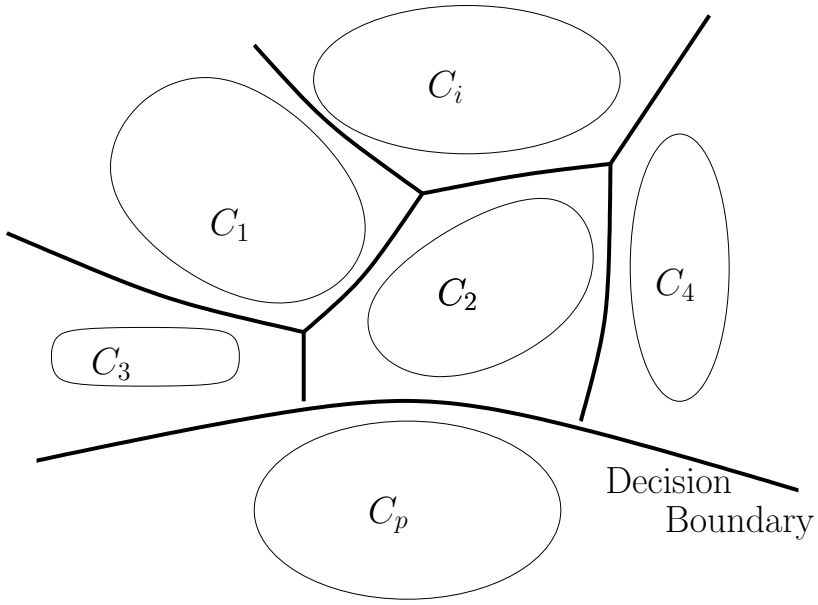


Figure 1.1: *Two-dimensional representation of the face space: each ellipse represents a class, that is, the set of all possible face images of the same person. The recognition problem is multi-class: the decision rule should separate all classes.*

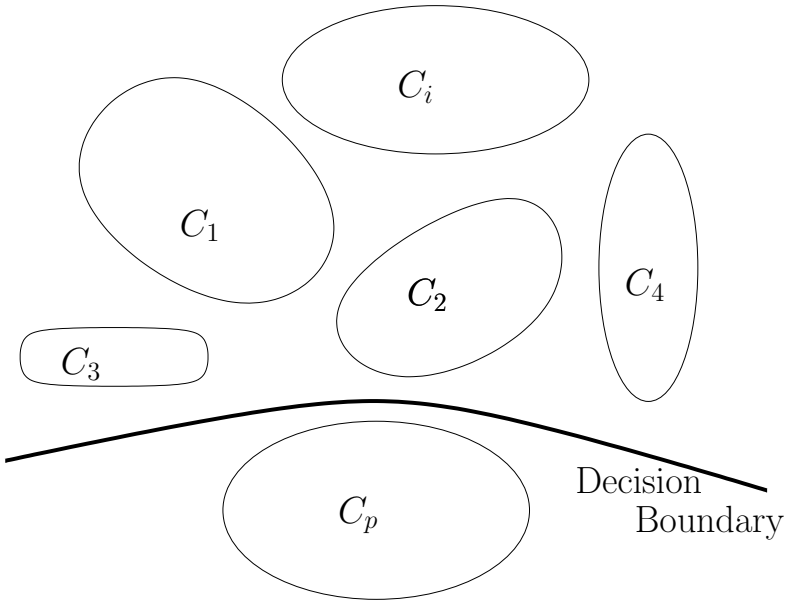


Figure 1.2: *Two-dimensional representation of the face space in the verification case. Here, the problem has two classes: the genuine class which is the set of faces of person p or class C_p, and the impostor "super-class" which is all other classes* $\{C_1, C_2, ...\}$

data than available. For this reason, in practice the decision rule and score function depend in general only on the first order moment of the probability density of the person $p(\mathbf{x}|\mathbf{x} \in C_p)$, the mean $\bar{\mathbf{x}}_p$.

Because of the sparsity of the training data, it is advantageous to assume that the class conditional probability densities $p(\mathbf{x}|\mathbf{x} \in C_i)$ differ only by the means $\bar{\mathbf{x}}_i$. It amounts to suppose that for $\mathbf{x} \in C_p$ the measured signal can be written $\mathbf{x} = \bar{\mathbf{x}}_p + \boldsymbol{\epsilon}$ where $p(\boldsymbol{\epsilon})$ is the same for all classes. A pooled estimation can then be used to estimate the common covariance matrix.

In any case, the score function is specific to the person p making the claim and we should write in this case $s_p(\mathbf{x})$. For example, s can be a distance measure between $\mathbf{x}$ and the mean $\boldsymbol{\mu}_p$

$$s_p(\mathbf{x}) = d(\mathbf{x}, \boldsymbol{\mu}_p).$$

The mean can be seen as a *template* of the biometric signal which the newly measured signal $\mathbf{x}$ is compared to.

Also, strictly speaking the genuine and impostor score distributions $p(s_p(\mathbf{x})|\omega_b)$ and $p(s_p(\mathbf{x})|\omega_a)$ depend on the person p. The error rates FAR and FRR should be computed for each person. This is of course unrealistic, therefore the FRR is computed by mixing all the genuine scores available and the FAR by mixing all the impostor scores. The resulting genuine and impostor score distributions are in fact mixtures of the distributions for individuals.

1.3 Confidence intervals for error rates

An important issue concerns the confidence we have in the verification system accuracy. When the verification system is tested, the FAR and FRR are estimated on a database of persons as described in Chapter 3. Each person recorded several biometric signals which are used to compute both genuine and impostor scores using Equations (1.7) and (1.8). As databases are inevitably of limited sizes, in order to gain confidence in the verification performance of the system once it is deployed in the real world (the population of impostors is virtually infinite), it is important to set confidence intervals around the estimated error rates. This will help to predict the verification performance for a real world application from the performance on a collected database.

Precaution must be taken as for the meaning of confidence intervals. The confidence interval does not represent *a priori* estimates of perfor-

mance in different environments and for different populations. They represent the inherent uncertainty in tests results owing to small sample size.

Two techniques to determine confidence limits for an estimation are presented here, one is parametric while the second is a bootstrap-based method.

1.3.1 Parametric confidence interval

By definition, the false acceptance rate R_A is the ratio of the number of impostor acceptances n_b to the total number of impostor accesses N_b that we have in an experiment:

$$R_A = \frac{n_b}{N_b}.$$

Likewise, the false rejection rate R_R is the ratio of the number of genuine rejections n_a to the total number of genuine accesses N_a

$$R_R = \frac{n_a}{N_a}.$$

Recall that e_A and e_R are respectively the true (unknown) probability of false acceptance and true probability of false rejection that we are trying to estimate for the system being tested. As the discussion is identical for the FAR and the FRR, we consider the case of the FAR only. As impostors have a probability e_A of being accepted, and there is a total of N_b trials, the rate R_A is a random variable following a binomial law $B(N_b, e_A)$ (each matching is a Bernoulli trial [90] with probability e_A), thus the probability that the FAR gets the value $\frac{k}{N_b}$ is

$$\mathrm{Prob}(R_A = \frac{k}{N_b}) = \left(\begin{array}{c} N_b \\ k \end{array} \right) e_A^k (1 - e_A)^{N_b - k}.$$

Using properties of the binomial law, the expectation of R_A is

$$\mathrm{E}[R_A] = e_A,$$

and the variance is

$$\mathrm{Var}[R_A] = \frac{e_A(1 - e_A)}{N_b},$$

which shows that the FAR R_A converges to e_A when N_b tends to infinity.

A confidence interval is an interval [a_α,b_α] (depending on R_A) in which e_A is included with probability α, i.e.

$$\mathrm{Prob}(a_\alpha < e_A < b_\alpha) = \alpha.$$

When N_b is large the binomial distribution may be approximated by a Gaussian distribution thus

$$\frac{R_A - e_A}{\sqrt{(1-e_A)e_A/N_b}} \sim \mathcal{N}(0,1),$$

and the confidence limits can be found to be [101]

$$\frac{1}{1+\epsilon^2/N_b}\left(R_A + \epsilon^2/2N_b \pm \epsilon\sqrt{R_A(1-R_A)/N_b + \epsilon/4N_b^2}\right), \qquad (1.16)$$

where ϵ satisfies

$$\phi(\epsilon) = \int_{-\infty}^{\epsilon} \mathcal{N}(x;0,1)dx = \frac{\alpha+1}{2}.$$

Usually α is taken to be 0.95 for a 95% confidence interval in which case we have $\epsilon = 1.96$.

Recall that $R_A(\eta)$ can be seen as an approximation of the impostor cumulative score distribution while $1 - R_R(\eta)$ an approximation of the genuine cumulative score distribution. A confidence interval around $R_A(\eta)$ and $R_R(\eta)$ can be obtained for any value of η, it corresponds to confidence intervals for the cumulative score distributions. Confidence intervals for the score densities can be deduced from there.

Assumptions

Note that this approach to find the confidence intervals does not assume any distribution for the impostor scores. This is important because this distribution is not known in practice.

The assumption that must be satisfied is firstly the *independence* between the impostor trials [101]. This means that if several biometric instances (face image, speech, fingerprint) of the same person are used to generate different impostor trials, the resulting scores distribution will reflect only the variation related to conditions of acquisition but not the variation in identity. For example, in face based biometrics, the scores of two matches corresponding to a client face image and two images of the same impostor are not independent. When estimating the false acceptance error, the data set should rather contain one image per impostor rather than several images of a small group of impostors. The same applies for genuine scores: a better estimate of false rejection probability e_R is obtained with genuine matches of many different people (ideally one match per person) rather than many matches a small group of people.

Secondly, it must be emphasised that each access will represent a Bernoulli trial only if e_A can be considered as constant across the impostor population. Doddington *et al.* have shown that for speaker verification, some people, called "wolves" by the authors, are very effective in impostoring, while some others, called "lambs" are very easily imposted [28]. Also, e_A depends on environmental conditions. That is, if the production data (i.e. data that is encountered by the system during its life-cycle [109]) differ from the test data because of the environment, the confidence intervals estimated may be not valid for the production data.

1.3.2 Bootstrap confidence interval

Bootstrap methods are very effective to determine confidence intervals in very general cases without hypothesis on distributions [98]. In [12], it is introduced to estimate confidence intervals in the field of biometrics.

The bootstrap principle is described in the following. We are interested in an estimation of a parameter θ from the observed samples $\mathbf{X} = \{X_1, X_2, ..., X_m\}$. Thus a statistic $T(\mathbf{X})$ will be used to estimate θ.

The bootstrap principle prescribes sampling, *with* replacement, the set $\mathbf{X}$ a large number B times, amounting to the sets $\mathbf{X}_i^*$, for $i = 1, ..., B$ and calculating many estimates $T_i^* = T_i^*(\mathbf{X})$, for $i = 1, ..., B$. Then one can compute a, say, 90% confidence interval by counting the bottom 5% and top 5% of the estimates T_i^* and subtracting these estimates from the set $\{T_i^*\}$. The leftover set determines the confidence interval. In other words, in the bootstrap method, we treat the observed samples as if they exactly represented the whole population.

We obtain a bootstrap confidence interval for $\hat{\theta} = T(\mathbf{X})$ as follows.

1. Calculate $\hat{\theta}$ from the set $\mathbf{X}$.
2. *Resampling*. Draw a random bootstrap sample $\mathbf{X}^* = \{X_1^*, X_2^*, ..., X_m^*\}$ by sampling $\mathbf{X}$ uniformly with replacement.
3. *Bootstrap estimate* Calculate $\hat{\theta}_i^* = T_i^*(\mathbf{X}^*)$ using the bootstrap sample $\mathbf{X}$.
4. *Repetition.* Repeat steps 2 and 3 a large number B times.
5. *Approximation of the distribution of* $\hat{\theta}$ Sort the bootstrap estimates into increasing order to obtain $\hat{\theta}_1^* < \hat{\theta}_2^* < ... < \hat{\theta}_B^*$.
6. *Confidence interval.* The desired α100% bootstrap confidence interval is $[\hat{\theta}_{q_1}^*, \hat{\theta}_{q_2}^*]$ where $q_1 = B(1-\alpha)/2$ and $q_2 = B - q_1 + 1$.

We are interested in estimating the false acceptance (or false rejection) probability e_A using the observed score set. In our case the estimator T is

$$R_A = \frac{n_b}{N_b},$$

where n_b is the number of successful impostor accesses and N_b the total number of accesses. To obtain the confidence interval, we resample a large number of times the set of accesses and we compute the FAR R_A on each resampled set. The obtained distribution defines the confidence interval. Figure 1.3 shows an example of distribution obtained using $B = 1000$ resampled sets of speaker impostor accesses (Banca database using P sub-configuration, see Chapter 3). In this case the FAR is 4.49% and 95% confidence interval is found to be [3.04,6.25].

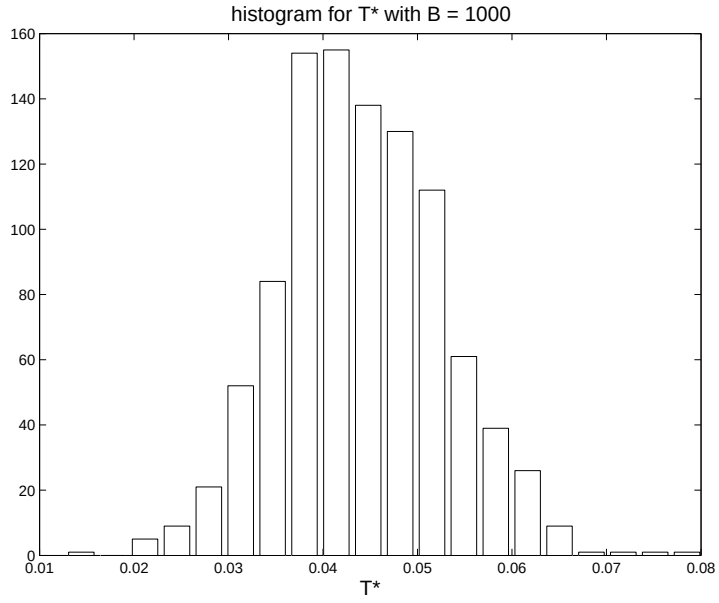


Figure 1.3: *Bootstrap estimates distribution of the false acceptance rate for speaker verification*

Chapter 2

Face verification

2.1 Introduction

While the preceding chapter was devoted to general issues of identity verification, in this chapter we focus on one particular biometric modality: frontal human faces. The large interest in identity verification using face images can be gauged by the organisation of biennial international conferences (AVBPA: Audio- and Video-based Biometric Person Authentication conference and FG: Face and Gesture Recognition international conference) devoted to this problem and the existence of several commercially available face recognition/verification products. A lot more applications and products will certainly arise once the technology becomes more reliable.

A face is a surface of a three dimensional solid having partially deformable parts. The images obtained from it depend upon pose, illumination condition, expression. In these respects, face recognition is paradigmatic for machine intelligence and pattern recognition [24]. This is why a considerable amount of research effort is devoted to automatic face recognition and verification.

The face modality is very important for real world applications because it has many advantages over other biometric modalities. The most important advantage is that face modality is very well accepted by the users [51]. Also, people use mainly face to identify each others in everyday life. The modality is non-intrusive and contactless, in the sense that there is no need to interfere with a sensor as for other modalities. Verification can be done without disturbing too much the user: the user only needs to be present in front of a camera. Another advantage of face modal-

ity is the low sensor costs: a cheap camera like a web-cam provides easily 25 frames per second at a resolution of 350×720 pixels.

This absence of constraints brings a lot of variability in the corresponding biometric signal. Face image pixels vary *drastically* under variable illumination (see Figure 2.1). A slight change in 3D pose transforms the corresponding pixel map. In fact, image variations due to illumination and viewing direction are almost always larger than variations due to change in identity [1]. Face images suffer from various occlusions due to glasses, hair, beards, make-up, etc. This variability or the "noise" on the image pixels, must be somehow modelled and removed before comparing the face image to its corresponding template. When the background is uncontrolled, the problem of automatically finding the face in the image is an indispensable pre-processing task and represents on its own a whole research area [123]. As we will see, the main problem of face verification in unconstrained environment comes more from high false rejection rates, that is when the template and the genuine test images are too "far" apart.

An important issue when designing a biometric system concerns protection against counterfeiting or "spoofing". The biometric characteristics used for verifying identity can be recorded (e.g. a picture for face modality) and supplied fraudulently to the verification system. The system therefore has to make sure that the input signal is alive, a task which may not be straightforward and which we will not look into in this work.

Face is intrinsically a low security modality. Monozygotic twins, i.e. twins who share the same DNA code, can fool people who do not know them (people familiar with the twins have received a big amount of data and are therefore much more trained to distinguish the twins). It is unlikely that automatic face verification, which must be quite lax to cope with variability of face images, will ever be able to detect the very subtle differences that exist between twins.

The concepts presented in the preceding chapter will be materialised for this modality. Given a face image $\mathbf{x} \in \mathbb{R}^n$ the central task consists of devising a score function $s(\mathbf{x})$ which leads to the best verification performance, that is the lowest False Acceptance (FAR) and False Rejection Rates (FRR). Up to now, the problem of face recognition has received most of the attention from the research community. As face verification and recognition have a lot in common, face verification algorithms are very often inspired by recognition algorithms and adapted to take into account of the particularities of the verification problem.

We briefly review the main algorithms that have been developed dur-

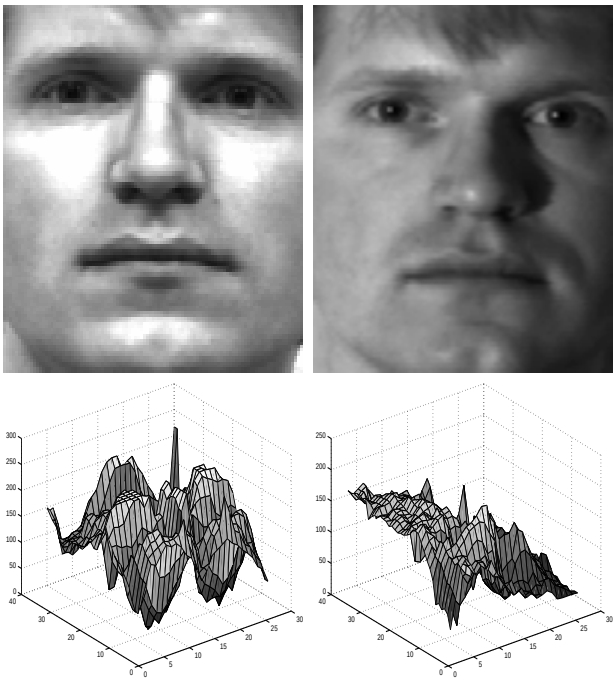


Figure 2.1: *Two face images of the same person under different illumination conditions, and the corresponding pixel intensity profiles. While the similarity between the two images is obvious for a human observer, the two image profiles are rather different. A similarity measure based on the Euclidean distance in the image space (or equivalently Mean Square Error between the two images) would fail to classify correctly the two images.*

ing the last two decades to tackle the challenging problem of face recognition or verification. We then describe in details the algorithms that we have used in this thesis.

2.2 Background

Automatic face processing is a quite vast area spreading over several engineering domains such as pattern recognition, computer vision and signal processing. It includes face recognition and verification, face detection or localisation, face tracking, facial expression analysis, facial pose estimation, etc. A very wide variety of approaches and models have been pro-

posed. We therefore describe in this section only the most representative approaches to face recognition and verification.

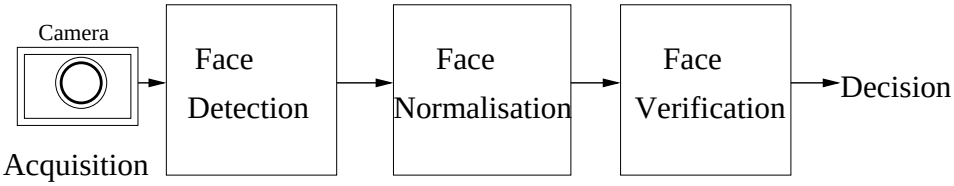


Figure 2.2: *The different sub-tasks of face verification leading to the final decision*

Very often in engineering, a complex problem is divided into several smaller problems which are more tractable. Traditional approaches divide the global task of face verification/recognition in modules. Each module performs a sub-task and passes the result to the next module in a sequential manner as depicted in Figure 2.2. Firstly the raw digital image is acquired by the camera. The second sub-task is the detection and localisation of the face sub-image in the raw image. The third sub-task is the normalisation of the face sub-image. The normalisation can be both geometric and photometric. It is meant to prepare the data for the next sub-task. The last task is the verification (or recognition) which results in the "accept" or "reject" decision.

In this thesis, the efforts are concentrated on the verification aspects of the problem. From the sequential process in Figure 2.2, it appears that the verification sub-task relies on the success of the localisation technique. In other words, if the localisation sub-task fails to find the face, the whole verification process fails. As a result, the performance of the verification system in its entirety is influenced by the localisation module performance and by the verification module performance. To gain understanding about our problem, it is wise to study the verification and localisation modules independently. The two problems can be decoupled by feeding perfectly located faces to the verification module. This way we can concentrate on the limitations proper to verification. However we do not neglect the localisation process as the verification system in its entirety rely on it. In order to present results that can be obtained with a completely automatic system, a face localisation method provided by University of Surrey [45, 44] has been used in our experiments. In the next paragraph we describe existing methods in face verification/recognition. Although not central to the thesis, some face localisation methods are very briefly reviewed after-

wards in order to give a context to the reader.

2.2.1 Face verification or recognition

Approaches to face recognition/verification are usually divided into two families: local feature based methods and holistic methods.

- The local feature based methods rely on a feature set small in comparison to the number of pixels in the image. Features can be defined manually [57, 15, 19], for example by choosing distances and angles between eye corners, ends of mouths, nostrils, etc. Features may correspond to Gabor wavelet coefficients (called *jets*) at predefined points in the image [67] or at *fiducial points* corresponding to specific facial landmarks [120]. In the latter approaches, elastic bunch graph matching is used to compare new images with templates. The use of Gabor wavelet coefficients is motivated by their characteristics of spatial locality and orientation selectivity and their biological relevance [25]. Alternatively, the Gabor jets may be replaced by multiscale morphological operations (dilation and erosion) on the nodes of a sparse grid placed over the face image [64]. The local feature method family provides intrinsically high flexibility in handling non-rigid facial features, and 3D pose variation. On the other hand, the success of these approaches relies highly on the accuracy of the feature detection process which requires good quality images (high resolution, good illumination) and may lack robustness.

- In the holistic or appearance based methods, the face image pixels are used directly as features. Images are seen as elements of a vector space called the image space. Template matching techniques and extensions [5, 15, 79] work directly in the image space. The image to be tested is compared to the template using an appropriate similarity measure. Template matching is quite effective when the images to compare have the same illumination and same pose. It fails on images such as in Figure 2.1.

  Because of the high dimensionality of the image space and the limited dataset sizes, a transform from the image space onto a low dimensional space is often performed with the hope that a decision boundary is easier to find in the low dimensional space. These very successful methods are called subspace methods. Kirby and Sirovitch proposed to use the Karhunen-Loève Transform or Princi-

pal Component Analysis (PCA) to represent faces in a low dimensional space [59]. The technique has been applied to face recognition by Turk and Pentland in their famous *eigenface* method [115]. Extensions of the eigenface method include view-based and modular eigenspaces [93] where multi-view face images are handled by constructing a PCA subspace for different views. Probabilistic visual learning [85, 86] originates from the eigenface method (this method is described in detail below). Linear Discriminant Analysis (LDA) has been proposed as a dimensionality reduction tool [33, 112, 7, 126] and many variants [73].

Other subspace methods include Local Feature Analysis (LFA) [92], evolutionary pursuit [75] where the optimal subspace basis is found using a genetic algorithm, the mixture of subspaces [36] where a mixture of factor analysers is fitted to model the face space, and the matching pursuit filter approach [94]. Independent Component Analysis (ICA), which searches for axes such that extracted features are statistically independent, has also been applied to define the face subspace [6, 46].

Active shape and appearance models constitute an important family of methods [17, 69]. In order to better represent faces, the shape and gray level appearance information is modelled separately. Both shape and appearance dimensionality is reduced by Principal Component Analysis. Thanks to the shape information, the appearance information is deformed using a thin-plate spline technique [34] to the mean shape before applying PCA. The method allows to model not only identity but also 3D pose, expression and even aging effects [68]. The difficulty is to fit the active shape model to a new image automatically as discussed in [18] and [17].

A classifier like the Support Vector Machine (SVM) which is trained so as to maximise the margin (which is the sum of distances to the closest positive and negative points, see [16] for an introduction to SVM) does not suffer from high dimensionality and can be applied directly in the image space for solving the verification problem [55, 107]. Since SVM's are intrinsically devoted to two-class problems, they can be use for verification directly. Note that in this method the decision function must be learned for each person separately.

2.2.2 Detecting faces in images

Numerous methods have been proposed to detect faces in a single image or in a video sequence. See [123] for a recent literature review. We give hereafter the general principles of face detection. Since the face can be positioned anywhere and at any scale in the image, the whole image is scanned using a sliding window at different resolutions. The extracted sub-image content is classified as face or non-face. Basically any classifier can be used for this purpose: Osuna *et al.* used an SVM classifier [89], multiple neural networks are used in [106]. In [111] the face and non-face classes are respectively represented by six multivariate Gaussian clusters. The pattern in the sub-image is labeled as face or non-face based on the distance to the clusters. The face/non-face classifier approach requires a dataset of non-face patterns for training the classifier. This kind of training set may be difficult to acquire. Alternatively, generative methods build a probabilistic model for face class. The face class is modelled by a mixture of Gaussians in [85], and a mixture of factor analysers in [122]. The sub-image contains a face if the density value exceeds a given threshold.

Instead of scanning the whole image using a rectangular window, facial features can be detected directly in the image. In [58] skin color is used to locate possible position of the face sub-image. In [45], local feature detectors are used to detect facial features such as corners of eyes, nostrils, etc. in the input image. Feature configurations that correspond possibly to a face are transformed into a canonical face image. In the canonical face image, the natural and geometric variability of faces is highly reduced because faces are registered in the same co-ordinates system. This allows to train the face/non-face classifier more accurately and therefore increases correct detection rates.

2.3 The face model

Any image can be re-arranged into a real vector, for example by lexicographic ordering of the image pixel values. Note that the lexicographic reordering changes the two-dimensional nature of image signals into one dimensional sequences. By doing so, the topological relationships between neighbour pixels are disregarded[1].

Seen as sequences, we may consider the set of all possible images with

[1] See for example [102] for how the topological relationships can be taken into account as prior knowledge in the case of face representation and recognition.

size $N \times M$ to span $\mathbb{R}^n$ where $n = MN$. Due to typical image sizes, the dimensionality n of this image space is very large, of the order of magnitude of 10^4. Images that may be encountered in reality (often referred to as natural images) represent an extremely small portion of the image space. In fact the image space is almost entirely populated with meaningless noise. The images that represent a plausible human face are again a tiny portion of the set of all possible natural images with size $N \times M$. Although vast, the set of human face images is likely to be governed by a limited number of parameters in comparison to the 10^4 numbers required to depict a face.

This face set may be seen as a low dimensional manifold (a manifold is a topological space which is locally Euclidean i.e., around every point, there is a neighbourhood which is topologically the same as the open unit ball in $\mathbb{R}^n$) in the image space.

From the discussion above, a way to model face images with a small number of parameters is to approximate the face manifold by a linear subspace of $\mathbb{R}^n$. Principal Component Analysis (PCA), also called Karhunen-Loève transform [37], has been shown to be adequate to find a basis of the linear subspace [59, 115] as it captures the part of the signal which has maximal energy.

2.3.1 Principal Component Analysis (PCA)

PCA is a standard technique in data analysis which is used for dimensionality reduction or equivalently for feature extraction for signal representation. Given a set of l data vectors $\mathbf{x}_i \in \mathbb{R}^n$ which are instances of a random vector $\mathbf{x}$, PCA looks for $m \leq n$ orthonormal vectors $\{\boldsymbol{\phi}_j\} \in \mathbb{R}^n$ which form an orthonormal basis of the subspace that captures maximal variance of the $\mathbf{x}_i$'s. It can be shown [37] that the $\{\boldsymbol{\phi}_j\}$'s are the eigenvectors of the sample covariance matrix Σ of the $\mathbf{x}_i$'s

$$\Sigma = \frac{1}{l} \sum_{i=1}^{L} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T, \tag{2.1}$$

where $\bar{\mathbf{x}}$ is the sample mean of the $\mathbf{x}_i$'s. Let λ_j be the eigenvalue associated with the eigenvector $\boldsymbol{\phi}_j$ we have

$$\Sigma \boldsymbol{\phi}_j = \lambda_j \boldsymbol{\phi}_j.$$

The approximation or *reconstruction* $\hat{\mathbf{x}}$ of the random vector $\mathbf{x}$ is the component of $\mathbf{x}$ that lies in the subspace expressed in $\mathbb{R}^n$ i.e.

$$\hat{\mathbf{x}} = \bar{\mathbf{x}} + \sum_{j=1}^{m} (y_j - \bar{y}_j)\boldsymbol{\phi}_j,$$

where

$$y_i = \boldsymbol{\phi}_i^T \mathbf{x}$$

is called the ith principal component of $\mathbf{x}$ and $\bar{y}_i$ is the ith principal component of the mean $\bar{\mathbf{x}}$, i.e.

$$\bar{y}_i = \boldsymbol{\phi}_i^T \bar{\mathbf{x}}.$$

By defining the $n \times m$ matrix Φ whose columns are m eigenvectors $\boldsymbol{\phi}_j$ of Σ we can write the principal component vector $\mathbf{y} \in \mathbb{R}^m$

$$\mathbf{y} = \Phi^T \mathbf{x},$$

and because the columns of Φ are orthonormal, we have

$$\hat{\mathbf{x}} = \bar{\mathbf{x}} + \Phi(\mathbf{y} - \bar{\mathbf{y}}),$$

where $\bar{\mathbf{y}} = \Phi^T \bar{\mathbf{x}}$. A very nice property of PCA is that the mean square error ϵ^2 between $\mathbf{x}$ and its reconstruction, which is

$$\epsilon^2 = E[\|\mathbf{x} - \hat{\mathbf{x}}\|^2],$$

can be written as

$$\epsilon^2 = \sum_{j=m+1}^{n} \tilde{\lambda}_j,$$

where $\tilde{\lambda}_j$ are the eigenvalues of the true (unknown) covariance matrix generating the $\mathbf{x}_i$'s. This last equation suggests that the mean square error between $\mathbf{x}_i$ and its reconstruction $\hat{\mathbf{x}}_i$ is minimised if the subspace basis contains the m eigenvectors $\boldsymbol{\phi}_j$ with the m highest eigenvalues. One must keep in mind that the transform Φ is the optimal *linear* transform under a mean square error criterion. This means that there may exist other non-linear transforms which leads to a better representation of the signal.

The PCA decorrelates the $\mathbf{x}_i$'s as the covariance expressed in the principal subspace

$$\Sigma_{\mathbf{y}} = \Phi^T \Sigma \Phi \tag{2.2}$$

is a $m \times m$ diagonal matrix whose diagonal elements are the λ_j's.

PCA can be interpreted geometrically as follows. Imagine a cloud of data points distributed on a hyperplane of dimension m in $\mathbb{R}^n$. The principal subspace defined by the m covariance eigenvectors will be the hyperplane which contains all the data points. If the data points are not exactly distributed on a hyperplane, the principal subspace will approximate a best fitting hyperplane under a mean square error criterion.

Application of PCA to face images

When applying PCA to face images, a numerical problem appears immediatly. The sample covariance matrix estimated using Equation (2.1) contains n^2 elements. Since n is typically $O(10^4)$, this matrix is beyond computation capabilities. Furthermore, it is easy to see that the rank of Σ is equal to $l-1$ or less. Therefore since the number of images is inferior to image space dimensionality for almost all face databases, Σ is singular in practice and has only $l-1$ eigenvectors. The other eigenvectors are not defined as they are associated to null eigenvalues.

Fortunately the $l-1$ eigenvectors can be computed without computing the complete covariance matrix [37]. Let U be a $n \times l$ matrix whose columns are the l centred data vectors $\mathbf{x}_i^C = \mathbf{x}_i - \bar{\mathbf{x}}$, i.e.

$$U = [\mathbf{x}_1^C \mathbf{x}_2^C ... \mathbf{x}_l^C].$$

Equation (2.1) becomes

$$\Sigma = \frac{1}{l} U U^T.$$

Let us call V and Λ respectively the matrices of eigenvectors and eigenvalues of the $l \times l$ matrix $\frac{1}{l} U^T U$. We have

$$\frac{1}{l} U^T U V = V \Lambda.$$

Multiplying both sides by U we get

$$\frac{1}{l} U U^T (UV) = (UV) \Lambda,$$

which shows that the l vectors in the matrix (UV) are the eigenvectors of Σ. Note that these vectors must be normalised in order to be orthonormal. The matrix of principal components is after normalisation

$$\Phi = \frac{1}{\sqrt{m}} U V \Lambda^{-1/2}.$$

Figure 2.3 shows a few eigenvectors or *eigenfaces* [115] computed using the $l = 600$ training images of the extended M2VTS database (see Chapter 3). The maximal number of eigenvectors is therefore 599. The image resolution is 64×64 leading to an image space dimensionality of 4096. The technique described above has been implemented to compute the principal subspace. In the figure, the eigenvectors are ranked by decreasing corresponding eigenvalue. Note that the eigenvectors corresponding to eigenvalues with high magnitude (ϕ_1 to ϕ_5) contain mainly global and low-frequency features. As the index increases, the eigenvectors contain gradually more and more high frequency information. The 200 last eigenvectors (ϕ_{400} to ϕ_{599}) are very noisy. Note also that the training faces have been carefully registered and cropped by pinpointing the centres of the eyes and applying an affine transformation that puts the two eyes respectively on the same positions for all images.

One may question the relevance of the estimation described above. If n is the covariance matrix size, the number of data samples needed to smooth out statistical fluctuations and obtain good estimates is usually several times the number n [30]. Clearly it is not the case here. Furthermore a set of l points in $\mathbb{R}^n$ is always embedded in a (at most) $(l-1)$-dimensional hyperplane. Therefore a $(l-1)$-dimensional PCA reconstruction will be exact (i.e. $\hat{\mathbf{x}}_i = \mathbf{x}_i$) for the l points used to determine the principal subspace (training points), and we hope that this subspace is valid for new points generated by the same probability law. In fact the principal subspace will be correctly determined by PCA if the intrinsic dimensionality of the signal $\mathbf{x}$ is several times smaller than the number of samples l. In this case, the true size of the covariance is equal to the intrinsic dimensionality m and good estimates of the principal components can be obtained (if the data distribution is similar to a Gaussian) if l is several times m.

Experiments show that the technique achieves good generalisation as a face image that was not included in the training set is correctly reconstructed, except if it contains variation not present in the training set. Figure 2.4 shows several reconstructions of a face image using a variable number of principal components. The reconstructed face image comes from a subject not present in the XM2VTS training set. From the figure, we see that 50 eigenvectors approximate very coarsely the identity of the subject. With 200 eigenvectors the information related to identity is present, so that the person is identifiable for a human observer but fine appearance details, such as the exact shape of mouth, or nose are missing. The

complete set of eigenvectors (599) allows an almost perfect reconstruction.

In contrast, Figure 2.5 shows the same image that has been rotated and scaled. The reconstructions are much less satisfactory in this case. Even the full reconstruction looks very noisy and does not capture the subject identity. The reason is that the training images are all registered in the same way. A rotated or scale face image does not belong to the same subspace because rotation deviates from the original subspace.

Determining the dimensionality of the principal subspace

The dimensionality n of the original signal $\mathbf{x}$ can be reduced to m using PCA. However, in general the choice of the dimensionality m of the principal is not easy. One can employ an energy criterion [112] : m is selected such that the retained eigenvalues contain 95% of the total energy, i.e.

$$\sum_{j=1}^{m} \lambda_j = 0.95 \sum_{j=1}^{\min(l,n)} \lambda_j.$$

In [126], the selection is based on the characteristics of the eigenvectors: above rank 300, the authors notice that the eigenvectors encode mainly noise and should be discarded for face classification tasks. In [87], based on empirical results on large face databases, it is suggested to choose m to be 40% of the total number of eigenvectors. Bayesian model selection is used in [83] to pick the correct dimensionality. A pragmatic approach consists in choosing m so as to minimise the classification error rate on a data set.

Whatever the technique for determining m, the intrinsic dimensionality of the face space depends on the preprocessing of the images. For example, if the faces are normalised by non-rigid transformation so that shape information is uniform like in [17], the face subspace is low-dimensional as only gray level information variation has to be explained by the model. In contrast, poorely registered faces lead to a higher dimensional face subspace as shown on Figure 2.5 (see also section on misregistration).

The part of the face that is present in the face image also plays a role in the value of m. In particular, the presence of hair in the images, very variable by nature, increases significantly m. In [40] the part of face useful for classification is determined by studying rank statistics between pixels in the image. Regions where pixels do not exhibit strong relationship are

disregarded. The face region found in this way, which excludes forehead and background, corresponds quite closely to the region used in our work.

2.3.2 Extensions of PCA and other face models

The face manifold, a subset of the image space, has been extensively studied in the framework of face detection. This is because membership of an image to the class of faces can be decided easily by computing a distance between the image and the face manifold. In its basic form, the face manifold is approximated by the principal linear subspace obtained by applying PCA on a training set of faces. In fact PCA can be reformulated as a high-dimensional density estimation technique and the probability of face class membership can be estimated directly [85]. Again for detection tasks, in [111] the face manifold is approximated by six Gaussian clusters (a modified version of k-means clustering algorithm is used to find the Gaussian centroids and the covariance matrices). Mixtures of factor analysis, a technique closely related to PCA, are used to represent the face manifold for recognition [36] and detection [123].

2.4 Classification in the face subspace

Principal component analysis finds a subspace which contains most of the variation present in the training face images. As a large part of the variation in face image gray levels is due to illumination differences, we expect that the first eigenvectors or eigenfaces encode illumination information. This is confirmed by our results presented in Figure 2.3. The third eigenvector encodes left-right shadowing due to illumination from the side. In other words, the PCA on faces does not provide more discrimination of identity than the image space as it explains the face data disregarding identity information. Classifying faces using Euclidean distance in the face space or in the image space should not differ very much as already pointed out in [15, 7].

On the other hand, the PCA low dimensionality of the face subspace allows to construct face classifiers.

Let $\mathbf{y}$ be the principal components of a face vector $\mathbf{x}$ of person C_p that is $\mathbf{y} = \Phi^T \mathbf{x}$. The fundamental hypothesis underlying many subspace methods is that the probability density $p(\mathbf{y}|C_p)$ is Gaussian with mean $\boldsymbol{\mu}_p$ and covariance matrix Σ_p. This probability density describes ideally the face appearance of C_p in all possible illumination conditions, expressions and

poses in the face subspace.

The ultimate goal of face verification is to find a discriminant score function $s(\mathbf{x})$ that will allow to classify the genuine ω_a and impostor ω_b classes by thresholding the score. Our face model allows us to look for the score function in the low-dimensional face subspace i.e. $s(\mathbf{y})$. Recall from Chapter 1 that the optimal score function is given by the likelihood ratio $l(\mathbf{y})$

$$l(\mathbf{y}) = \frac{p(\mathbf{y}|\omega_a)}{p(\mathbf{y}|\omega_b)} \underset{\omega_a}{\overset{\omega_b}{\lessgtr}} \eta. \tag{2.3}$$

For person C_p, the genuine density is simply $p(\mathbf{y}|C_p)$ while the impostor density is the mixture made up of all densities corresponding to subjects other than from C_p

$$p(\mathbf{y}|\omega_b) = \sum_{i \neq p} \pi_i p(\mathbf{y}|C_i),$$

where the π_i's as the mixture weights. In this case a set of training impostors is required.

Alternatively, we may assume the impostor density to be uniformly distributed. This assumption is natural as the number of impostors is virtually infinite and one may assume that they populate the face space uniformly. In that case, the likelihood ratio test of Equation (1.1) reduces to

$$p(\mathbf{y}|C_p) \underset{\omega_a}{\overset{\omega_b}{\lessgtr}} \eta, \tag{2.4}$$

where η is a threshold. This equation states that the user is accepted if the probability density exceeds a certain pre-defined value. This is similar to a Maximum Likelihood (ML) estimation (where the estimated parameter maximises the likelihood of the observed data) versus Maximum a Posteriori (MAP) estimation (where the parameter to estimate has a *a priori* distribution).

Interestingly, many face verification methods can be derived from this model by making different assumptions on the density $p(\mathbf{y}|C_p)$. As the number of samples available to estimate this density is too small to derive a non-parametric estimate (even in the face subspace), we must assume a particular form for $p(\mathbf{y}|C_p)$ and convert the general density estimation into a parametric one. Multivariate normal density is a natural choice to model the density $p(\mathbf{y}|C_p)$. This density can be written

$$p(\mathbf{y}|C_p) = \frac{1}{(2\pi)^{N/2}|\Sigma_p|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_p)^T \Sigma_p^{-1} (\mathbf{y} - \boldsymbol{\mu}_p)\},$$

where $\boldsymbol{\mu}_p$ is the mean of $\mathbf{y}|C_p$. Different hypotheses on Σ_p lead to different score definitions as detailed below:

- When Σ_p is assumed to be the identity matrix, it is easy to show that the rule (2.4) is equivalent to

$$\|\mathbf{y} - \boldsymbol{\mu}_p\| \underset{\omega_a}{\overset{\omega_b}{\lessgtr}} \eta. \tag{2.5}$$

  This is simply a Euclidean distance classifier which corresponds in the face recognition scenario to the eigenface technique used by Turk and Pentland [115]. Similarly Equation (2.3) leads to a decision rule where each impostor of the training set contributes to the decision.

- We can assume Σ_p to be equal to the general covariance matrix Σ defined by (2.1). In this case the Euclidean distance in Equation (2.5) is replaced by the Mahalanobis distance [30] leading to the following decision rule

$$(\mathbf{y} - \boldsymbol{\mu}_p)^T \Lambda^{-1} (\mathbf{y} - \boldsymbol{\mu}_p) \underset{\omega_a}{\overset{\omega_b}{\lessgtr}} \eta, \tag{2.6}$$

  where Λ is the diagonal matrix defined by (2.2). The Mahalanobis distance has been shown to outperform the Euclidean distance [87, 121, 88, 35]. However, assuming Σ to be equal to the total covariance matrix amounts to treat intra-personal variation (i.e. variation described by Σ_p) as equal to the total variation, both inter- and intra-personal (described by Σ). This may be the best hypothesis to make if only one image per person is available for training, but if more images are available other hypotheses can be done on Σ_p in order to learn a better decision function.

- If $p(\mathbf{y}|C_p)$ is assumed to be the same for all subjects except for the mean

$$p(\mathbf{y}|C_p) = \boldsymbol{\mu}_p + \boldsymbol{\epsilon},$$

  where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Sigma_p)$, we have that Σ_p is equal to the same covariance matrix Σ_w for all person p. Because the covariance matrix S_w is equal for all subjects, a pooled estimation can be used to estimate S_w

$$S_w = \frac{1}{qr} \sum_{i=1}^{q} \sum_{j=1}^{r} (\mathbf{y}_{ij} - \boldsymbol{\mu}_j)(\mathbf{y}_{ij} - \boldsymbol{\mu}_j)^T, \tag{2.7}$$

  where $\mathbf{y}_{ij}$ is the jth training image of person i there are q persons or classes and r images per person. The rank of this matrix is at most

$q(r-1)$, this may create computational problems where trying to inverse S_w. One way to solve this problem is to project $\mathbf{y}$ onto a lower-dimensional subspace (i.e. further extract some features) in order (i) to remove the null space of S_w so that it is full rank (ii) to increase class separability in the subspace. This leads to the well known Linear Discriminant Analysis (LDA) described in detail below.

- In the general case, we should assume that different subject's face appearance are governed by different covariance matrices. Unfortunately, the number of training images per person required to estimate a particular covariance matrix for each person is too large even under the Gaussian hypothesis. It would require too many enrolment recordings and is unrealistic.

  An intermediate case consists of assuming Σ_p different for each subject but constraining it to be diagonal or even proportional to the identity matrix. Eventhough these assumptions might not be satisfied, they usually lead to more robust estimates when training is very limited as the number of degree of freedom of the covariance is highly reduced.

  Another possibility is to implement a special setup for enrolment like the one used to record the PIE database at Carnegie Mellon University (USA) [110] in which many images of the same person in different illumination and pose conditions are recorded in a very short time. Another possibility is to acquire range images of the subject's face during the enrolment phase using one of the many commercially available 3D scanners. The 3D information then allows to render facial appearance of the individual in different pose and illumination conditions (see for example [10]).

Due to data constraints, the approach taken in this thesis is based on the common covariance matrix assumption, i.e. the covariance matrix of $\mathbf{y}|C_p$ is the same for all subjects. Under this assumption, performing a Linear Discriminant Analysis on the data appears to be very natural.

2.4.1 Linear Discriminant Analysis (LDA) for face images

Linear Discriminant Analysis [37] looks for a linear transform W from $\mathbb{R}^m$ to $\mathbb{R}^{m_{lda}}$ where $m_{lda} \leq q-1$ (q is the number of classes available for training) which maximise the class separability J. Among the several ways to

define the class separability associated to the linear transform W, the most common is

$$J(W) = \frac{\det |W^T S_b W|}{\det |W^T S_w W|},$$

where S_b is the inter-class scatter matrix defined by

$$S_b = \sum_j (\boldsymbol{\mu}_j - \bar{\mathbf{y}})(\boldsymbol{\mu}_j - \bar{\mathbf{y}})^T,$$

and S_w is defined in Equation (2.7) and is called intra-class scatter in classical presentation of LDA. Note that by definition, the rank of S_b is at most $q - 1$. The dimensionality of the LDA subspace is therefore limited by this number. It can be shown that finding W that maximises the separability criterion $J(W)$ is equivalent to solve the following generalised eigenproblem [37].

$$S_b W = S_w W \Lambda_{lda}, \tag{2.8}$$

where the diagonal matrix Λ_{lda} contains the eigenvalues. The eigenproblem (2.8) is solved by first applying a whitening transformation P to S_w, i.e.

$$P^T S_w P = I, \tag{2.9}$$

and then solving

$$P^T S_b P \; W' = W' \Lambda'.$$

The projection matrix is found to be $W = PW'$. In the LDA projection subspace, the intraclass scatter matrix is the identity matrix: after the projection S_w becomes

$$W^T S_w W = W'^T P^T S_w P W' = W'^T W',$$

and because W' is an orthonormal matrix we have $W'^T W' = I$.

Therefore the new representation of the signal $\mathbf{z} = W^T \mathbf{y} = W^T \Phi^T \mathbf{x}$ has an identity scatter S_w and spans the directions where $P^T S_b P$ has maximal variance.

The power of LDA lies in the fact that intra-personal variations are modelled by averaging the variation over all classes. This allows to find a good representation for classification of high dimensional data as faces if the assumption S_w constant across classes is satisfied. However, the training set should contain at least two images per person in order to compute the LDA basis. LDA has been successfully applied to face recognition and verification by many authors [112, 33, 7, 126].

2.4.2 Matching distances in the LDA subspace

Any type of matching distances can be used in the LDA subspace to define the score between a probe face image $\mathbf{x}$ and a template $\bar{\mathbf{x}}_p$

$$s(\mathbf{x}) = d(W^T\Phi^T\mathbf{x}, W^T\Phi^T\bar{\mathbf{x}}_p) = d(\mathbf{z}, \bar{\mathbf{z}}_p).$$

We describe hereafter some possible distances.

Euclidean distance

After projection on the LDA subspace, S_w is the identity matrix, therefore the likelihood test of Equation (2.4) becomes

$$\|W^T(\mathbf{y} - \boldsymbol{\mu}_p)\| \underset{\omega_a}{\overset{\omega_b}{\lessgtr}} \eta,$$

in which the score $s(\mathbf{x}) = s(\Phi^T\mathbf{y})$ is simply the Euclidean distance (or L_2 norm) between the probe vector $\mathbf{x}$ and the template $\bar{\mathbf{x}}_p$ in the LDA subspace, i.e.

$$s(\mathbf{x}) = \|W^T\Phi^T(\mathbf{x} - \bar{\mathbf{x}}_p)\| = \|\mathbf{z} - \bar{\mathbf{z}}_p\|,$$

where $\bar{\mathbf{z}}_p$ is the template $\bar{\mathbf{x}}_p$ of person C_p expressed in the LDA subspace. This shows that under our hypotheses, the Euclidean distance is optimal according to the minimum error decision rule (2.4).

Normalised correlation

Although Euclidean distance is optimal in theory, various experiments have shown that the Euclidean distance is outperformed by other matching distances [61]. One of them is the *normalised correlation* which is defined by

$$s(\mathbf{x}) = \frac{\mathbf{z}^T\bar{\mathbf{z}}_p}{\|\mathbf{z}\|\|\bar{\mathbf{z}}_p\|}.$$

This function computes simply the cosine of the angle between the probe and the template vector in the LDA subspace. A high value of correlation corresponds to a good match between the two vectors which is opposite to our convention (good match correspond to low score). Hence we will simply change the sign of the normalised correlation in order to conserve the same convention, i.e.

$$s(\mathbf{x}) = -\frac{\mathbf{z}^T\bar{\mathbf{z}}_p}{\|\mathbf{z}\|\|\bar{\mathbf{z}}_p\|}.$$

In comparison to the Euclidean distance, the normalised correlation is independent of the norm of the two vectors. In fact, both matching distances are equivalent when the vectors are normalised to have unit norm. Let $\mathbf{x}$ and $\mathbf{y}$ be two vectors such that $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$, we have

$$\|\mathbf{x} - \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\mathbf{x}^T\mathbf{y} = 2 - \mathbf{x}^T\mathbf{y}, \tag{2.10}$$

which shows that only the threshold differs between a score based on normalised correlation and Euclidean distance.

Gradient direction metric (GDM)

The idea behind the GDM method is to compute the distance between the probe $\mathbf{z}$ and the reference $\bar{\mathbf{z}}_p$ in the LDA subspace along the most discriminative direction [61]. Recall from Chapter 1 that the optimal decision for person C_p in the LDA subspace is given by

$$P(C_p|\mathbf{z}) \underset{\omega_a}{\overset{\omega_b}{\lessgtr}} \frac{1}{2}.$$

Therefore, in a given point $\mathbf{z}$ of the subspace the most discriminative direction is the one along which the function $P(C_p|\mathbf{z})$ varies the most. This direction is in fact given by the gradient of the *a posteriori* probability $P(C_p|\mathbf{z})$ of person C_p given $\mathbf{z}$. Since the scatter S_w, which is a pooled estimation of the covariance of $\mathbf{x}$ conditioned to one of the N classes $\{C_1, C_2, ..., C_N\}$, is the identity matrix after LDA projection, the distribution $p(\mathbf{z}|C_j)$ is assumed to be Gaussian with mean $\bar{\mathbf{z}}_j$ and identity covariance matrix. Under this assumption the *a posteriori* probability can be written

$$P(C_p|\mathbf{z}) = \frac{p(\mathbf{z}|C_p)}{\sum_{j=1}^{N} p(\mathbf{z}|C_j)},$$

where equal prior have been assumed for all classes. The gradient of the *a posteriori* probability of person C_p given $\mathbf{z}$ can be easily computed

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

where a is a scalar that does not influence the gradient direction. See [61] for further details. The GDM score for person C_p is therefore given by

$$s(\mathbf{x}) = \frac{|(\mathbf{z} - \bar{\mathbf{z}}_p)^T \nabla P(C_p|\ \mathbf{z})|}{\|\nabla P(C_p|\ \mathbf{z})\|}.$$

Nearest mean based metric

Instead of using the rule given by Equation (2.4), one can use a set of training impostors (which may be simply the other users) in order to learn the decision boundary between genuine and impostor domain given by Equation (2.3). Basically this amounts to implement a classifier in the LDA subspace, a classifier which is trained using class ω_a and ω_b training samples.

We propose to use the Nearest Mean Classifier, which is a very simple linear classifier that assigns to a test pattern the class label of the nearest mean of the training samples for that class. In other words, each class is represented by one template : the mean of the training samples. In the case of face verification, the nearest mean classifier will assign the label genuine to the sample which is closer to the mean of person C_p samples $\bar{\mathbf{z}}_p$ and the impostor label to the sample closer to the mean of all other persons. This leads to the following decision rule

$$s(\mathbf{x}) = \frac{\|\mathbf{z} - \bar{\mathbf{z}}_p\|}{\|\mathbf{z} - \bar{\mathbf{z}}_{all}\|} \underset{\omega_a}{\overset{\omega_b}{\lessgtr}} \eta,$$

where $\bar{\mathbf{z}}_{all}$ is the mean of the impostor samples. This rule is in fact the same as the one defined in Equation (2.3) when both $p(\mathbf{z}|\omega_a)$ and $p(\mathbf{z}|\omega_b)$ are Gaussian with identity covariance matrix.

Nearest feature line

The nearest feature line distance [71, 72] makes use of the available information about classes contained in the multiple prototypes of each class. A subspace is constructed in the LDA subspace based on the multiple prototypes of the class. The NFL metric is defined as the Euclidean distance between the query and its projection in the subspace representing the class.

The idea behind the NFL metric is that a face image (or an image of any object) of a person changes continuously from one prototype to another in some way, drawing a trajectory in the feature space. The set of all such trajectories constitute a subspace in the feature space. A new face image of the same person should be close to the subspace although not necessarily close to the prototype images.

Let $\mathbf{z}_1$ and $\mathbf{z}_2$ be two training images of the same person, and $\mathbf{z}$ a test image, the Feature Line (FL) score is defined by

$$s(\mathbf{x}) = \|\mathbf{z} - \mathbf{p}\|,$$

where $\mathbf{p}$ is the projection of $\mathbf{z}$ on the feature line $\mathbf{z}_1\mathbf{z}_2$. The projection point $\mathbf{p}$ can be computed as

$$\mathbf{p} = \mathbf{z}_1 + \nu(\mathbf{z}_2 - \mathbf{z}_1),$$

where the paratemeter ν which describes the position of $\mathbf{p}$ relative to $\mathbf{z}_1$ and $\mathbf{z}_2$, is given by

$$\nu = \frac{(\mathbf{z} - \mathbf{z}_1)^T(\mathbf{z}_2 - \mathbf{z}_1)}{\|\mathbf{z}_2 - \mathbf{z}_1\|^2}.$$

Clearly when $\nu = 0$ we have $\mathbf{p} = \mathbf{x}_1$ and when $\nu = 1$ we have $\mathbf{p} = \mathbf{x}_2$. See [72] for further details.

2.4.3 Probabilistic Matching for face verification

Another face recognition technique that we have implemented in this thesis is the Probabilistic Matching (PM) algorithm that has been proposed by Moghaddam *et al.* [86, 84]. Although it does not define explicitely a face subspace, the algorithm relies strongly on PCA.

The principle is as follows. Let $\mathbf{x}_r$, $\mathbf{x}_t$ be respectively the face template image and test face images. Instead of classifying the image $\mathbf{x}_t$ directly as genuine or impostor, the method classifies the image difference $\mathbf{\Delta} = \mathbf{x}_r - \mathbf{x}_t$ as *intrapersonal variations* Ω_I, if $\mathbf{\Delta}$ corresponds to a difference between images of the same person, or *interpersonal variations* Ω_E, if $\mathbf{\Delta}$ corresponds to a difference between images of different persons.

Having observed a certain image difference $\mathbf{\Delta}$, we estimate the probability of image variation $P(\Omega_I|\mathbf{\Delta})$, i.e. the probability that $\mathbf{\Delta}$ is due to intrapersonal image variation, and base our decision on this probability. Note that because $P(\Omega_I|\mathbf{\Delta}) + P(\Omega_E|\mathbf{\Delta}) = 1$, we can equivalently base our decision on $P(\Omega_E|\mathbf{\Delta})$. The probability of intrapersonal variation can be expressed thanks to Bayes formula:

$$P(\Omega_I|\mathbf{\Delta}) = \frac{p(\mathbf{\Delta}|\Omega_I)P(\Omega_I)}{p(\mathbf{\Delta}|\Omega_I)P(\Omega_I) + p(\mathbf{\Delta}|\Omega_E)P(\Omega_E)}. \tag{2.11}$$

To compute $P(\Omega_I|\mathbf{\Delta})$ we need to estimate the density $p(\mathbf{\Delta}|\Omega_I)$ and $p(\mathbf{\Delta}|\Omega_E)$ for intraclass (or intrapersonal) and interclass (interpersonal) variation which is difficult because of the high dimensionality of $\mathbf{\Delta}$ ($\mathbf{\Delta} \in \mathbb{R}^n$ where n is the number of pixels in the image). Interestingly, $\mathbf{\Delta}|\Omega_I$ and $\mathbf{\Delta}|\Omega_E$ are not linearly separable as they are both *zero-mean*.

As the discussion is identical for $p(\mathbf{\Delta}|\Omega_I)$ and $p(\mathbf{\Delta}|\Omega_E)$ we drop the subscripts I and E for clarity. We assume that $p(\mathbf{\Delta}|\Omega)$ is Gaussian with

covariance matrix Σ_Δ

$$p(\boldsymbol{\Delta}|\Omega) = \frac{1}{(2\pi)^{n/2}|\Sigma_\Delta|^{1/2}} e^{-d(\boldsymbol{\Delta})/2},$$

where $d(\boldsymbol{\Delta})$ is the quadratic function

$$d(\boldsymbol{\Delta}) = \boldsymbol{\Delta}\Sigma_\Delta^{-1}\boldsymbol{\Delta}.$$

The function $d(\boldsymbol{\Delta})$ can be rewritten in the coordinate system in which the covariance matrix is diagonal, i.e.

$$d(\boldsymbol{\Delta}) = \boldsymbol{\xi}\Lambda_\Delta^{-1}\boldsymbol{\xi},$$

where $\boldsymbol{\xi}$ is the expression of the image difference $\boldsymbol{\Delta}$ in the new coordinate system and Λ_Δ the diagonal matrix containing the eigenvalues of Σ_Δ. The function $d(\boldsymbol{\Delta})$ can be written

$$d(\boldsymbol{\Delta}) = \sum_{i=1}^{n} \frac{\xi_i^2}{\lambda_{\Delta i}},$$

where the $\lambda_{\Delta i}$ are the eigenvalues of Σ_Δ. Even in this simplified form, $d(\boldsymbol{\Delta})$ is difficult to evaluate. For this reason, $d(\boldsymbol{\Delta})$ is estimated using a lower dimensional representation in the principal subspace (PCA subspace), which captures the majority of the variance of the data. Therefore the summation can be divided in two independent parts corresponding to the principal subspace with dimensionality m and its orthogonal complement with dimensionality $m - n$, that is

$$d(\boldsymbol{\Delta}) = \sum_{i=1}^{m} \frac{\xi_i^2}{\lambda_{\Delta i}} + \sum_{i=m+1}^{n} \frac{\xi_i^2}{\lambda_{\Delta i}}.$$

The first term can be computed by applying PCA to image differences (either intrapersonal or interpersonal). As the second term contains unknown eigenvalues, the authors of the method propose to approximate this term by

$$\frac{1}{\rho} \sum_{i=m+1}^{n} \xi_i^2$$

where ρ is the parameter that can be adjusted on training data. It can be easily shown that $\sum_{i=m+1}^{n} \xi_i^2$ is simply the squared reconstruction error $\boldsymbol{\epsilon}^2(\boldsymbol{\Delta})$, i.e.

$$\sum_{i=m+1}^{n} \xi_i^2 = \|\boldsymbol{\Delta}\|^2 - \sum_{i=1}^{m} \xi_i^2 = \boldsymbol{\epsilon}^2(\boldsymbol{\Delta}),$$

and can be computed easily. Therefore we can write our estimate of the probability density $p(\mathbf{\Delta}|\Omega)$ as

$$\hat{p}(\mathbf{\Delta}|\Omega) = \left[\frac{\exp\left(-\frac{1}{2}\sum_{j=1}^{m}\frac{\xi_j^2}{\lambda_j}\right)}{(2\pi)^{m/2}\prod_{j=1}^{m}\lambda_j^{1/2}}\right] \cdot \left[\frac{\exp\left(-\frac{\epsilon^2(\mathbf{\Delta})}{2\rho}\right)}{(2\pi\rho)^{(n-m)/2}}\right]. \tag{2.12}$$

This equation states that the density estimate is formed by a product of two marginal densities, one in the principal subspace and the other in its orthogonal complement. The authors showed that the optimal value of ρ is theoretically the average of eigenvalues in the orthogonal subspace. The estimate given by (2.12) is then used in Equation (2.11) (for both intrapersonal and interpersonal variation) to compute $P(\Omega_I|\mathbf{\Delta})$. More details on the high dimensional density estimation can be found in [85].

Practical considerations

In practice, the exact value of the intrapersonal variation probability given by Equation (2.11) is not important, as the likelihood ratio is rather used and compared to a threshold η determined on training data, i.e.

$$l(\mathbf{\Delta}) \underset{\omega_a}{\overset{\omega_b}{\lessgtr}} \eta.$$

By taking the logarithm of the likelihood ratio, we obtain the *probabilistic matching distance* $S(\mathbf{x}_r, \mathbf{x}_t)$ between the two images $\mathbf{x}_t$ and $\mathbf{x}_r$, which can be written

$$\begin{aligned} S(\mathbf{x}_r, \mathbf{x}_t) = & \\ & (\boldsymbol{\xi}_{rI} - \boldsymbol{\xi}_{tI})^T \Lambda_I^{-1} (\boldsymbol{\xi}_{rI} - \boldsymbol{\xi}_{tI}) + \frac{\epsilon_I^2(\mathbf{\Delta})}{\rho_I} \\ & -(\boldsymbol{\xi}_{rE} - \boldsymbol{\xi}_{tE})^T \Lambda_E^{-1} (\boldsymbol{\xi}_{rE} - \boldsymbol{\xi}_{tE}) - \frac{\epsilon_E^2(\mathbf{\Delta})}{\rho_E}, \end{aligned}$$

where $\boldsymbol{\xi}_{rI}$, $\boldsymbol{\xi}_{tI}$ and $\boldsymbol{\xi}_{rE}$, $\boldsymbol{\xi}_{tE}$ are the PCA projections of vector $\boldsymbol{x}_r$ and $\boldsymbol{x}_t$ in the intrapersonal space and interpersonal space respectively, Λ_I and Λ_E are the intra- and interpersonal eigenvalue matrices. Note that $S(\boldsymbol{x}_r, \boldsymbol{x}_t)$ is a distance measure, i.e. it is small when the images come from the same

person. By developing $\epsilon_I^2(\mathbf{\Delta})$ and $\epsilon_E^2(\mathbf{\Delta})$ the distance can be written

$$\begin{aligned} S(\mathbf{x}_r, \mathbf{x}_t) = & \\ & (\boldsymbol{\xi}_{rI} - \boldsymbol{\xi}_{tI})^T \Lambda_I^{-1} (\boldsymbol{\xi}_{rI} - \boldsymbol{\xi}_{tI}) - \frac{1}{\rho_I} \|\boldsymbol{\xi}_{rI} - \boldsymbol{\xi}_{tI}\|^2 \\ & -(\boldsymbol{\xi}_{rE} - \boldsymbol{\xi}_{tE})^T \Lambda_E^{-1} (\boldsymbol{\xi}_{rE} - \boldsymbol{\xi}_{tE}) + \frac{1}{\rho_E} \|\boldsymbol{\xi}_{rE} - \boldsymbol{\xi}_{tE}\|^2 \\ & +(\frac{1}{\rho_I} - \frac{1}{\rho_E}) \|\mathbf{x}_t - \mathbf{x}_r\|^2. \end{aligned}$$

Note that the probabilistic matching distance has five contributing terms. The first term is the distance in the intrapersonal subspace, as it is positive (Λ_I contains positive terms) it increases the distance between reference and test images. The third term is the counterpart in the interpersonal subspace, it reduces the distance as it counts for how likely is that the images come from different person. The three remaining terms take into account the error committed by working in the principal component subspaces. These terms are weighted by two parameters. Interestingly, this error contains a term with the Euclidean distance in the image space.

From the last equation we see that four parameters need to be adjusted: ρ_E, ρ_I and the eigenvector numbers m_I, m_E for the intra- and interpersonal PCA which define the sizes of matrices Λ_i and Λ_e.

The optimal values of these parameters were found by varying them and choosing the ones that give the minimal error on the validation set (see Chapter 3). Then the optimal parameter values are used to assess generalisation performance on the test set. To find the minimal EER, we performed an exhaustive search in the parameter space to find an approximate minimum. The corresponding parameters are then used as initial values for Nelder-Mead simplex algorithm [100]. Note that the EER is very sensitive to ρ_I and ρ_E, it is then very difficult to find the global minimum.

2.5 Handling mis-registered images

2.5.1 Registration technique

Up to now we have considered that the face images to be processed by the verification algorithm are perfectly registered. The term registration is somewhat vague as "perfectly registered images" may have different meanings. For example shape information can be used to normalise by non-rigid mapping the face image so as all normalised faces images share

the same average shape. This approach can be quite time consuming as many points have to be localised by a human operator in order to extract a precise shape information. Another possibility is to use a reduced number of manually detected (i.e. by a human operator) facial feature coordinates in order to determine the parameters of an affine transformation from the original image to a registered image. The affine transformation parameters are determined so that manually detected facial features are on the same locations for all normalised images. A rigid transform is characterised by 4 parameters (rotation angle, scale factor, translation in the x- and y-axis) so only two points are required to estimate the transform. This is the approach taken in this thesis. However, variants are possible. If the coordinates of a third point non-collinear to the two others is available, we may determine six transform parameters : for example by determining different scale factors in the x- and y-axis.

The chosen facial features should be in rigid part of the face, to avoid undesirable feature displacements due to face expression which lead to inaccurate affine parameters estimation. In this work, two manually detected facial features (the two centres of the pupil for the XM2VTS database and the middle point between the corners of each eye for BANCA database, the latter features can be manually pin-pointed with more accuracy) giving the coordinates x_1, y_1, x_2, y_2, were used to determine the 4 parameters t_x, t_y, θ, α for transform from the initial face image $I(x, y)$ to the registered face image $I(x', y')$. The transform is given by

$$x = t_x + \alpha(x' cos\theta + y' sin\theta),$$
$$y = t_y + \alpha(-x' sin\theta + y' cos\theta),$$

where t_x, t_y are translations in the image plane, θ the rotation angle and α the scaling factor. The target pupil centre locations on the registered image are depicted in Figure 2.6. This leads to four equations with four unknown parameters which can be solved easily. Because images are discrete signals, we must interpolate $I(x, y)$ to find the values of $I(x', y')$. This has been done using a classical bilinear interpolation [39].

2.5.2 Mis-registrations

The variability of facial images of the same individual, due to illumination rotations in 3D, expression, etc., makes the task of face verification and recognition very challenging. On top of these sources of variability, it has been shown that registration errors or geometric distortions severely

degrade the performances of face recognition [124] and verification [56]. Since perfect face localisation technology has not been achieved yet, a realistic face recognition/verification system has to handle the variability induced by mis-registrations.

In this Section, we study how mis-registrations affect the performance of LDA in face verification, i.e. when the images to be matched are not perfectly aligned or registered. Our experiments show that the degradation in performance is substantial even for moderate amplitude distortions unless the variability added by the imperfect localisation is incorporated into the calculation of the LDA vectors.

To evaluate the sensitivity of LDA to image registration, we added random displacements to the ground-truth eye coordinates (located manually) generating two-dimensional translations, rotations and scalings. These simulated mis-registrations offer the following advantages: 1) The results are independent of a particular registration algorithm; 2) the mis-registrations are easily controlled. The random displacements are drawn from a two-dimensional Gaussian variable $N(0, \Sigma)$ with zero mean and covariance

$$\Sigma = \begin{pmatrix} 0 & \sigma^2 \\ \sigma^2 & 0 \end{pmatrix}$$

and are added to each eye coordinates independently. The statistical distribution of the errors of an automatic eye locator is unlikely to be very far from our perturbation model. The length r of the displacement from the ground-truth eye coordinates is known to be described by a Rayleigh random variable whose probability density function is given by

$$p_R(r) = \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}}.$$

The mean of the distribution is $\bar{r} = \sqrt{\frac{\pi}{2}}\sigma$, the median $r_{\frac{1}{2}} = \sqrt{2\ln 2}\sigma$ and the upper quartile $r_{\frac{3}{4}} = 2\sqrt{\ln 2}\sigma$. The statistics of the displacement length distribution depend linearly on the standard deviation σ of the Gaussian distribution. Consequently, this parameter controls the displacement magnitude distribution.

2.5.3 The hypersurface of mis-registered images

Consider a face image $\mathbf{x}$ as a vector of $\mathbb{R}^n$, perfectly registered, which undergoes an affine transformation $T_{\vec{a}}$ with parameter $\vec{a} = (t_1, t_2, \theta, \alpha)$ where

t_1, t_2 represent translations in the image plane, θ represents the rotation angle and α the scaling factor. This is what happens if the face image is not perfectly localised or if manually located eye coordinates are perturbed to simulate this. The result is a mis-registered image $\tilde{\mathbf{x}}$

$$\tilde{\mathbf{x}}(\vec{a}) = T_{\vec{a}}(\mathbf{x}).$$

In the image space, the set of all transformed faces lies on a 4-dimensional non-linear hypersurface $S(\vec{a})$ for a given face image $\mathbf{x}$. This should prevent the use of a linear method such as LDA for handling the mis-registration problem and for this reason, alternative solutions have been proposed such as the mixture of local linear subspaces [36] and the multiple distorted LDA subspaces [124]. However, the use of multiple subspaces or mixtures of local subspaces increases a lot the complexity, the number of parameters and the storage space needed for the system (For example the distorted subspace method requires an exhaustive search in the transformation parameter space). In fact, these methods have been devised to handle large image distortions.

For small displacements $\vec{a}$, which are likely to happen with an automatic face localisation system, we may expect that the 4-dimensional non-linear hypersurface is approximately contained in a linear subspace (we expect the linear subspace dimension to be higher than the hypersurface dimension). Thus for any $\vec{a}$ in a certain range, a transformed image $\tilde{x}(\vec{a})$ can be approximated by a linear combination of vectors ϕ_j independent of $\vec{a}$:

$$\tilde{x}(\vec{a}) = T_{\vec{a}}(x) \approx \sum_j \alpha_j(\vec{a})\phi_j.$$

To find the ϕ_j's, one option is to perform a PCA on a set of transformed images of the same face image. Any $\tilde{\mathbf{x}}(\vec{a})$ with a displacement $\vec{a}$ in the same range can be reconstructed faithfully using a limited number of PCA vectors. Figure 2.7(a) and (b) show a mis-registered image and its reconstruction using 30 PCA eigenvectors. A 3 pixel displacement has been added to left eye vertical ground-truth coordinate and subtracted to the right eye coordinate. The resulting rotation angle is about 3.5 degrees. When the amplitude of the displacement increases, the dimensionality of the linear subspace that approximates the distortion increases too, i.e. more eigenvectors are needed to reconstruct the distorted image. 100 eigenvectors (Figure 2.7(d)) have been used to reconstruct the image on Figure 2.7(c) which is rotated by 5.5 degrees. When the displacement magnitude becomes too large, the linear subspace approximation is insufficient to reconstruct the mis-registered image (Figure 2.7(e) and (f)).

Two conclusions can be drawn:

- When the displacement with respect to perfectly registered face image is small, the mis-registered face can be approximated by a linear subspace. The basis vectors of this subspace can be found performing PCA on registered and mis-registered images.
- Small displacements lie in a low dimensional subspace for each individual. If mis-registered images are included in the calculation of intraclass and interclass scatter matrices of LDA, the LDA will discard the variation due to mis-registration.

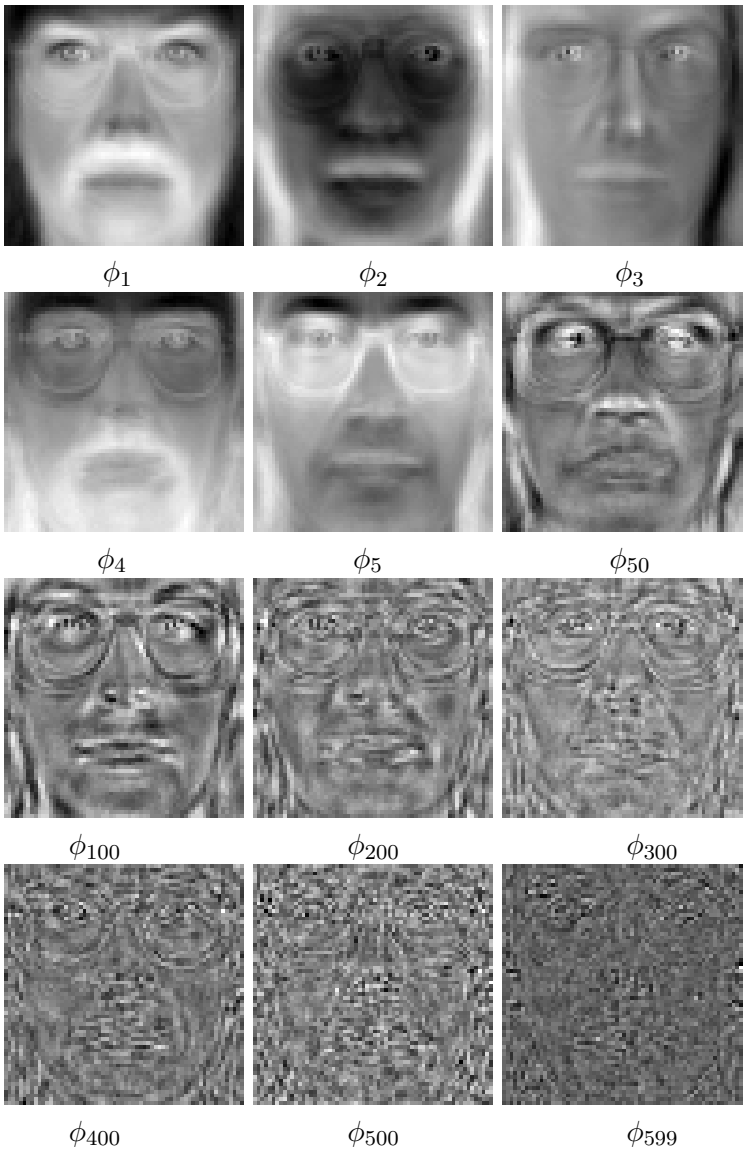


Figure 2.3: *Basis vectors of the principal subspace or eigenfaces [115] computed from the training set (600 face images) of the extended M2VTS database. The eigenvectors are ranked by decreasing corresponding eigenvalue.*

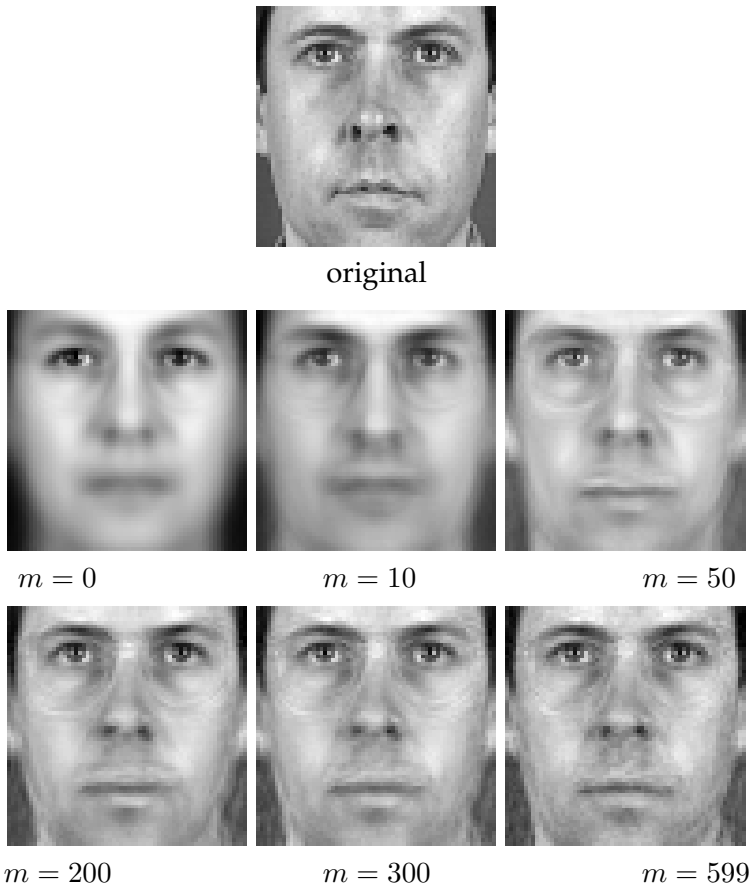


Figure 2.4: *Reconstructed face images from the extended M2VTS database using variable number of principal components m. The top image is the original image. Note that the subject is not present is the training set used to compute the principal subspace basis.*

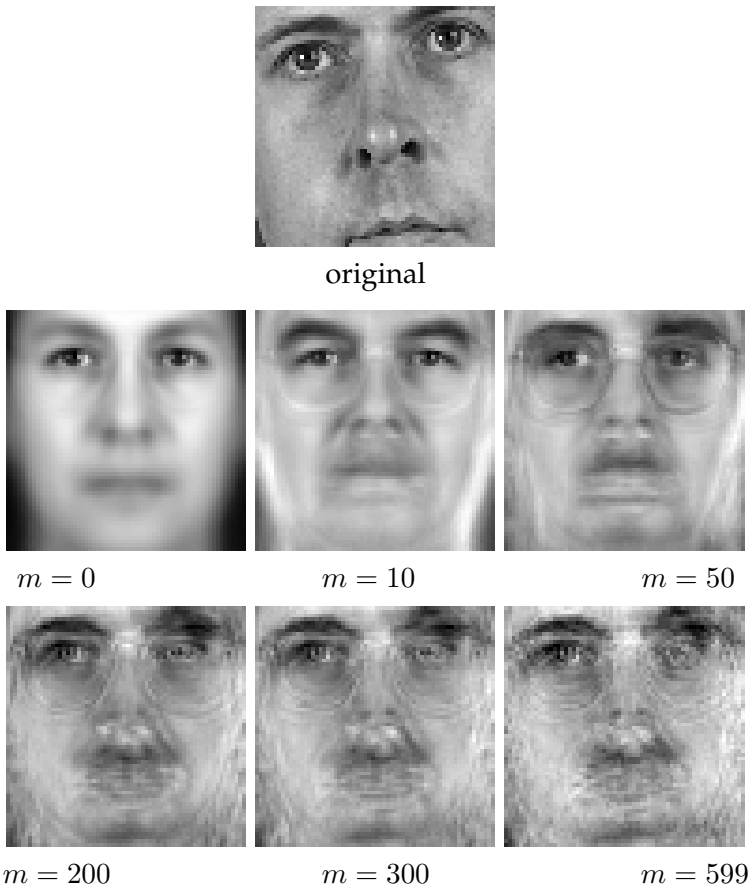


Figure 2.5: *Effect of mis-registration on the reconstruction of a face images. The principal subspace is the same as for Figure 2.4, i.e. it models only properly registered faces. It fails to model correctly a rotated and/or scaled face image.*

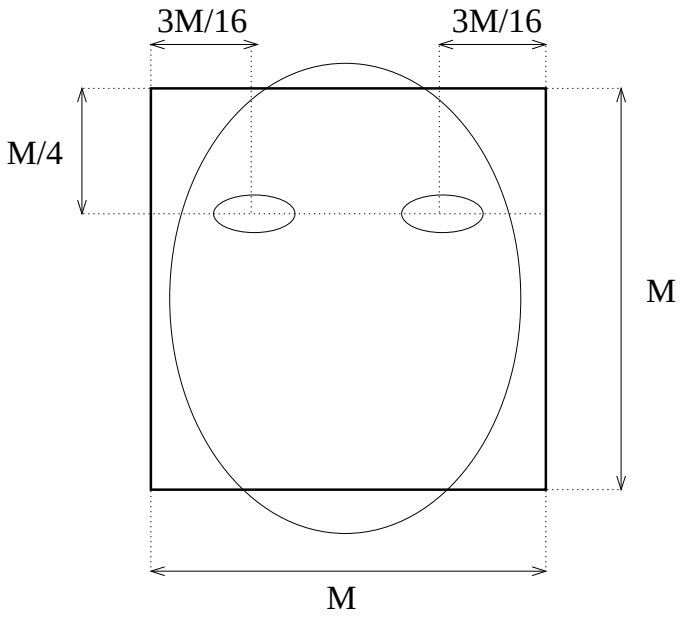


Figure 2.6:

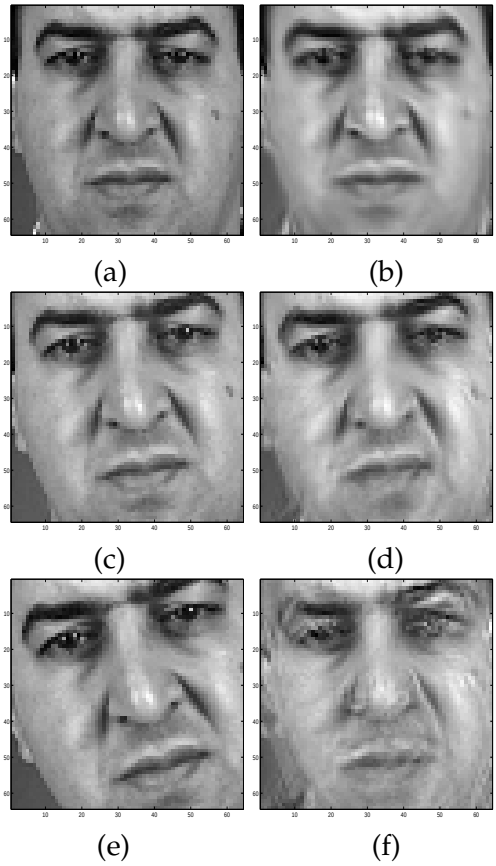


Figure 2.7: *Examples of mis-registered images from XM2VTS database (a - c - e) with different eye coordinate displacement magnitudes and their reconstructions (b - d - f).*

Chapter 3

Face data sets and experimental protocols

3.1 Introduction

The practical goal of pattern recognition is to design algorithms which perform their classification task by committing as little errors as possible. A key parameter of such algorithms is the probability of error. In biometric verification, this error breaks down in a probability of false acceptance e_A and a probability of false rejection e_R defined in Chapter 1.

Due to its probabilistic nature, the error probabilities can be computed analytically only when the probability law generating the patterns is known completely. In most practical cases this law is unknown. The easiest way to evaluate the algorithm remains the estimation of the error on a test dataset of the targeted patterns, e.g. a set of face images to estimate the error rates of a face verification algorithm.

We stress here the importance of having test data representing as closely as possible the data encountered in the application. In statistical words, the data sample should represent the population. For face verification it means that the recording conditions of biometric test data should be ideally identical to the recording conditions expected in the application. The term recording conditions should be understood in a broad sense: it includes the hardware (sensor quality), the environmental conditions (background noise, part of the body present in the signal, illumination conditions), as well as temporal changes (e.g. changes due to illness, aging, hair style, etc.)

This shows that, due to the design process, performance of biometric

algorithms is inherently bound to the recording conditions. For instance, a given speaker verification algorithm may show a performance level when used with a given microphone but the level may change quite drastically with another microphone. Experimental results are only valid for a given dataset, i.e. for given recording conditions in the biometric context. Error rates should always be accompanied by datasets. Furthermore, the comparison of two algorithms in terms of error rates can only be done in the rates were obtained on the same data using strictly the same testing protocol.

Very often the design of classification algorithm requires data as well. As data is expensive to aquire, process and store, it should be used in an efficient way to be able to design and test the algorithm with limited data. Testing protocols must be devised carefully to avoid methodological flaws, which happen if the algorithm is tested using the same data that was used for the design. Strategies and guidelines for the design testing protocols can be found is [30].

In this chapter, we describe in detail the two face databases that were chosen to conduct the experiments. Experimental protocols that were carefully designed for these databases are also described and the motivations behind the protocols are underlined.

Many face databases exist; one of the major choice factors is the adequacy with application requirements. The designer should answer questions such as: is the application intended to work in controlled or uncontrolled lighting? Is the background controlled or not?, Is the number of sessions sufficient for simulating temporal effects? etc. Other factors include the size of the database, its availability, the existence of a well defined testing protocol and its popularity for comparing results.

The first database that we have worked on is the extended M2VTS database, that was recorded during the European Union ACTS project M2VTS (Multi Modal Verification for Teleservices and Security applications). This project deals with access control by the use of multimodal identification. The goal of using a multimodal recognition scheme was to improve the recognition efficiency by combining single modalities, namely face and voice features.

The second database used in this work is the BANCA database that was recorded in the context of the European project successor of M2VTS namely the project BANCA (Biometric Access Control for Networked and e-Commerce Applications). The objective of this project is to develop and implement a complete secured system with enhanced identi-

fication, authentication and access control schemes for applications over Internet such as tele-working and Web-banking services. One of the major innovations of this project is to obtain an enhanced security system by combining classical security protocols with robust multimodal biometric verification schemes based on speech and facial images. No existing face database fullfils the particularities of the applications targeted by the project BANCA. This is why a new database, called the BANCA database, had to be recorded under precise specifications to develop and evaluate the biometric technology meant for network and Internet security.

3.2 The extended M2VTS database (XM2VTS)

The extended M2VTS database [82], recorded at the University of Surrey (UK), is an extended version of the M2VTS database [96], recorded at the Université catholique de Louvain. The extension was required because of the relatively small size of the M2VTS database (37 persons)

3.2.1 Database description

XM2VTS is a publicly available multimodal database recorded specifically for assessing the performance of biometric approaches to identity verification. It contains synchronised 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 recording is a speech sequence and a head rotation sequence. The images contain head and shoulder parts of the body, on a blue uniform background. The total database reaches 4000 GBytes in size.

Our experiments were performed on the still frontal face images gathered in the sets called CDS001 and CDS006, which were extracted from the video sequence. The original size of the colour images is 720x576 pixels coded on 24 bits in RGB (255 possible colour values per channel). Note that images are stored in raw format to avoid lossy compression artifacts. In these face sets, the lighting is controlled. Figure 3.1 shows typical examples of XM2VTS images.

The motivation for the choice of this database is mainly its quite large size (295 persons and 2360 images in total), and popularity. Since it was recorded, the extended M2VTS database has become a standard in the audio- and video- biometric identity verification community. The rigorous protocol allows fair comparison between verification algorithms and two

face verification contests have been organised [78] and [81]. We can therefore determine where our algorithms stand with respect to the state-of-the-art. Also, since the database has been recorded over a long period of time, the appearance of the subjects varies due to changing in hair style, aging, expression, and facial hair. Furthermore, the database is multimodal which allows fusion of speech and face modality experiments. However, a major drawback of XM2VTS is that the lighting and the background are controlled. The results obtained on this database cannot be extrapolated to unconstrained environment.

Figure 3.1: *Typical images from the extended M2VTS database. Note how general appearance (hair style, make up, etc.) varies between images of the same person in different session (Images are taken from recording 1 of sessions 1 to 4).*

3.2.2 The XM2VTS protocol or "Lausanne Protocol"

A detailed description of the standard experimental protocol can be found in [76]. The protocol, especially designed for multimodal fusion exper-

iments, divides the database into 200 clients or system users (used for determining false rejection rate) and 95 impostors (used for determining false acceptance rate). For each client/impostor one frame per recording per session is selected thus giving 8 images per subject in total. The protocol specifies a partitioning of the database into disjoint sets for training, validation (referred to as evaluation set in [76]) and testing. Two configurations are defined.

In configuration I, three images are used for training the 200 clients or system users, that is for creating the templates or client models. Three other images of the 200 clients constitute the validation set, and are used to set algorithm parameters and the threshold. Recall from Chapter 1 that an unbiased threshold is obtained with independent data. The two remaining images, corresponding to the last session, constitute the test set. This independent data is used to compute the error rates. No optimisation is allowed on this data. The impostor set is divided in 25 impostors which are used for validation and the remaining 70 for testing. This partitioning is depicted in Figure 3.2. In this configuration, 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).

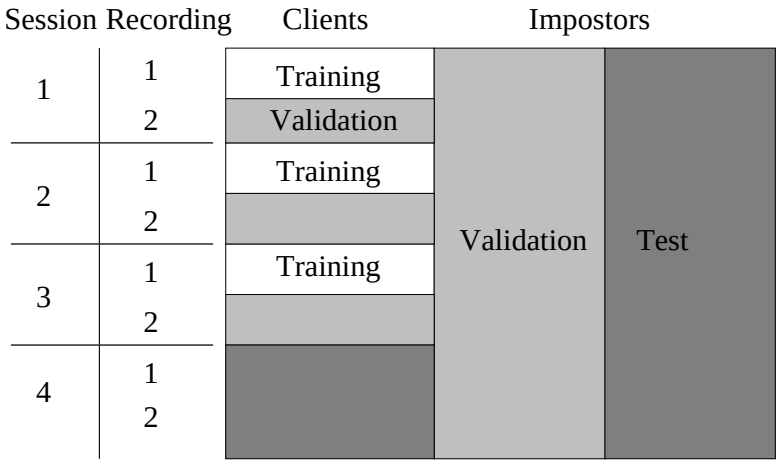


Figure 3.2: *Database partitioning as defined by the Lausanne protocol in configuration I*

In configuration II, the genuine users and impostors are partitioned in the same way. This time, four images corresponding to two sessions are used for training the clients or system users, Two other images from one session constitute the validation set. As for configuration I, the two remaining images, corresponding to the last session, constitute the test set. This partitioning is depicted in Figure 3.3. In this configuration, the validation set leads to 400 genuine or client matchings and 40,000 impostor matchings. The test set generates 400 genuine matchings and 112,000 impostor matchings.

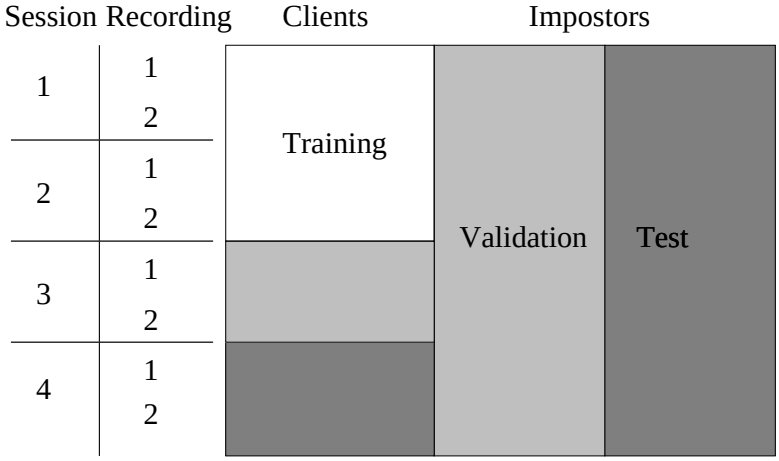


Figure 3.3: *Database partitioning as defined by the Lausanne protocol in configuration II*

As mentioned in Chapter 1, the evaluation of a biometric verification algorithms can be done by measuring two error rates: the False Acceptance (FAR) and the False Rejection Rate (FRR). As these two numbers depend on the threshold η (see Equations (1.7)), one method of evaluation is to plot the FAR on the x-axis and FRR on the y-axis parametrised by the threshold. The corresponding curve is the Receiver Operation Characteristic (ROC) curve. Apart from the complete ROC curve, the system can be assessed globally and coarsely through the Equal Error Rate (EER). At the threshold η_{EER},the FAR(η_{EER}) is equal to the FRR(η_{EER}), and they defines the EER.

Another approach consists of setting the threshold so that a satisfying trade-off is reached between the FAR and the FRR, and re-compute the FAR and FRR on an independent dataset. This approach required more

data but it is more realistic as it allows to assess if the threshold generalises well. The methodology adopted in the Lausanne protocol is the following. The threshold η_{EER} corresponding to the EER is set on the validation dataset. This *a priori* threshold choice implements the minimax test (see Chapter 1). The same threshold is then used to computed the FAR and FRR on the test set using Equations (1.7). Since η_{EER} is an estimate, there is no guarantee that the FAR and FRR calculated on the test data will be equal. The rates can be summed up in a single number: the Half Total Error Rate (HTER) which is defined as

$$\mathrm{HTER} = \frac{\mathrm{FAR} + \mathrm{FRR}}{2}.$$

Note that the 70 test impostors are completely unknown to the system which is a very important requirement for any rigorous verification protocol. In configuration I, it is worth noticing that the client validation set contains data very similar to the client training set as it is made up of the second recordings of the first three sessions. This lead likely to an optimistic threshold choice. In contrast, the test client images are made up of the last session, that is images which are quite different from the training images. In configuration II, more data is available for creating user templates (4 images) but in return only two images per system user are available for setting the threshold. However, the two validation images per user are coming from a different session. The threshold is thus likely to be more accurate in this case. We believe that data is employed more efficiently in configuration II: there is a better balance between training and validation data. This shows that the splitting of the available data into training and validation should be done carefully as it has an impact on the test performance. The study of how the dataset should be split for optimal performance goes beyond the scope of this work. Therefore we have limited ourselves to the standard configurations I and II of the Lausanne protocol.

Both configuration assumes that three sessions are available to train completely the system (i.e. to determine the client models and the threshold). This leads to an enrolment process which may be quite tedious as users have to give their biometric features on three different occasions. Furthermore, as enrolment must be done in a trusted environment, it is likely that users would have to get to a special "enrolment office" which may be expensive to run so many times. Clearly, a viable biometric system must limit the number of enrolment sessions.

3.2.3 Scalability issues

A major drawback of the Lausanne protocol is that it leads to results that are not scalable with the number of users.

Whereas the difficulty and complexity of the task increases in an identification task (where the identity accompanying the claim is unknown and the claim must therefore be compared to every client), the difficulty of the verification task stays constant since it involves only a one to one matching.

However, care must be taken when we analyse the effect of adding a new user to the system. If the client model of the new enrolment is built not independently from the existing clients, then all the models of the clients belonging to the system must be re-built. In [42] these kind of non-independent client model are said to be "template known to the system". Otherwise, the accuracy (false acceptance and false rejection) of the verification may be severely affected for this new client. The task of rebuilding may be very time consuming and may cause integration and implementation trouble. For example in a LDA-based face verification, often the LDA or PCA basis vectors (matrices Φ and W, see Chapter 2) are computed using the client set. In this case, these basis vectors should be re-computed after the face images of the new user has been added to the face data set. For p face images, the PCA eigenvector computation is $O(p^3)$ and $O(q^3)$ for LDA basis vectors where q is the number of PCA eigenvectors kept. These computations may become extremely heavy when the number of clients grows.

In order to obtained scalable results with growing number of users, the PCA and LDA bases should be computed using another data set than the client set in order to avoid the re-computation of eigenvectors. More generally, the model or template $\mathbf{y}_i$ of client i must be a function of the raw measurement $\mathbf{x}_i$ only, i.e.

$$\mathbf{y}_i = f(\mathbf{x}_i)$$

as opposed to

$$\mathbf{y}_i = f(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p)$$

as it is the case in the Lausanne protocol.

3.3 The BANCA database

The shortcomings of the extended M2VTS database have lead the research team of the European project BANCA to record another multimodal database.

The specifications of the BANCA database rest on the same basis as the extended M2VTS: the database has been recorded over several months, the number of users is relatively large, the database is multimodal and well founded test protocols have been defined.

The major innovation of the BANCA database is that it has been designed to reproduce as closely as possible the acquisition conditions (voice and video recordings) in 3 different realistic automatic identity verification scenarios. For example, user authentication during a network transaction is among the targeted applications of the BANCA project. In this scenario, the user whose identity is to be verified is typically in an unconstrained environment for example at home using a cheap web-cam. It means that during the acquisition of the data needed to authenticate the user, surrounding noise is present, as well as an unknown possibly moving background, and unknown variable illumination. It is therefore important that the database used for designing and testing biometric algorithms, reflects the noise and variability that would be present in reality. The other considered scenario is a ATM (automatic teller machine) simulation. The user is standing in a room adjacent to the bank. Therefore, in contrast to the XM2VTS database, the video recordings of the BANCA database contains high variability in illumination, pose, resolution, acquisition camera and in the background.

3.3.1 Database description

The BANCA database contains voice and video recordings of 260 people in several scenarios (controlled, degraded and adverse). The database is multi-lingual, i.e. 5 different languages (Spanish, Italian, French, German and English) are present with 52 subjects per language (26 males and 26 females). Each language specific population is itself subdivided into 2 groups of 26 subjects (13 males and 13 females), denoted in the following g1 and g2. Each subject recorded 12 sessions, each of these sessions containing 2 records: one true client access and one impostor attack. The impostor attacks are attempted only for subjects of the same sex, within the same group.

The 12 sessions were separated into 3 different scenarios: controlled

for sessions 1-4, degraded for sessions 5-8 and adverse for sessions 9-12. A low-cost camera has been used to record the session in the degraded scenario (simulating a user authentifying himself using its home PC and a low cost web-cam), while the expensive camera was used for controlled and adverse scenarios (simulating an ATM machine). From the 20-30 seconds of video recording, 5 frames were randomly selected, giving thus 5 frames per person per record for processing. Figure 3.4 shows the first frame of the 12 sessions of a subject.

In a given session, the impostor accesses by subject X were successively made with a claimed identity corresponding to each other subject from the same group (as X). In other words, all the subjects in group g recorded one (and only one) impostor attempt against each other subject in g and each subject in group g was attacked once (and only once) by each other subject in g. Moreover, the sequence of impostor attacks was designed so as to make sure that each identity was attacked exactly 4 times in the 3 different scenarios (hence 12 attacks in total). For each language, an additional set of 30 other subjects, 15 males and 15 females, recorded one session (audio and video). This set of data is referred to as world data.

3.3.2 The BANCA protocol

A protocol has been devised to have a realistic assessment of biometric algorithms and to be able to compare fairly different algorithm performances. A detailed description can be found in [4] and [8]

The database subjects are divided in a development set (used to set and adjust parameters of the algorithm) and an evaluation set (used to assess the algorithm on a separate data). It is very important that the evaluation set consists of different subjects than the development set (A condition that is not fulfilled in the XM2VTS database), because in real world applications, the algorithm parameters cannot be adjusted on users that will use the system. The group g1 and g2 are used successively as development and evaluation set.

Inside each set, we specify which sessions are to be used to create the client model (training sessions) and which ones are to be used to simulate access attempts (testing sessions). Several configurations were defined in this way:

- Matched controlled configuration (Mc) consists of training and testing images acquired in the controlled scenario.

Conf.name	Training sessions	Testing sessions
Matched controlled	1	2,3,4
Unmatched degraded	1	6,7,8
Unmatched adverse	1	10,11,12
Matched degraded	5	6,7,8
Matched adverse	9	10,11,12
Pooled test	1	2,3,4,6,7,8,10,11,12
Grand test	1,5,9	2,3,4,6,7,8,10,11,12

Table 3.1: *Sessions used for training and testing in the different configurations defined in the BANCA protocol*

- Matched degraded configuration (Md) consists of training and test images acquired in the degraded scenario.
- Matched adverse configuration (Ma) consists of training and test images acquired in the degraded scenario.
- Unmatched degraded configuration (Ud) consists of training images in controlled scenario and test images in degraded scenario
- Unmatched adverse configuration (Ua) consists of training images in the controlled scenario and test images in the adverse scenario
- Pooled test configuration (P) consists of the union Md, Ud and Ua
- Grand test configuration (G): one session in each in scenario for training and the rest for testing

Table 3.1 shows the sessions used to define each configuration.

The threshold and the algorithm parameters are first adjusted on group g1, so as to be at equal error rate (EER(g1)). The same parameters are then used on g2, leading thus to an a priori false acceptance and false rejection rates (FAR(g2) and FRR(g2)), reflecting faithfully the expected performance of a real system. The same procedure is carried out after swapping g1 and g2, leading then to EER(g2), FAR(g1) and FRR(g1).

The BANCA protocol is quite flexible and allows evaluation of verification algorithm in different conditions and with different constraints. Clearly the most demanding configuration is configuration P as it contains only one session (thus 5 very similar frames) to create the user template

and three different environment. At the same time it is the most realistic configuration: one training session means one enrolment session and therefore less trouble for users.

The strict BANCA protocol requires that the user template is built independently from the existing clients. This feature lead to scalable results with respect to the number of user as discussed in Section 3.2.3.

Figure 3.4: *First frame of the 12 sessions of a subject of the BANCA database. Note how illumination and pose change across the sessions. The image quality of sessions 5 to 8 recorded with a low-cost web cam is clearly inferior to the other sessions.*

Chapter 4

Face verification: experimental study

4.1 Introduction

Up to now we have approached the automatic verification of faces from a theoretic point of view. We have discussed the plausible hypotheses that can be made on the statistics governing the problem and how this leads to a decision rule. As face verification is a practical problem it is now time to test the different alternatives and try to draw some conclusions. Unfortunately, the conclusions that can be drawn are dependent on the face database that has been used for making the experiments. We must be very careful before generalising empirical results. For example in the early developments of LDA for face recognition, LDA has been reported to outperform pure PCA-based algorithm (i.e. eigenfaces) [7] and later to be outperformed by PCA [74]. In fact, it has been shown recently that the relative merits of the two methods (PCA and LDA) depend essentially on the size of the training set, more specifically the number of images per person in the database [77].

In this chapter we present our experimental study of the various face verification algorithms introduced in Chapter 2. The experiments were performed on the XM2VTS and BANCA face databases described in Chapter 3.

The chapter is organised as follows. In the first part, we present results from experiments performed on the XM2VTS database. Photometric normalisation techniques used in the experiments are described. Afterwards experimental results for template matching, eigenface, LDA and Proba-

bilistic matching algorithms are given and discussed. Impact of image size on the performance is studied. In the second part of the chapter, results obtained on the BANCA database are presented. As it performed the best on the previous experiments, the LDA based verification algorithm is evaluated in detail on this database. A photometric normalisation based on face symmetry is also presented.

4.2 Results on the XM2VTS database

4.2.1 Photometric normalisation

In order to remove as much as possible the influence of non-constant illumination between two images (see Figure 2.1), a photometric normalisation is applied to the raw gray level images. For these experiments, we have studied traditional normalisation techniques used in image processing, namely pixel value reduction and histogram equalisation. In pixel value reduction, the pixel values are translated and rescaled so as to have zero mean and unit variance (ZMUV). We will refer to this normalisation as ZMUV. This very basic photometric normalisation removes the effect of global uniform lighting differences, but of course the normalisation is ineffective if the illumination directions are different in the two images to compare.

In histogram equalisation, the pixels values are transformed non linearly so that the histogram of the image gray levels is as uniform as possible. A detailed description of the method is described in [39]. Basically, histogram equalisation increases the dynamic range of the pixels, i.e. all pixel values are equally distributed between 0 and 255.

More advanced photo-normalisation are possible. For example we can exploit face symmetry to normalise azimuthal direction illumination (see Section 4.3.3). In [41], another approach is taken, which tries to extract image reflectance independently of the illuminance. Based on biological arguments and smoothness constraints, an approximation of the reflectance is found. It is shown that the recognition rate using facial images normalised with this technique is much higher.

4.2.2 Z-normalisation or client specific threshold

As explained in Chapter 3, the threshold η from Equation 1.2 is set on the validation set at equal error rate (EER) and used on the test set to obtain the half total error rate (HTER) which is half of the sum of the FAR

and FRR. Following the testing protocol, each subject p of the training set has 200 impostor scores. These 200 scores can be used to estimate a *client specific threshold* $\eta^{(p)}$ for subject p. The client specific threshold allow more lax thresholds for persons who are difficult to impose while allowing more strict thresholds for persons who are easy to impose ("sheep" in [28]). This is a step towards a person specific approach as discussed in Section 1.2. Using the genuine scores for finding a client specific threshold is not advised because of the small number of genuine score per person.

There are many ways to define the threshold using the 200 impostor scores. From preliminary experiments that we performed, the following technique appeared to be the more robust. It is also the technique used in [54]. The specific threshold $\eta^{(p)}$ for subject p is obtained from the global threshold η by

$$\eta^{(p)} = \eta\sigma_s^{(p)} + \mu_s^{(p)},$$

where $\sigma_s^{(p)}$ and $\mu_s^{(p)}$ are the standard deviation and mean of the 200 impostor scores of person p. It is easy to see that using $\eta^{(p)}$ is equivalent to use η and rescale the scores $s^{(p)}$ related to person p so that the impostor scores for this person have zero mean and unit variance. This amounts to use $\tilde{s}^{(p)}$ instead of $s^{(p)}$ in Equation (1.2) with $\tilde{s}^{(p)}$ defined as

$$\tilde{s}^{(p)} = \frac{s^{(p)} - \mu_s^{(p)}}{\sigma_s^{(p)}}.$$

This normalisation is called Z-normalisation. If the impostor scores were Gaussian distributed, with different means and variances for each person, the Z-norm rescales the scores so that all impostor scores are distributed according to $\mathcal{N}(0, 1)$.

4.2.3 Baseline face verification algorithm

As a baseline face verification algorithm, we have chosen to match the images directly in the image space. This amounts to traditional template matching [15, 97].

The face images are cropped using the technique described in Section 2.5.1. For the XM2VTS database we have used manually located eye coordinates in order to focus on the purely verification aspects. Image size for this experiment is 64×64 pixels. This may seem quite low resolution, but as shown in Figure 2.4, image quality is more than acceptable. After cropping, images are photo-normalised using either histogram equalisation or pixel value reduction.

The training set of the XM2VTS database contains 3 images per subject in configuration I and 4 images per subject in configuration II. The incoming test image $\mathbf{x}_t$ claiming identity p is compared to all reference images $\mathbf{x}_{pi}$ of the subject p in the training set and the score is based on the minimum Euclidean distance, i.e.

$$s = \min_i \|\mathbf{x}_t - \mathbf{x}_{pi}\|.$$

The decision is based by comparing the score to a threshold as in Equation 1.2. Another very simple matching scheme is the normalised correlation introduced in Section 2.4.2.

Results of matchings (Euclidean distance (EUCL) and normalised correlation (NOC)) in the image space are given in Table 4.1 using a global threshold and 4.2 using client specific thresholds, or equivalently, Z-normal-ised scores. Each row reports the results on the evaluation data set and test data set. The acceptance (FAR) and the false rejection rates (FRR) are listed on the two middle columns of Table 4.1. The next column of Table 4.1 shows the half-total error rate (HTER). From the tables, we can see that

Configuration I

		Validation	Test		
Photo.	matching	EER	FAR	FRR	HTER
HIST	EUCL	9.83	11.54	8.50	10.12
	NOC	9.82	11.40	8.25	**9.82**
ZMUV	EUCL	11.83	13.30	10.00	11.65
	NOC	11.83	13.30	10.00	11.65
Configuration II					
HIST	EUCL	7.27	8.58	7.50	8.04
	NOC	7.47	8.38	7.00	**7.69**
ZMUV	EUCL	8.60	9.87	9.50	9.68
	NOC	8.60	9.87	9.50	9.68

Table 4.1: *Matching in the image space: verification results on the XM2VTS database using a global threshold.*

with ZMUV normalisation, the Euclidean distance and normalised correlation give the same results. Because of Equation (2.10), the two score functions are completely equivalent. Likewise, in the case of histogram equalisation there is almost no difference between the two matching schemes.

Configuration I

Photo.	matching	Validation EER	Test FAR	Test FRR	Test HTER
HIST	EUCL	6.75	8.75	5.00	6.87
	NOC	6.67	8.12	5.50	**6.81**
ZMUV	EUCL	7.91	9.56	5.50	7.53
	NOC	7.91	9.56	5.50	7.53
Configuration II					
HIST	EUCL	4.48	6.36	3.50	4.93
	NOC	4.50	5.63	3.75	**4.69**
ZMUV	EUCL	7.32	8.58	7.50	8.04
	NOC	7.32	8.58	7.50	8.04

Table 4.2: *Matching in the image space: verification results on the XM2VTS database with Z-normalised scores*

In contrast, there is an advantage of using histogram equalisation over ZMUV. Also, the Z-normalisation helps in reducing the error rates.

4.2.4 Eigenface based face verification

As described in Chapter 2, in the eigenface or PCA face verification approach, the classification is performed in the principal subspace of faces. This subspace is obtained on the 600 (800) training images of the XM2VTS database in configuration I (configuration II). Its dimensionality is therefore at most 599 (799) (see Section 2.3.1). For choosing the dimensionality m of the PCA subspace we choose the value of m that minimises the EER on the validation set. The variation of the EER with respect to the PCA dimensionality is shown in Figure 4.1 for the various matching schemes and photo-normalisation techniques presented above. In all cases, the EER decreases very fast with the first 50 eigenvectors and either it stabilises above 50 or increases. It can be seen that, as for the matching in the image space, the histogram equalisation gives better results over the pixel value reduction. This time the normalised correlation outperforms the Euclidean distance.

The error obtained using the Mahalanobis distance decreases up to 50-70 eigenvectors before increasing dramatically. When two vectors are compared using Mahalanobis distance, each component difference is weighted by the inverse of the eigenvalue λ_i (see Equation (2.6)). As the

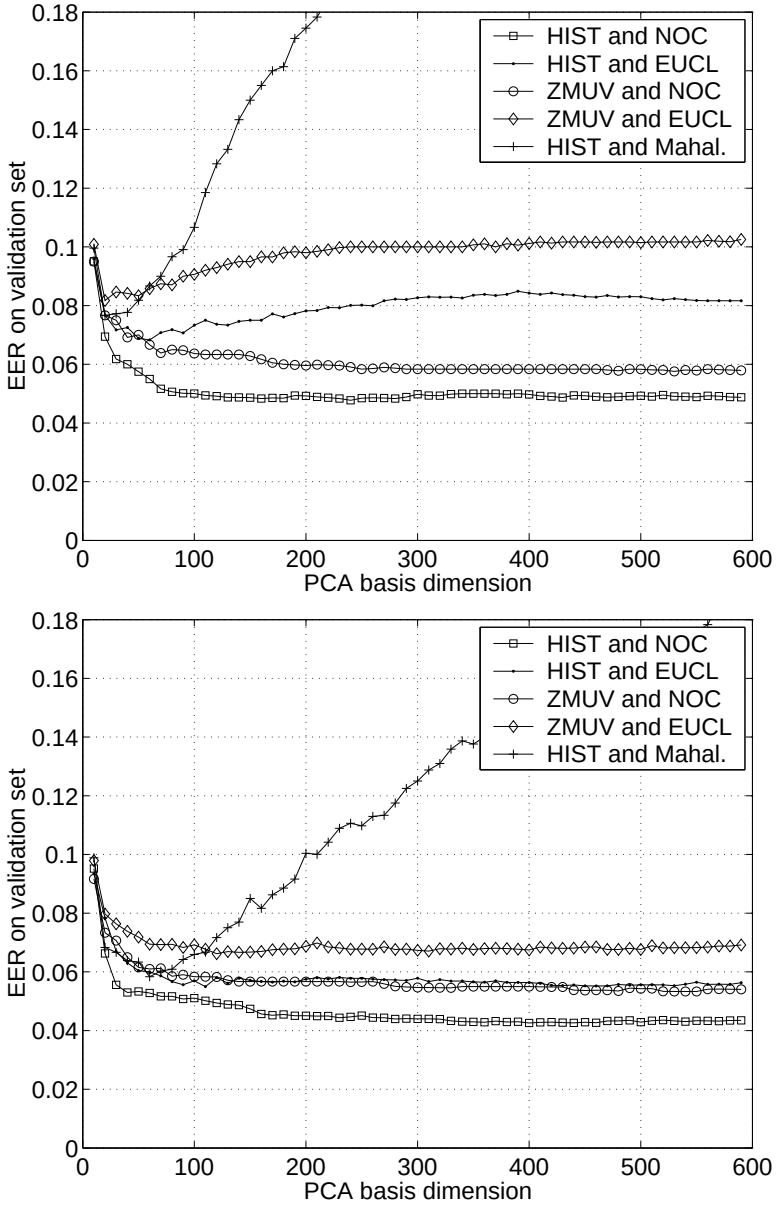


Figure 4.1: *EER on the XM2VTS configuration I as a function of the PCA subspace dimension, for various matching schemes. A global threshold (above) and Z-normalised scores (below) have been used*

eigenvectors are ranked by decreasing eigenvalues, when the eigenvector index i increases, the corresponding component difference is more and more amplified. As an example in this experiment $\frac{1}{\lambda_{100}} \approx 70 \times \frac{1}{\lambda_1}$ and $\frac{1}{\lambda_{200}} \approx 160 \times \frac{1}{\lambda_1}$. It is likely that the eigenvectors corresponding to large indexes ($i > 100$) encode mainly noise that is overfitted to the training data. As this noise is amplified by the Mahalanobis distance, the error rate increases with the eigenvector index i.

Interestingly, although the PCA approximates the face images with small reconstruction error, it seems that making the verification in the PCA subspace is more advantageous than in the image space. It is like if the reconstruction error, which has been removed by the PCA process, is perturbing the decision boundary when making the verification in the image space. The EER when using Euclidean distance and global threshold decreases quite fast with the first eigenvectors, before steadily increasing after $m = 50$. Interestingly the results using Euclidean distance are significantly improved when scores are Z-normalised. In contrast, the normalised correlation is almost equally efficient with global and specific thresholds. The results on the test set, using the PCA dimensionality that

Configuration I

Photo.	matching	Valid. EER	Test FAR	Test FRR	Test HTER	PCA dim.
HIST	EUCL	5.50	5.48	5.00	5.24	110
	NOC	4.26	4.52	4.25	**4.39**	400
ZMUV	EUCL	6.62	7.46	5.50	6.48	120
	NOC	5.37	5.47	4.25	4.86	450
Configuration II						
HIST	EUCL	3.75	4.54	2.75	3.65	160
	NOC	4.00	4.65	2.25	**3.45**	270
ZMUV	EUCL	6.31	7.09	4.45	5.92	380
	NOC	6.53	7.48	5.00	6.24	230

Table 4.3: *Influence of the matching distance technique on the verification rate using different metric in the PCA subspace (Z-normalised scores and configuration I).*

minimises the EER on the validation set, are shown in Table 4.3. Only Z-normalised scores are shown because they give consistently better results. The influence of the photo-normalisation on the rates is similar to the tem-

plate matching case. Note that the normalised correlation requires a significantly larger number of eigenvectors for optimal performance. Also, it is worth mentioning that in table 4.3 the test error is often inferior to the validation error. This shows that, by chance, the test data are easier to discriminate using eigenfaces than the validation data. To avoid this effect, the database can be randomly resampled (see Section 4.2.5).

4.2.5 LDA-based face verification

The performance can be further improved by projecting the PCA face vectors on a subspace were discrimination is enhanced. This is done by Linear Discriminant Analysis which is described in Chapter 2. The training set is used to estimate the within S_w and between class S_b scatter matrices. As it contains only 200 subjects, there are 200 classes. Therefore the rank of S_b is 199, which is the maximal dimensionality of the LDA subspace. Since the rank of S_w is 400 (600 in configuration II) and S_w cannot be singular (otherwise the whitening in Equation (2.9) is problematic), the PCA intermediate dimensionality is 400 (600 in configuration II) at the maximum.

As explained in Chapter 2, the selection of PCA intermediate dimensionality and LDA dimensionality is a difficult problem. We have adopted a pragmatic approach to determine these numbers: they are selected so as to minimise the EER on the validation set. Both LDA and PCA basis vectors are ranked according to the decreasing magnitude of their corresponding eigenvalue. The magnitude of the LDA eigenvalue is an indicator of the discriminatory "power" of the corresponding eigenvector. Figure 4.2 shows the variation of the EER as a function of the intermediate PCA dimensionality and the LDA dimensionality for the normalised correlation and the Euclidean distance as well as the two photo-normalisation techniques.

Again the histogram equalisation gives better results than ZMUV. The difference between normalised correlation and Euclidean distance is more difficult to see. Normalised correlation shows better rates in combination with ZMUV photo-normalisation while Euclidean distance shows better rates with HIST.

It can be seen from the figure that the EER is not a very smooth function of the two parameters. Choosing the global minimum may not be robust. Instead we find the minimum by visually observing large variation trends on the bi-dimensional EER plot. The selected PCA and LDA basis vector numbers is reported in Table 4.4 with the corresponding errors. An important point to mention is that the best performing matching scheme

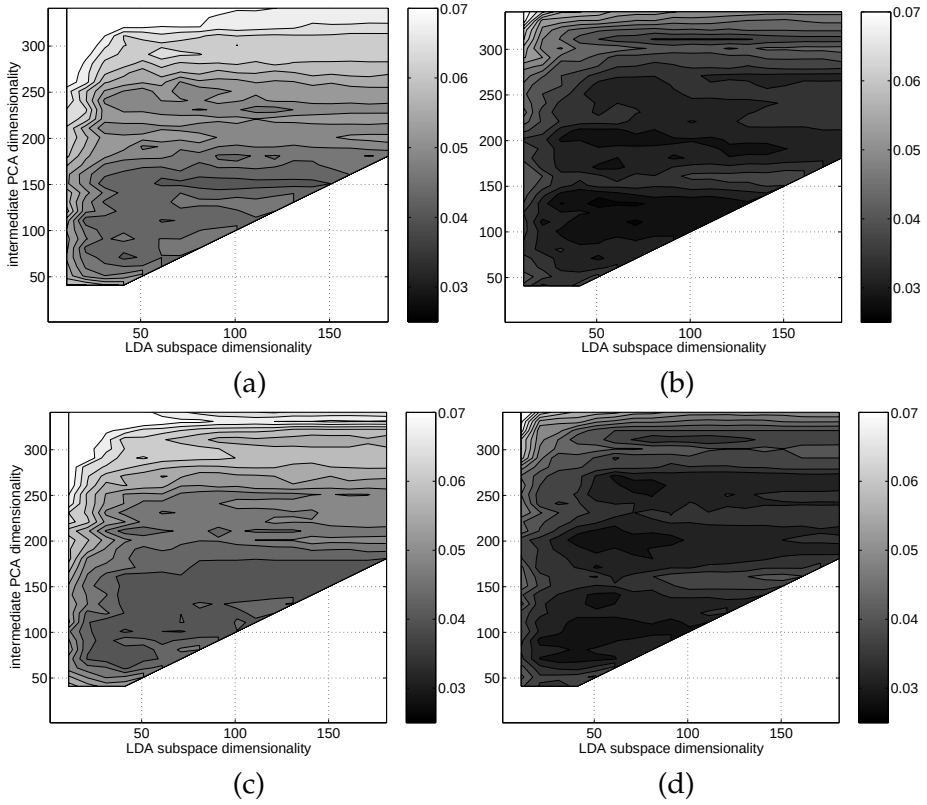


Figure 4.2: *EER on the XM2VTS configuration I as a function of the PCA intermediate dimensionality. (a) ZMUV and Eucl. distance. (b) HIST and Eucl. distance. (c) ZMUV and normalised correlation (d) HIST and normalised correlation. Z-normalised scores have been used in the four cases.*

is the Euclidean distance on the validation set, while it is the normalised correlation on the test set.

Configuration I

Photo.	matching	Valid. EER	Test FAR	Test FRR	Test HTER	LDA dim.	PCA dim.
HIST	EUCL	2.67	3.53	3.50	3.52	60	120
	NOC	2.93	3.83	2.75	**3.29**	60	140
ZMUV	EUCL	4.00	5.37	4.00	4.68	90	160
	NOC	3.67	4.51	4.25	4.38	100	150

Configuration II

Photo.	matching	Evaluation EER	Test FAR	Test FRR	Test HTER	LDA dim.	PCA dim.
HIST	EUCL	2.25	3.51	2.00	2.75	100	200
	NOC	2.04	2.94	1.75	**2.35**	130	240
ZMUV	EUCL	3.13	3.93	3.00	3.47	90	230
	NOC	2.50	3.40	2.25	2.82	160	230

Table 4.4: *Influence of the matching distance technique on the verification rate using different metric in the LDA subspace (XM2VTS database).*

Resampling the XM2VTS database

Although the Lausanne protocol specifies a large number of client and impostor matchings, we can increase this number by randomly resampling the training, validation and test set. Then, the algorithm can be re-trained and error rates re-computed on the resampled datasets. This procedure can be repeated several times. It ensures that the results do not depend on a particular choice of training, validation and test subjects.

The LDA algorithm has been evaluated by repeating 25 times the following procedure

1. Resample the XM2VTS database randomly in 200 clients, 25 validation impostors and 70 test impostors. Database recordings are allocated according to Lausanne protocol in configuration I.

2. Train the LDA algorithm on the training set. Use the validation set to find the threshold and LDA algorithm parameters (PCA and LDA dimensionality).

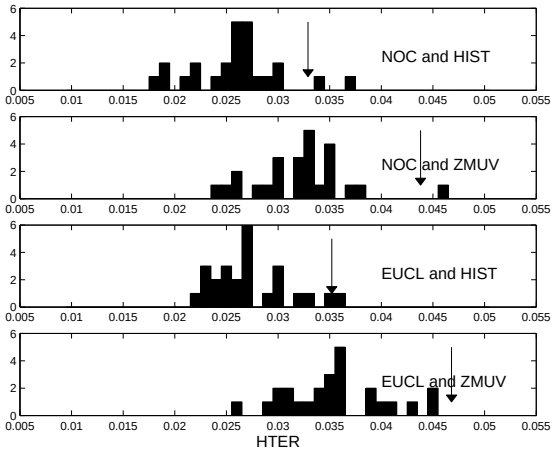


Figure 4.3: *Histogram of HTER values for the LDA algorithm obtained by resampling several times the XM2VTS database.*

3. Compute the HTER using the test set.

Figure 4.3 shows the histograms of the 25 values of HTER obtained using normalised correlation and Euclidean distance in combination with either ZMUV or HIST photo-normalisation. The arrow in the figure indicates the HTER obtained using the Lausanne protocol client-impostor partitioning. Interestingly, the HTER is most of the time better for other partitionings than the Lausanne partitioning. It suggests that the Lausanne client impostor partitioning is, by chance, unusually difficult.

The resampling experiments comfirm the ranking of the rates listed in Table 4.4. The LDA algorithm combined with normalised correlation and histogram equalised images gives the lowest error rates (average HTER = 2.60%). It is immediatly followed by Euclidean distance (average HTER = 2.74%). Using ZMUV normalisation, although rates are higher, the normalised correlation (average HTER = 3.22%) does outperform the Euclidean distance (average HTER = 3.55%).

4.2.6 Probabilistic matching face verification

Probabilistic matching (see Chapter 2) has been proposed by Moghaddam *et al.* [86] as an alternative to LDA. It requires the estimation of an intraclass and an interclass eigen-subspace. Both subspaces are estimated using difference images $\mathbf{\Delta}$ in the training set of the database. As training

contains 3 images per person for 200 person (4 in configuration II), we can generate 600 (800) intraclass difference images and much more interclass images. To keep a reasonable computational time, we randomly select 600 (800) interclass difference images. The intraclass and interclass eigen-subspaces are then estimated using the 600 (800) intraclass and interclass difference images respectively.

From Section 2.4.3 , four parameters need to be adjusted: ρ_E, ρ_I and the eigenvector numbers m_I, m_E for the intra- and inter-personal PCA. The optimal values of these by parameters were found so as to minimise the EER on the validation set. Then the optimal parameter values are used to assess generalisation performance on the test set.

To find the minimal EER, we performed a coarse exhaustive search in the parameter space to find an approximate global minimum. Then we used these values as initialisation of the Nelder-Mead simplex [100] which finds the local minimum in this area of the parameter space. Note that the EER is very sensitive to ρ_i and ρ_e.

Results are given in Table 4.5 for the two photo-normalisation techniques. Again, histogram equalisation allows lower error rates, but the

Configuration I

Photo	Valid.	Test			Parameters			
norm.	EER	FAR	FRR	HTER	m_I	m_E	ρ_I	ρ_E
ZMUV	4.50	4.89	4.75	4.82	52	59	6.7	7.39
HIST	3.43	4.71	3.00	3.86	40	65	0.1037	0.098

Configuration II

Photo	Valid.	Test			Parameters			
norm.	EER	FAR	FRR	HTER	m_I	m_E	ρ_I	ρ_E
ZMUV	4.50	5.70	3.75	4.73	40	160	1	1
HIST	2.50	3.65	2.50	3.08	175	99	0.3159	0.3925

Table 4.5: *Probabilistic matching results for different normalisations on XM2VTS database.*

HTER is slightly higher in comparison with the HTER obtained using LDA and normalised correlation. Note that the values of ρ_I and ρ_E are much smaller in the case of HIST photo-normalisation. This is because in the ZMUV case, the similarity measure is $O(10^2)$, while it is $O(5.10^4)$ in the HIST case. The drawback of the probabilistic matching algorithm is that its optimisation is very time consuming. Furthermore, a wrong choice of

parameters leads to high error rates.

4.2.7 Comparison with benchmark results

As we use a standard protocol for our experiments, we can compare the results with other approaches. Table 4.6 lists some of the best error rates in configuration I, obtained recently by other researchers during a face verification contest organised by the University of Surrey (UK). The complete results of the contest can be found in [81]. The first three lines show error

Approach (Institution)	Test (Config. I)		
	FAR	FRR	HTER
Neural Net. (IDIAP)	1.75	2.00	1.88
Shape Trace Trans. (MUT)	0.97	0.50	0.74
ECOC (UniS)	0.86	0.75	0.81
LDA (UCL)	3.83	2.75	3.29
Intramodal comb. (UCL)	1.12	1.25	1.31

Table 4.6: *Face verification contest results*

rates obtained by different institutions using different approaches. In the Neural Network approach proposed by IDIAP, the face image is used as an input to an MLP. The MLP is trained to discriminate clients from impostors. In the approach proposed by MUT, faces are represented using the Shape Trace Transform. The Hausdorff context, a dissimalirity measure related to the Hausdorff distance, is introduced to measure similarity between two shapes. In the UniS approach, an Error Correcting Output Codes (ECOC) classifier is used to separate client from impostor. The last line shows the error rates obtained in this work using intramodal combination (see Chapter 5).

4.2.8 Reduced image size

In this experiments, we are interested in evaluating the influence of the image size on verification rates. One way to reduce the image size is to filter and down-sample the images. By doing so, the high frequency components of the face images are removed and possibly discriminative information is lost. We thus expect the verification performance to be worse with low-resolution than with high-resolution images. We have down-sampled the face images from 64×64 pixels to 8×8 in three steps. Before

Figure 4.4: *Examples of images from the XM2VTS database at different resolutions.*

down-sampling the images are filtered using a 11 by 11 low pass FIR filter. Examples of images from the XM2VTS database at these resolutions are shown in Figure 4.4. Clearly, for a human observer the full resolution images provide much more discriminative information.

Verification results at various resolutions are shown in Table 4.7 for template matching and eigenface methods. In both cases, the images were histogram equalised and normalised correlation was used for matching. For the LDA-based verification, the results obtained are shown in Table 4.8. For each image resolution, results using two matching measures in the LDA subspace, namely Euclidean distance and normalised correlation, are listed.

size	Method	Valid. EER	Test FAR	Test FRR	Test HTER	PCA dim.
64x64	Temp.match.	6.67	8.12	5.50	6.81	
	Eigenface	4.26	4.52	4.25	4.39	400
32x32	Temp.match.	6.50	7.56	4.75	6.15	N/A
	Eigenface	4.75	4.87	4.25	4.56	450
16x16	Temp.match.	6.61	7.27	4.25	6.10	N/A
	Eigenface	4.14	4.42	5.75	5.09	170
8x8	Temp.match.	9.74	9.71	8.75	8.98	N/A
	Eigenface	12.36	11.92	6.75	9.33	64

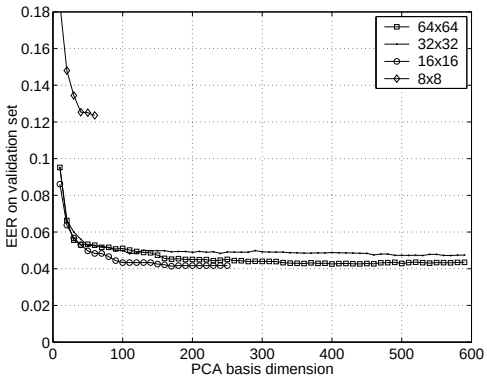
Table 4.7: *Template matching and eigenface error rates on the XM2VTS database (configuration I) at several image resolutions. Histogram equalisation and normalised correlation were used.*

size	matching	Valid. EER	Test FAR	Test FRR	Test HTER	LDA dim.	PCA dim.
64x64	EUCL	2.67	3.53	3.50	3.51	60	120
	NOC	2.93	3.89	2.75	3.32	60	140
32x32	EUCL	3.01	3.57	3.50	3.53	70	100
	NOC	3.07	3.65	3.25	3.45	40	110
16x16	EUCL	3.50	4.28	3.35	3.77	40	110
	NOC	3.31	3.97	2.50	3.24	40	130
8x8	EUCL	6.33	8.06	22.00	15.03	20	40
	NOC	3.83	4.59	7.25	5.92	40	40

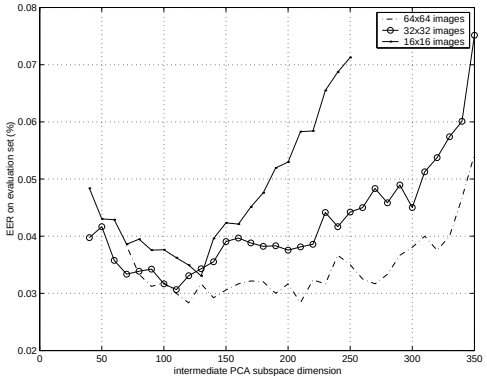
Table 4.8: *Verification error rates on the XM2VTS database (configuration I) using several image resolutions and two matching measures: Euclidean distance (EUCL) and normalised correlation (NOC). The LDA and the PCA subspace dimensions are also listed.*

From the Tables 4.7 and 4.8, we can see that the performance stays almost unchanged for the first three resolutions. The degradation appears only when processing 8×8 images. Nevertheless, the total error rate on 8×8 images when using normalised correlation is 5.77% which is surprisingly good with respect to the information that these images carry (see Figure 4.4). In fact it is better than the baseline template matching technique which is give an half total error rate of about 7% using 64×64 images. Note that for the first three resolutions, the intermediate PCA subspace dimension, leading to the best results on the evaluation set, is more or less constant. Figure 4.5(a) shows the variation of the EER on the validation set versus the number of PCA basis vectors at various resolution. It is clear that the verification performance are more of less constant down to 16×16 images. It is only with 8×8 images that the performance degrades.

Figure 4.5(b) shows the variation of the EER on the validation set vs. the intermediate PCA dimension and for 3 different image resolutions. It appears that the minimum EER is always located in the same region. At higher resolution the admissible range for the PCA intermediate dimension (leading to a low EER) is larger than at lower resolution. Note also that the verification results we obtained with these three image resolutions are more or less constant.



(a)



(b)

Figure 4.5: *(a) Eigenfaces: EER versus PCA dimensionality and image size (b) LDA: EER versus PCA intermediate subspace dimension and image size.*

4.2.9 PCA on subband images

In this section, we show that the PCA representations of a face and its half resolution version are very similar. This explains the constancy of verification rates across different resolutions observed above. For simplicity of the notations, the discussion is done for one-dimensional signals only.

From subband coding, it is well known that a one-dimensional signal can be decomposed in a *low-pass version* and a *high pass version* having both half the size of the original signal. In the two channel subband decomposition case, a half-resolution version of the signal is obtained by filtering and two-fold downsampling. The filtering followed by two-fold down-

sampling can be expressed in matrix notation as

$$\mathbf{x}_l = H\mathbf{x},$$
$$\mathbf{x}_h = G\mathbf{x},$$

where H and G are $\frac{n}{2} \times n$ matrices with the following structure

$$H = \begin{pmatrix} & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \dots & h(L-1) & \dots & h(2) & h(1) & h(0) & \dots \\ \dots & 0 & 0 & h(L-1) & \dots & h(2) & \dots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix},$$

and

$$G = \begin{pmatrix} & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \dots & g(L-1) & \dots & g(2) & g(1) & g(0) & \dots \\ \dots & 0 & 0 & g(L-1) & \dots & g(2) & \dots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix},$$

where L is the filter length. By writing the filtering and downsampling in this way, it is clear that this operation is a linear projection from $\mathbb{R}^n$ to $\mathbb{R}^{n/2}$. Let matrix A be

$$A = \begin{pmatrix} H \\ G \end{pmatrix}.$$

The linear transformation of $\mathbf{x}$ by matrix A leads to a new representation of $\mathbf{x}$

$$\tilde{\mathbf{x}} = A\mathbf{x} = \begin{pmatrix} \mathbf{x}_l \\ \mathbf{x}_h \end{pmatrix}.$$

If the decomposition filters are chosen so that they constitute a orthonormal (or paraunitary) filter bank, the new representation $\tilde{\mathbf{x}}$ is strictly equivalent to $\mathbf{x}$ in the sense that the original $\mathbf{x}$ can be recovered by upsampling and filtering using time-reversed version of the decomposition filters $h(n)$ and $g(n)$. In matrix notation this yields

$$\mathbf{x} = A^T\tilde{\mathbf{x}},$$

or $A^TA = I$, which shows that A is othornormal.

The covariance matrix $\tilde{\Sigma}$ in the new representation becomes

$$\tilde{\Sigma} = A\Sigma A^T,$$

or

$$\tilde{\Sigma} = \begin{pmatrix} \Sigma_L & \Sigma_{HL} \\ \Sigma_{HL}^T & \Sigma_H \end{pmatrix}$$

where

$$\begin{aligned} \Sigma_L &= E[(\mathbf{x}_L - \bar{\mathbf{x}}_L)(\mathbf{x}_L - \bar{\mathbf{x}}_L)^T], \\ \Sigma_H &= E[(\mathbf{x}_H - \bar{\mathbf{x}}_H)(\mathbf{x}_H - \bar{\mathbf{x}}_H)^T], \\ \Sigma_{HL} &= E[(\mathbf{x}_H - \bar{\mathbf{x}}_H)(\mathbf{x}_L - \bar{\mathbf{x}}_L)^T]. \end{aligned}$$

Using the eigen-decomposition if Σ, we have

$$\tilde{\Sigma} = A\Phi\Lambda\Phi^T A^T,$$

which shows that the eigenvectors of $\tilde{\Sigma}$ are the columns of $A\Phi$ since $(A\Phi)^{-1} = \Phi^T A^T$. Let $\tilde{\boldsymbol{\phi}}_j$ be the jth column of $A\Phi$, we have

$$\tilde{\boldsymbol{\phi}}_j = \begin{pmatrix} \boldsymbol{\phi}_{Lj} \\ \boldsymbol{\phi}_{Hj} \end{pmatrix} = A\boldsymbol{\phi}_j.$$

Since $\tilde{\boldsymbol{\phi}}_j$ is an eigenvector of $\tilde{\Sigma}$ we have

$$\begin{pmatrix} \Sigma_L & \Sigma_{HL} \\ \Sigma_{HL}^T & \Sigma_H \end{pmatrix} \begin{pmatrix} \boldsymbol{\phi}_{Lj} \\ \boldsymbol{\phi}_{Hj} \end{pmatrix} = \lambda_j \begin{pmatrix} \boldsymbol{\phi}_{Lj} \\ \boldsymbol{\phi}_{Hj} \end{pmatrix},$$

which yields

$$\Sigma_L\, \boldsymbol{\phi}_{Lj} + \Sigma_{HL}\, \boldsymbol{\phi}_{Hj} = \lambda_j \boldsymbol{\phi}_{jL}. \tag{4.1}$$

From Section 2.3.1 we know that the eigenvectors $\boldsymbol{\phi}_j$ corresponding to largest eigenvalues, i.e. for j smaller than a certain indice m_0, encode the part of the signal with maximal energy. Since energy is located mainly in low frequencies for images, these eigenvectors contain mainly low frequencies. Consequently, we can assume that the high pass version of these eigenvectors $\boldsymbol{\phi}_{Hj} \approx 0$ for $j \leq m_0$. Therefore Equation (4.1) becomes

$$\Sigma_L\, \boldsymbol{\phi}_{Lj} \approx \lambda_j \boldsymbol{\phi}_{Lj}, \quad \text{for } j \leq m_0,$$

which shows that the eigenvectors of Σ_L (i.e. the PCA basis of the low resolution faces) corresponding to the m_0 highest eigenvalues, can be approximated with the low resolution versions $\boldsymbol{\phi}_{Lj}$ of the original eigenvectors $\boldsymbol{\phi}_j$. Note that the $\boldsymbol{\phi}_{Lj}$'s must be first normalised since they do not satisfy the condition $\|\boldsymbol{\phi}_{Lj}\| = 1$. Note also that the eigenvalues are approximatively the same for original and low resolution images. The eigenvector ranking is thus the same for $j \leq m_0$.

This explains the similarity between the first eigenvectors depicted on Figures 4.6 and 4.7.

Furthermore, the PCA representation $\tilde{\mathbf{y}}$ of $\tilde{\mathbf{x}}$ is equal to the PCA representation $\mathbf{y}$ of the original face image $\mathbf{x}$ since

$$\tilde{\mathbf{y}} = \tilde{\Phi}^T \tilde{\mathbf{x}} = \Phi^T A^T A \mathbf{x} = \mathbf{y}.$$

Thus the jth component y_j of $\mathbf{y}$ can be written

$$y_j = \mathbf{x}_L^T \, \boldsymbol{\phi}_{Lj} + \mathbf{x}_H^T \, \boldsymbol{\phi}_{Hj}.$$

Assuming again that $\boldsymbol{\phi}_{Hj} \approx 0$ for $j \leq m_0$ we have

$$y_j \approx \mathbf{x}_L^T \, \boldsymbol{\phi}_{Lj}.$$

This last equation shows that, up to a scaling factor due to eigenvector normalisation, the PCA representation of $\mathbf{x}$ and its low resolution version $\mathbf{x}_L$ are very similar as long as a small number of eigenvectors is used. Figure 4.8 illustrates this phenomenon using a set of $N_0 = 600$ face images of the XM2VTS database. The solid curve in the figure shows the mean square error (MSE) between the original PCA representation and the low resolution PCA representation versus the eigenvector index j. The MSE has been defined as

$$\frac{1}{N_0} \sum_{i=1}^{N_0} \frac{(y_{ij} - \mathbf{x}_{Li}^T \, \boldsymbol{\phi}_{Lj})^2}{y_{ij}^2},$$

where y_{ij} is the jth principal component of face image $\mathbf{x}_i$, and $\mathbf{x}_{Li}$ is the low resolution version of $\mathbf{x}_i$. The dashed curve in the figure shows the MSE when the principal components y_{ij} of the face image $\mathbf{x}_i$ and the principal components of the low resolution image $\mathbf{x}_{Lk}$ correspond to different images, i.e.

$$\frac{1}{N_0} \sum_{i=1}^{N_0} \frac{(y_{ij} - \mathbf{x}_{Lk}^T \, \boldsymbol{\phi}_{Lj})^2}{y_{ij}^2}, \quad \text{with } i \neq k.$$

We will refer to this number as *average distance* between distinct face principal components. From Figure 4.8, it can be seen that the average distance (dashed curve) is more or less constant for all eigenvector indexes. For low eigenvector indexes, the MSE between low resolution and original image principal components is small in comparison with the average distance between principal components. This is in accordance with Equation 4.1. It shows that verification rates using low resolution or original images are similar because face principal components do not differ very much as long as a small number ($m < 100$) of eigenvectors is used for face representation.

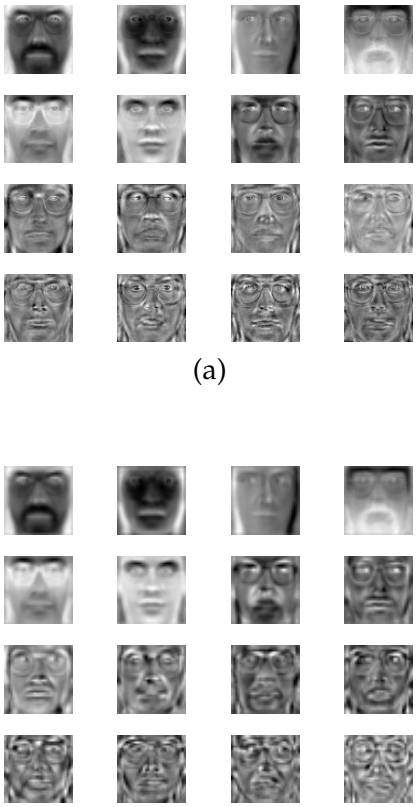


Figure 4.6: *Eigenfaces number 1,2,3,4,5,6,10,20,30,40,50,60,70,80,90,100,110 at resolution 64 by 64 (a) and 32 by 32 (b).*

4.2.10 Direct LDA

Results using 16 $\times$ 16 images suggest that there is enough information in these small images to get similar error rates than with 64 $\times$ 64 images. With 16 $\times$ 16 images, it is no longer mandatory to project the images on a PCA subspace before applying LDA in order to get a full rank matrix S_w. Indeed, the rank of S_w is equal to the number of training images minus the number of classes thus 400 in our case. In the case of 16 $\times$ 16 images, the image space dimensionality is 256 which means that LDA can be applied in the image space directly. However, such system lead poor results (EER of around 8%). Moreover, the eigenvectors obtained with direct LDA don't have the aspect of faces as for the "indirect" LDA. It seems that direct

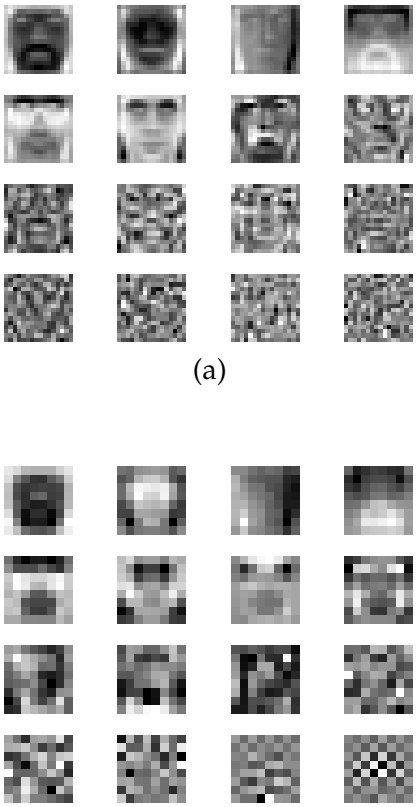


Figure 4.7: *Eigenfaces number 1,2,3,4,5,6,10,20,30,40,50,60,70,80,90,100,110 at resolution 16 by 16 (a) and number 1,2,3,4,5,6,7,8,9,10,20,30,40,50,60,64 at resolution 8 by 8 (b) .*

LDA extracts non relevant information like noise instead of facial feature. We conclude that PCA is not only a preprocessing stage allowing a full rank S_w matrix, but a mandatory step that extracts meaningful features for subsequent classification.

4.3 Experiments on the BANCA database

4.3.1 Introduction

In contrast to the XM2VTS database, the BANCA database images contain a lot of variability mainly due to illumination, user pose and camera qual-

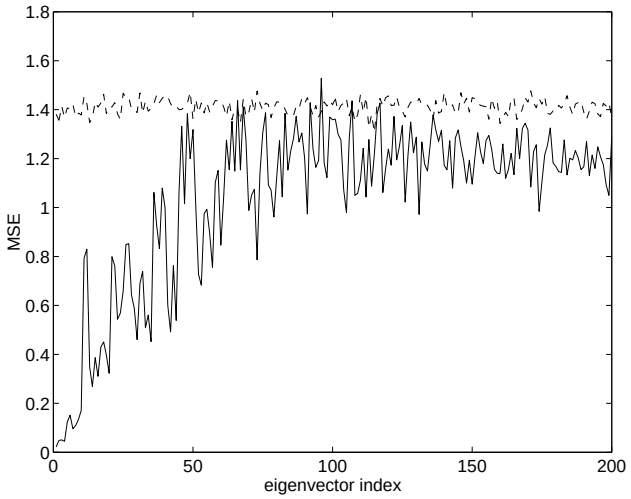


Figure 4.8: *Mean Square Error (MSE) between the original PCA representation and the low resolution PCA representation versus the eigenvector index j for the same image (solid curve) and different images (dashed curve).*

ity. The other main difference with the XM2VTS is the very small amount of training data available to create the user template. This prevents the computation on the face subspace basis and the LDA basis on the BANCA images themselves. We must use an external dataset to obtain the face subspace. The drawback is that it is hard to find a face dataset with similar image acquisition conditions as for the BANCA. The representation capability of the face subspace found using the external dataset is lower.

As a result, the verification results presented in this section are more realistic, but also worse than for XM2VTS database. We first present the result of our study on images with manually detected eye coordinates. Results obtained with fully automatic system are given in Chapter 7 for two different face localisation algorithms.

From the experience gained on the XM2VTS database, we know that HIST photometric normalisation is consistently superior to the ZMUV normalisation. In this section, we exclusively use the histogram equalisation. Due to the small number of impostor matches in the BANCA protocol, Z-normalisation or client specific thresholds are not possible anymore. Recall that the normalised correlation has shown quite good results independently of the threshold technique, which clearly favours the nor-

malised correlation in the BANCA database case.

Each access attempt of the BANCA database consists of 5 frames selected randomly. The 5 scores generated by the 5 frames must be somehow integrated to generate the decision for one access. This topic is treated in Chapter 7. In the experiments presented in this section, the decision is based on the frame the gives the minimum score. Also, as for the XM2VTS experiments, if several training images are available, the score comes from the best match between the test image and the training images.

According to the BANCA testing protocol, the threshold and the algorithm parameters are first adjusted on group g1, so as to be at equal error rate (EER(g1)). The same parameters are then used on g2, leading thus to an *a priori* false acceptance and false rejection rates (FAR(g2) and FRR(g2)), reflecting faithfully the expected performance of a real system. The same procedure is carried out after swapping g1 and g2, leading then to EER(g2), FAR(g1) and FAR(g1).

4.3.2 Matching in the image space

In order to compare the difficulty of the BANCA database to the XM2VTS, we used the same baseline face verification algorithm as defined in Section 4.2.3. Results are reported in Table 4.9 for several metrics and the two major sub-protocols of the BANCA database. Detailed results for the other protocols can be found in Appendix C. The last column of the table shows the average HTER on the two groups g1 and g2.

Clearly the baseline algorithm give much worse results than for the XM2VTS database. Even in the controlled sub-protocol Mc the HTER is around 12% as it is 6% for the XM2VTS database. As expected, when the environmental conditions in the enrolment do not match the conditions in testing, as it is the case in sub-protocols Ud and Ua, the results are very poor. The most realistic sub-protocol is the protocol P as it allows only one session for creating the user template, and testing is done in the three scenarios. The best result in this sub-protocol is about 25% which is of course unacceptable for the majority of real world applications. Note that manually located eye coordinates were used in this experiments, worse results are therefore expected on automatically located faces. Note that as in the experiments on XM2VTS database, the normalised correlation matching leads to almost always the best results. The nearest mean metric uses the world model images (100 images of 20 persons recorded in controlled condition) to create the impostor class mean. It shows in general slightly better results than Euclidean distance but it is not as good as normalised

Prot.	Metric	FAR(g2)	FRR(g2)	FAR(g1)	FRR(g1)	HTER
G	NOC	13.46	21.37	19.87	19.87	**16.35**
	EUD	13.46	35.47	30.13	30.13	22.33
	NMM	14.10	31.20	29.17	29.17	20.75
P	NOC	17.95	31.62	30.13	30.13	**25.48**
	EUD	21.47	34.62	30.45	30.45	26.66
	NMM	28.85	36.75	33.97	33.97	32.05

Table 4.9: *Matching in the image space: BANCA database (English subset), rates for the different configuration and for normalised correlation (NOC), Euclidean distance (EUD) and nearest mean metric (NMM).*

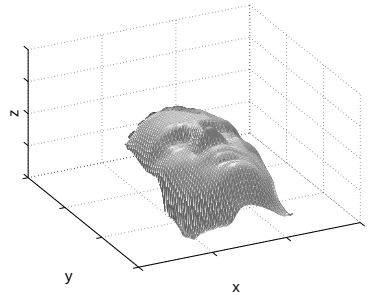


Figure 4.9: *The surface* $z = z(x, y)$

correlation.

4.3.3 Exploiting face symmetry for illumination

Face can be seen as a symmetric 3-dimensional surface. It may be described by a function $z = z(x, y)$ where the xy-plane is vertical and z axis is pointing towards the camera as shown in Figure 4.9. Under the hypothesis of Lambertian surface, the image $I(x, y)$ obtained from surface z under a single distant light source is [125]

$$I = \rho \cos \theta,$$

where $\rho = \rho(x, y)$ is the albedo of the surface and θ the angle between the surface normal and the illumination direction $\vec{L} = (-P, -Q, 1)$. This equation can be explicited using the fact that the direction of the surface

Prot.	Metric	FAR(g2)	FRR(g2)	FAR(g1)	FRR(g1)	HTER
G	NOC	10.90	17.95	15.06	15.06	**13.22**
	EUD	10.58	26.07	20.83	20.83	16.40
	NMM	16.67	26.92	24.68	24.68	20.70
P	NOC	15.38	29.91	26.92	26.92	**22.44**
	EUD	18.91	28.21	25.64	25.64	23.32
	NMM	19.23	30.34	25.96	25.96	23.90

Table 4.10: *Matching in the image space: BANCA database (English subset), rates for the different configuration and for normalised correlation (NOC), Euclidean distance (EUD) and nearest mean metric (NMM). The images were symmetrised before matching.*

normal is $(p, q, 1)$ where $p = \frac{\partial z}{\partial x}$ and $q = \frac{\partial z}{\partial y}$.

$$I = \rho \frac{1 + pP + qQ}{\sqrt{1 + p^2 + q^2}\sqrt{1 + P^2 + Q^2}}.$$

Thanks to face symmetry we have

$$\begin{aligned} z(-x, y) &= z(x, y), \\ p(-x, y) &= -p(x, y), \\ q(-x, y) &= q(x, y). \end{aligned}$$

Using these equations we can compute the symmetrised image I_s

$$I_s = \frac{I(x, y) + I(-x, y)}{2} = \rho \frac{1 + qQ}{\sqrt{1 + p^2 + q^2}\sqrt{1 + P^2 + Q^2}}. \quad (4.2)$$

Clearly, I_s is, apart from a multiplicative constant, the image of surface z obtained with the illumination direction $\vec{L} = (0, -Q, 1)$. This means that symmetrising the face image is equivalent to removing the effect of azimuthal illumination direction (component P of $\vec{L}$). The technique can be therefore used to normalise face images and obtain robustness against azimuthal illumination direction.

Symmetrised images according to Equation (4.2) were used to obtain verification results in Table 4.10. An improvement of about 3% is observed in most cases.

4.3.4 Matching in the LDA subspace

From the experimental results on the XM2VTS database, we observed that the lowest error rates were obtained by normalised correlation matching in the LDA subspace. In this section we study the performance of LDA on the BANCA database.

The BANCA protocol specifies that the user templates must be built independently of each others (see Chapter 3). This means that we are not allowed to use the training set to compute the PCA and LDA subspaces. In fact, as we are using the English subset, the amount of data available for training is too small to obtain a meaningful face subspace. Therefore, we used the complete XM2VTS database (295 persons, 8 images per person) to compute the PCA and LDA bases. This leads to a 294 LDA basis vectors at the most. Table 4.11 shows the results for the two major sub-protocols. Results for the other protocols can be found in Appendix C. Note that all the images, including the XM2VTS images needed for LDA subspace computation, were symmetrised. Because images are symmetric, half of the image is sufficient for the matching process. Using half image reduces obviously the computational cost by a factor two. It also reduces the dimensionality of the initial image space and therefore doubles the sample number to dimensionality ratio, which is favourable for classification.

As for the XM2VTS experiments, it can be seen that the representation of faces in the LDA subspace allows significantly lower error rates than in the images space. Although computed on different images of different people, the LDA representation of faces is somehow more robust to image variability. This is in accordance with how the LDA subspace is constructed: in the LDA subspace the intraclass variation are reduced while the interclass variation are maximised (see Chapter 2). However, the matching using the Euclidean distance is worse in the LDA subspace than in the image space. This is due to the fact that the LDA basis vector are scaled so that the intraclass covariance matrix is unity *for the XM2VTS users*. It is likely that this scaling is inadequate for BANCA users. The normalised correlation, which does not take the magnitude of the basis vectors into account, is superior in this case. It shows that the possible advantage of the LDA representation is determined by the matching technique, in the sense that matching must be chosen carefully to get an improvement.

Once again, the lowest error rates are obtained using the normalised correlation. We have added results using the Gradient Direction Metric (GDM), which is specially designed for matching in the LDA subspace.

Prot.	Metric	g2		g1			PCA	LDA
		FAR	FRR	FAR	FRR	HTER	dim.	dim.
G	EUD	11.86	29.49	25.64	12.82	19.95	330	140
	NOC	4.81	8.55	8.01	4.70	**6.52**	330	150
	NMM	10.90	27.35	22.12	9.40	17.44	330	140
	GDM	13.46	26.07	20.19	9.83	17.39	330	140
P	EUD	21.15	31.20	32.37	20.09	26.20	330	140
	NOC	16.99	18.38	17.95	14.10	**16.85**	360	230
	NMM	21.79	27.35	25.64	20.51	23.82	330	140
	GDM	23.08	28.21	28.21	19.66	24.79	330	140

Table 4.11: *Matching in the LDA subspace: BANCA database (English subset), rates for normalised correlation (NOC), Euclidean distance (EUD), nearest mean metric (NMM) and gradient direction metric (GDM). The images were symmetrised before matching.*

Unfortunately the results are poor in comparison with normalised correlation. This may be due to the fact that the intraclass covariance is not the identity matrix in the LDA subspace as assumed in the GDM method because the LDA subspace is not computed using the same user images.

It can be seen that in the most demanding sub-protocol, the sub-protocol P which allows only one training session, the HTER is about 16%. From the Table C.3 in Appendix C we observe that most of the errors come from the unmatched protocols: i.e. when the enrolment conditions do not match the testing conditions. This is why the HTER goes down to 6% when three images recorded in the three different environmental conditions are available for training (protocol G).

4.4 Conclusions

In this chapter, we have presented the various experiments performed on two large face databases. Face verification algorithm performance depends on a large number of factors. The search for the best performance can be seen as an optimisation problem in the space of all algorithm parameters. As the effects of the different factors such as pre-processing technique, matching, representation space, are not independent, in theory an exhaustive search for the optimum parameter set should be done. Obviously, these exhaustive results would be difficult to present. For this rea-

son we presented results that have a practical impact on the performance.

The aim of the experiments were: (i) to test the various face verification algorithms presented in Chapter 2. (ii) To determine the influence of algorithm parameters (iii) to determine heuristically pre- and post-processing techniques that improve verification accuracy.

Although empirical conclusions should be considered with caution, a certain number of observations can be done.

- Normalised correlation (NOC) consistently outperforms the other matching techniques. Interestingly, the NOC matching does not require any training in contrast to matchings such as nearest mean metric or gradient direction metric.

- LDA representation is superior to other representations tested provided that parameters are chosen carefully. Its performance is also intimately linked to the matching schemes.

- All tested algorithms, which are appearance based, seem to be unaffected by low-pass filtering and down-sampling of the images. This suggests that only low frequency information is used for classification. This may be due to the fact that the high frequencies (small details, edges, etc) are difficult to register and vary with facial expression and pose. It is therefore difficult to extract discriminative high frequency information, and the appearance based algorithms base their decision on low-frequencies only. This is probably why the recently proposed *component based techniques* [49], in which the face image is decomposed in small patches containing facial feature and registered separately, seem to beat traditional global techniques.

- Histogram equalisation in combination with symmetrisation of the face images are two heuristic pre-processing steps that help in reducing the verification error rates. The photometric normalisation preprocessing influences substancially the verification algorithm performance. For this reason, techniques such as [41] are worth investigating.

- The best HTER in the most realistic scenario (protocol P of the BANCA database) are about 16% with manual localisation and 25% using a fully automatic system. This high error rate is due to the different factors. One of them is the inadequacy of the enrolment and testing conditions. The second one is the high variability of the

images in the database. This error is quite high for realistic applications but, as we will see in Chapter 6, the HTER can be significantly decreased if the face verification is combined with a speaker verification algorithm.

Chapter 5

Intramodal fusion of face verification experts

5.1 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 will be 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 recognition/classification rates than any of the classifiers taken individually [60, 32, 52]. A major factor behind any improvement is the diversity in the classifier opinions [2], and methods such as boosting and bagging have been introduced to create diversity [13]. 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 chapter.

Among all possible combination rules, the sum, product, maximum, minimum and median rules, which effectively combine *a posteriori* class probabilities given by each classifier, have received very much attention [60, 65, 3, 113, 114], probably because of their simplicity and because they do not require training. If the classifiers operate in the same measurement space, i.e. they relate to the same phenomenon through the same sensor, different classifiers provide different estimates of the same posterior probability.

An example of such classifiers is an ensemble of neural networks with a fixed number of inputs but 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 even if this noise is positively correlated [114]. If the classifiers operate in different measurement spaces, i.e. they related 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 [60].

Although not optimal, the sum or averaging rule, in which the combined decision is based on the sum of the posteriors, has been shown to be remarkably robust to estimation noise [60]. 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 [113]. Experiments presented in [3] 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 because an estimate of the score class conditional density functions is needed.

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 [66].

This approach appears to be especially efficient in the context of biometric identity verification. The reasons are that:

- 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.
- 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 [97, 29, 105, 117, 9], where scores coming from multiple modalities or measurements like face and voice are combined (see chapter 6), and intramodal fusion [99, 53, 38, 21], where scores coming from the same modality but different matchers are combined.

In the latter approach, the second level classifier will require some training data to learn the combination decision rule. Also, this combination method 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 aim of this chapter 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. We propose a simple method for obtaining *a posteriori* probabilities estimates from raw scores. 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. In fact, estimation errors are almost always unknown and difficult to estimate in contrast to one-dimensional score distributions, which can be estimated more easily. This gives practical relevance to our study when choosing a fusion strategy for a given application. 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 *a posteriori* probability given by each classifier 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 chapter is organised as follows. In the next section, we introduce the problem of identity verification and its formalism. Probability-based

and second level classifier-based combination rules for two-class problems are discussed in Section 5.2. The effect of correlation between classifiers on probability combination rules is emphasised. In Section 5.3, after describing the face verification algorithms used, we present the combination results for face-based identity verification. Conclusions are given in the last section.

5.1.1 Multimodal and Intramodal Expert Fusion

In an attempt to verify the identity of a person, one can employ multiple measurements of different biometric traits, or modalities, in order to improve the results. For example a score coming from a face verification algorithm can be combined with a score coming from a speaker verification algorithm, in order to generate a more accurate decision. This is the topic of chapter 6.

The other possibility is to take advantage of multiple experts that provide their opinions on the same biometric data, and perform *intramodal fusion* as opposed to multimodal fusion described in chapter 6. Here, we focus on the intramodal expert fusion, i.e. when all experts base their opinion on the same face image (same modality, same sensor, i.e. same measurement space). 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 [114, 62], i.e.

$$f_i^c(\mathbf{x}) = P(\omega_c|\mathbf{x}) + \epsilon_i(\mathbf{x}, \omega_c),$$

where $\epsilon_i(\mathbf{x}, \omega_c)$ is the noise for class ω_c and expert i. This formulation suggests that a better estimation can be obtained by for example averaging the different estimates. In fact, for unbiased experts of equal accuracy ($E[\epsilon_i^2(\mathbf{x}, \omega_c)]$ is constant $\forall\ i$), averaging the estimates reduces the variance and therefore improves the posterior probability estimate [114] even if the noise exhibits some correlation between the experts. However, although we are assured to get an improved decision by using the averaging rule, there may exist other combination rules that outperform the averaging rule.

Except in some special cases such as k-NN classifiers or neural networks, experts usually do not output posterior probabilities but scores. 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 [31].

Here we propose to use the score *a posteriori* probability as the estimate $f_i^c(\mathbf{x})$, that is

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

The score *a posteriori* probability is given by the Bayes rule, leading to

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

Clearly this mapping from s_i to f_i^c is non-linear, hence averaging the scores $s_i(\mathbf{x})$ and the $f_i^c(\mathbf{x})$ does not have in general the same effect on the combined decision. When combining several scores using posterior probability, 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 Chapter 1. In fact, the probability distribution function $p(s|\omega_c)$ can be readily estimated from the FAR and FRR estimations, noticing from Equation (1.5) that

$$p(\eta|\omega_b) = \frac{de_A(\eta)}{d\eta} = \frac{d}{d\eta}\int_{-\inf}^{\eta} p(s_i|\omega_b)ds_i,$$

and

$$p(\eta|\omega_a) = -\frac{de_R(\eta)}{d\eta} = -\frac{d}{d\eta}\int_{\infty}^{\eta} p(s_i|\omega_a)ds_i.$$

5.2 Combination Rules

In this section we start by presenting and discussing the different score combination rules that use posterior probabilities. We continue the discussion by comparing 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 that have been employed for the problem of identity verification.

5.2.1 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 which reflects its opinion about pattern $\mathbf{x}$. Let $s_i(\mathbf{x})$, $i = 1, 2, ..., R$ be these scores. 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 label with maximum posterior probability, i.e. choose ω_a if

$$P(\omega_a|s_1, s_2, ..., s_R) > P(\omega_b|s_1, s_2, ..., s_R). \tag{5.1}$$

This is the best combination that can be made using the scores as it minimises the probability of error. However optimality requires in general the exact knowledge of the class conditional probability density functions, which are almost always unknown. Under the assumption that the scores s_i are class conditionally independent, for both classes i.e.

$$p(s_1, s_2, ..., s_R|\omega_c) = \prod_{i=1}^{R} p(s_i|\omega_c),$$

where $c \in \{a, b\}$, the posteriors of inequality (5.1) can be expressed as

$$P(\omega_c|s_1, s_2, ..., s_R) = P(\omega_c)^{-R+1} \frac{\prod_i^R p(s_i)}{p(s_1, s_2, ..., s_R)} \prod_{i=1}^{R} P(\omega_c|s_i).$$

Injecting this last equation into the decision rule (5.1) and noticing that $P(\omega_b|s_c) = 1 - P(\omega_a|s_c)$, we obtain the product rule, that is, choose class ω_a if

$$\prod_{i=1}^{R} P(\omega_a|s_i) > \prod_{i=1}^{R} (1 - P(\omega_a|s_i)), \tag{5.2}$$

otherwise choose ω_b. Note that equal priors have been assumed. 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. The effect of score dependence can be qualitatively evaluated if a score dependence function γ_c with $c \in \{a, b\}$ is introduced

$$\gamma_c(s_1, s_2, ..., s_R) = \frac{p(s_1, s_2, ..., s_R|\omega_c)}{\prod_i^R p(s_i|\omega_c)}.$$

The probability of class c given the R scores can be written as

$$P(\omega_c|s_1, s_2, ..., s_R) = \gamma_c(s_1, ..., s_R) P(\omega_c)^{-R+1} \frac{\prod_i^R p(s_i)}{p(s_1, s_2, ..., s_R)} \prod_{i=1}^{R} P(\omega_c|s_i),$$

and the product rule (5.2) becomes

$$\prod_{i=1}^{R} P(\omega_a|s_i) > \prod_{i=1}^{R}(1 - P(\omega_a|s_i))\frac{\gamma_b(s_1, ..., s_R)}{\gamma_a(s_1, ..., s_R)}. \tag{5.3}$$

It appears that the decision boundary from (5.2) is affected by the ratio of γ_b to γ_a. Since γ_b and γ_a are usually difficult to compute or estimate, this rule may be not applicable in practice. However, the decision rule given by (5.3) changes with respect to rule (5.2) only in the region of the score space $\mathbb{R}^R$ where $\frac{\gamma_b}{\gamma_a} > 1$ and $\prod_{i=1} P(\omega_a|s_i) > \prod_{i=1}(1-P(\omega_a|s_i))$ and in the region where $\frac{\gamma_b}{\gamma_a} < 1$ and $\prod_{i=1} P(\omega_a|s_i) < \prod_{i=1}(1 - P(\omega_a|s_i))$. Therefore, rule (5.2) and (5.3) will not differ very much if the decision boundaries defined by (5.2) and $\frac{\gamma_b}{\gamma_a} = 1$ are close to each other, even if the scores are not independent.

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 : $\min_k P(\omega_a|s_k) > \min_k P(\omega_b|s_k)$,
- Vote : $\sum_k u(P(\omega_a|s_k) - 0.5) > \sum_k u(P(\omega_b|s_k) - 0.5)$,
- Median $\text{med}_k P(\omega_a|s_k) > \text{med}_k P(\omega_b|s_k)$.

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 [65]. 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.

5.2.2 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^2_{ci} and the means μ_{ci}

($c \in \{a, b\}$). The class conditional joint score densities $p(s_1, s_2, ..., s_R|\omega_c)$ are also Gaussian with covariance matrices Σ_c, and means $\boldsymbol{\mu}_c$ with

$$\Sigma_c = \begin{pmatrix} \sigma_{c1}^2 & c_{12} & c_{13} & \dots & c_{1R} \\ c_{12} & \sigma_{c2}^2 & c_{23} & \dots & c_{2R} \\ \dots & \dots & \dots & \dots & \dots \\ c_{1R} & c_{2R} & c_{3R} & \dots & \sigma_{cR} \end{pmatrix},$$

and $\boldsymbol{\mu}_c = (\mu_{c1}, \mu_{c2}, \dots, \mu_{cR})^T$ where $c \in \{a, b\}$ The class ω_a *a posteriori* probability for expert i is given by

$$P(\omega_a|s_i) = \frac{p(s_i|\omega_a)P(\omega_a)}{p(s_i|\omega_a)P(\omega_a) + p(s_i|\omega_b)P(\omega_b)} = \frac{1}{1 + e^{-(\alpha_i s_i^2 + \beta_i s_i + \delta_i)}},$$

with

$$\alpha_i = \frac{1}{2\sigma_{bi}^2} - \frac{1}{2\sigma_{ai}^2}, \qquad \beta_i = \frac{\mu_{ai}}{2\sigma_{ai}^2} - \frac{\mu_{bi}}{2\sigma_{bi}^2}, \qquad \delta_i = \frac{\mu_{bi}^2}{2\sigma_{bi}^2} - \frac{\mu_{ai}^2}{2\sigma_{ai}^2},$$

and equal priors are assumed. Clearly $P(\omega_a|s_i)$ depends only on the one-dimensional statistics of score s_i. The statistical properties of the experts are determined by the covariance matrices Σ_c and means μ_c:

- Correlation between the experts i and j is reflected in the off-diagonal element c_{ij} of Σ_c.
- Accuracy of the ith expert depends on the σ_{ci}^2 and $\mu_{ai} - \mu_{bi}$. Hence for a fixed difference $\mu_{ai} - \mu_{bi}$, a larger variance σ_{ci}^2 amounts to a larger overlap between the two score distributions and therefore a higher error rate (see Appendix A).

In this particular case where the exact joint score densities are known, the minimal combination error which is the Bayes error, can be computed using Equations (1.6), (1.3) and (1.4), and compared to the error given by probability combination rules cited above. Note that the Bayes error for the original two-class problem defined in the space of $\mathbf{x}$ may be smaller (but not greater) as discriminatory information has been lost during the score calculation mapping from the high-dimensional space of $\mathbf{x}$ onto $\mathbb{R}$. By varying these parameters, we can evaluate the efficiency of the combination rules versus optimal combination 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. Therefore the off-diagonal elements of Σ_c can be written $c_{ij} = \rho_c \sigma_{ci} \sigma_{cj}$. As scores are likely

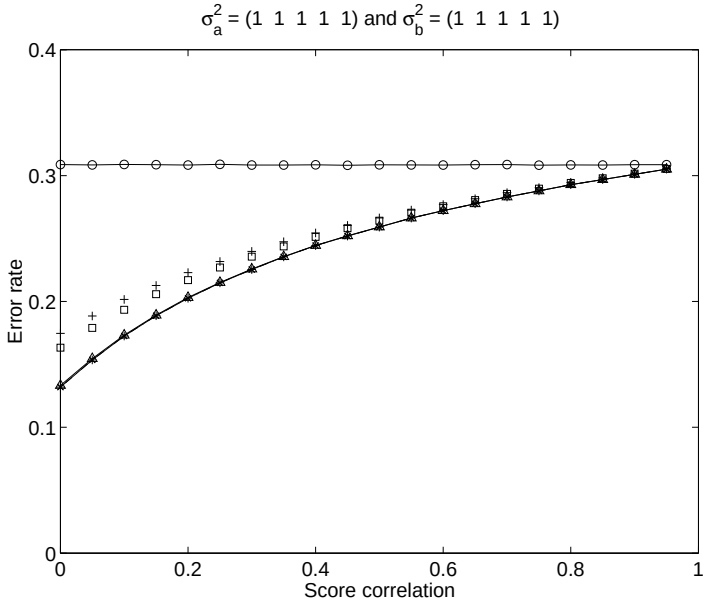


Figure 5.1: *Combination error versus expert correlation* ρ_c *in the case* $\sigma_{ai} = \sigma_{bi} = 1$ *for* $i = 1, 2, ..., 5$. *Key:* $\circ$ *best single expert;* $\square$ *max/min; + vote/median;* $\triangle$ *sum; * product;* $\times$ *Bayes error.*

to be positively correlated, the case $\rho_c > 0$ only is presented. Class ω_a is centred at the origin and class ω_b is centred at $(1,1,1,1,1)^T$. Figures 5.1, 5.2, 5.3 show the classification error obtained by the probability combination rules versus the Gaussian parameters for several representative situations. In Figure 5.1 the combination error versus the score correlation is shown in the case $\sigma_{ai}^2 = \sigma_{bi}^2 = 1$ for $i = 1, 2, ..., 5$ 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

$$r = \frac{\sigma_i(\mu_{aj} - \mu_{bj})}{\sigma_j(\mu_{ai} - \mu_{bi})} \tag{5.4}$$

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 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. Figure 5.2 shows the error rate in the case $\rho_a = \rho_b$ = 0.6 and $\sigma_{ai}^2 = 1$ for $i = 1, 2, ..., 5$ 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 Figure 5.3), 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.

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), Figure 5.3 shows that for moderate values of correlation (below 0.3), product may be close to optimality. In case (b), Figure 5.2 shows that, with high correlation between

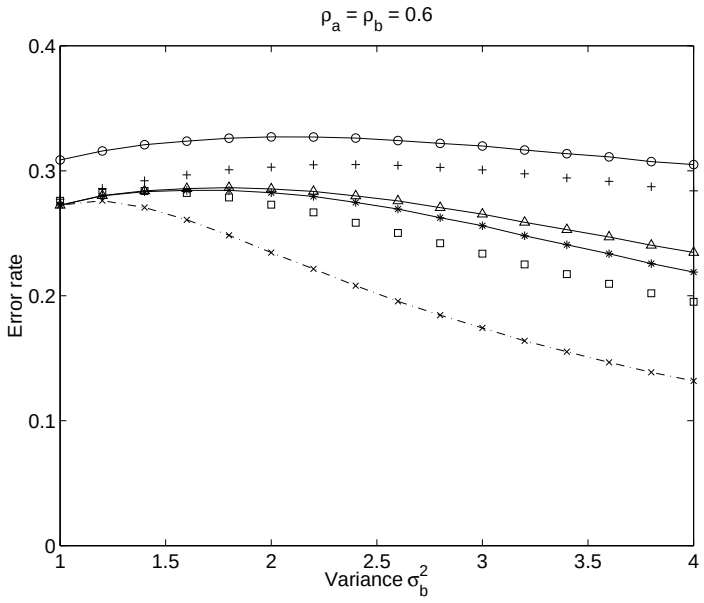


Figure 5.2: *Combination error versus σ_{bi}^2 in the case $\sigma_{ai} = 1 \; \forall \, i$ and $\rho_c = 0.6$. Key: ∘ best single expert; □ max/min; + vote/median; △ sum; * product; × Bayes error.*

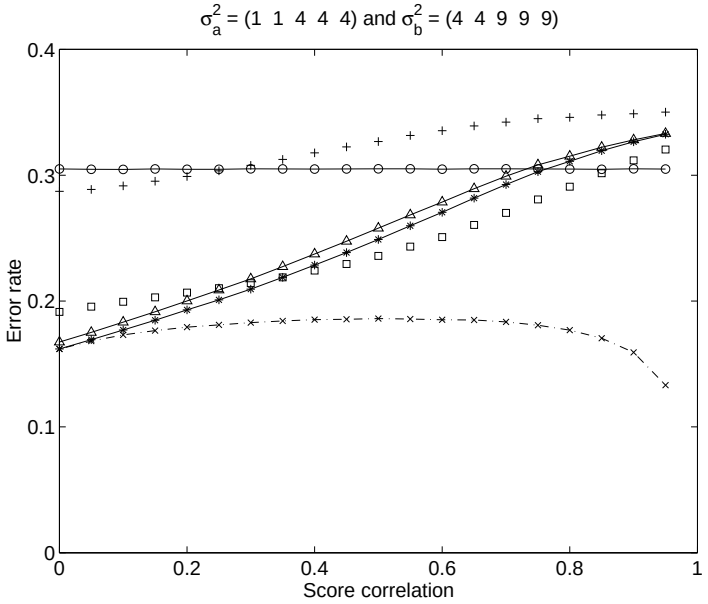


Figure 5.3: *Combination error versus expert correlation ρ_c in the case $\sigma_{ai} = 1$ and $\sigma_{bi} = 4$ when $i \in \{1, 2\}$, and $\sigma_{ai} = 4$ and $\sigma_{bi} = 9$ when $i \in \{3, 4, 5\}$. Key: $\circ$ best single expert; $\square$ max/min; + vote/median; $\triangle$ sum; * product; $\times$ Bayes error.*

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.

5.2.3 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. This classifier (or combiner) learns the optimal decision boundary defined by (5.1) from the training samples. It is therefore implicitly learning the dependencies between the different experts. Several issues concerning the training data needed for training the combining classifier are discussed in [32]. Basically for combining the scores any classifier can be used, provided that the capacity of the classifier is correctly chosen with respect to the available training. In [99], a non-parametric Parzen estimation technique is used to estimate the joint score density and rule (5.1) is used directly for combining several fingerprint matchers. A support vector classifier and a neural network fusion are compared in [9] for audio-visual

authentication. A weighted averaging of scores, which corresponds to a linear discriminant function in the score space is proposed in [29, 105, 21]. In [53] and [117], a logistic regression based combination is used. Note that the logistic regression also defines a linear discriminant function in the score space. 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.

5.3 Face Verification Expert Combination

We combined five different face verification experts. Based on the same face image, the experts use different feature extraction and matching algorithms to generate the scores. After describing the experts, we report combination results using both *a posteriori* probability combination rules and trainable combiners. All experiments presented in this section have been performed on the XM2VTS database (see chapter 3).

5.3.1 Face Experts

In this section, we describe the five face experts available for combination. Among them two 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. After localisation, the face image is cropped and photometric normalisation is applied to reduce the effect of lighting variation. Note that all the experts locate the face based on the same eye coordinates, however they operate on different image sizes and croppings.

LDA-based experts

The first expert referred to as **LNC** uses normalised correlation in the LDA subspace to compute the score. The second expert referred to as **GDM** uses the gradient direction metric. Both experts are described in Chapter 2.

Probabilistic Matching Experts

Probabilistic Matching (see Chapter 2 for a description) associated with two different photometric normalisation techniques was 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.

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. Colour histograms are used in image database retrieval. To limit the effect of lighting variation, image pixels are first rescaled to have zero mean and unit variance. Also a segmentation is applied to remove background colour and hair parts from the image to be processed. 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.

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

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.

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 expected to be low. However as expert HST operates on colour only, it may provide information little correlated to the other experts.

5.3.2 Results

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

Individual Expert Performance

Table 5.1 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 5.1, it can be seen that expert GDM offers the best verification

	Eval.	Test		
expert	EER	FAR	FRR	HTER
GDM	2.17	3.00	2.69	2.84
LNC	2.93	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

Table 5.1: *Individual verification results on the XM2VTS database. The verification result is measured with the equal error rate on the evaluation set (EER) and the false acceptance (FAR), false rejection (FRR) and half total error rate (HTER) on the test set.*

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 5.2 shows the 2nd 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 Equation (5.4).

Trainable Combiners

To compare the probability combination rules to trainable combiners, the scores were combined with the following classifiers: weighted averaging, Bayes quadratic classifier 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 [99] using a Gaussian kernel. The joint densities are used to compute the *a posteriori* probabilities and rule (5.1) leads directly

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

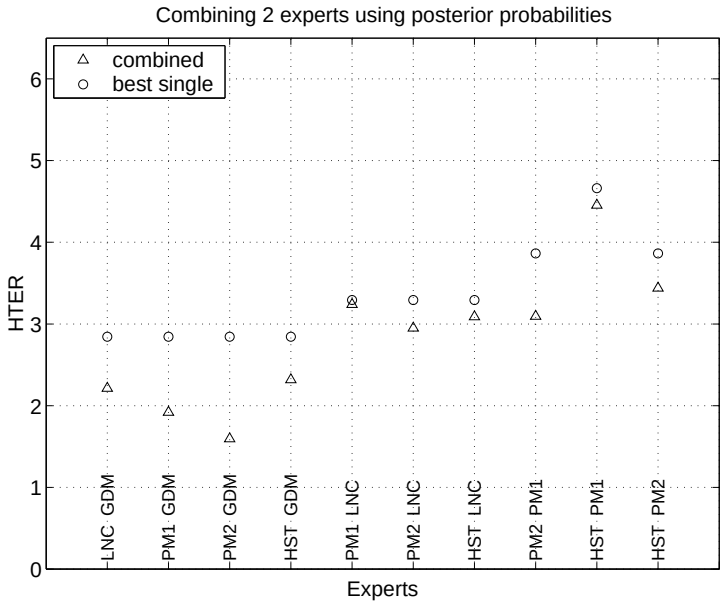
Table 5.2: *Expert score statistics*

to the decision. The combiners are trained on the 600 genuine matching scores and the 40,000 impostor matching tests forming the validation set.

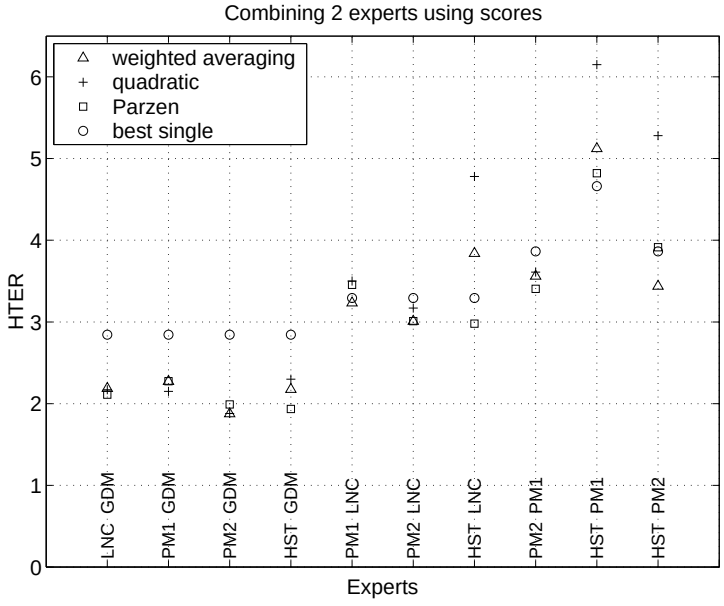
Combination results and discussion

The combination results are shown in Figure 5.4, 5.5, 5.6 for respectively two, three and four experts for both probability combination rules (a) and for trainable combiners (b). Each point on these figures corresponds to a 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 on Figure 5.4(a). 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 appreciable 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 5.2. For example PM1 and PM2 have approximatively the same correlation as PM1 and LNC. Also LNC and PM2 have similar error

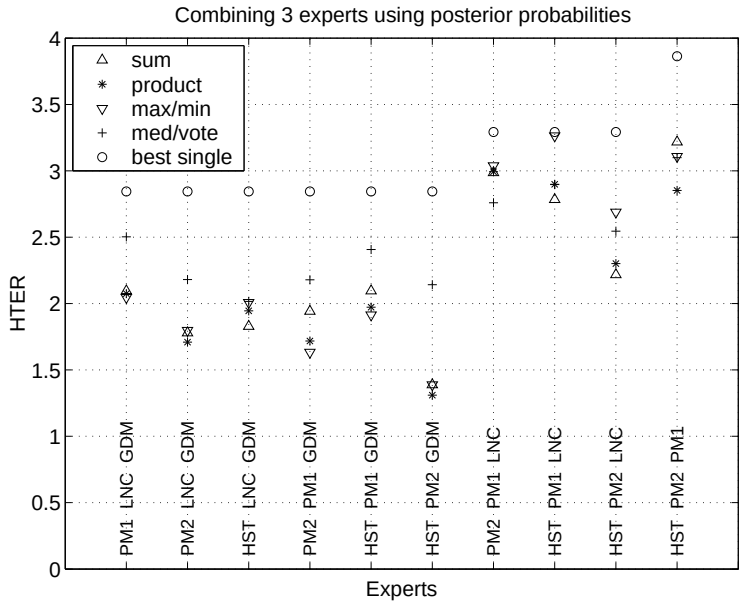


(a)

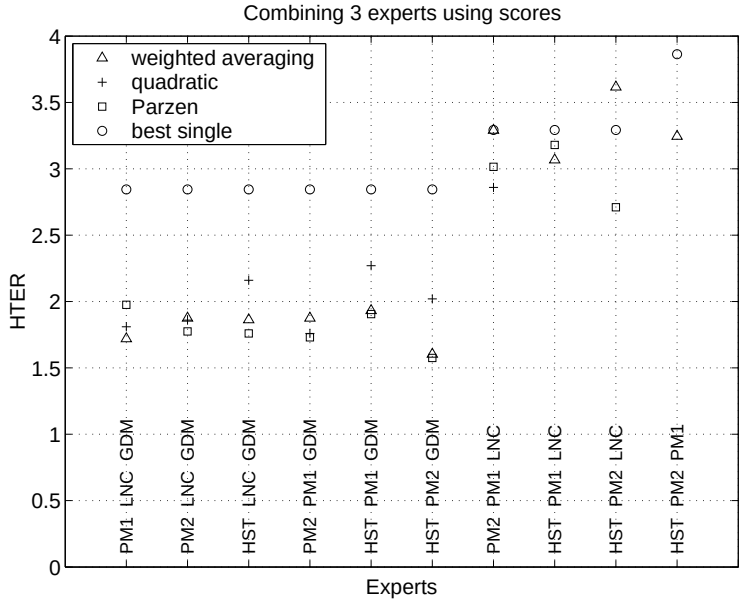


(b)

Figure 5.4: *Combination of two face verification experts: (a) a posteriori probability combination rules; (b) trainable combiners.*

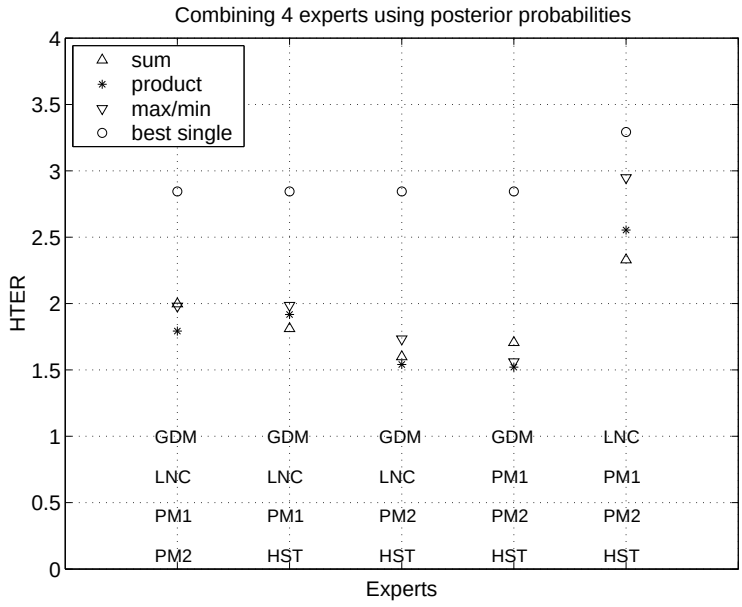


(a)

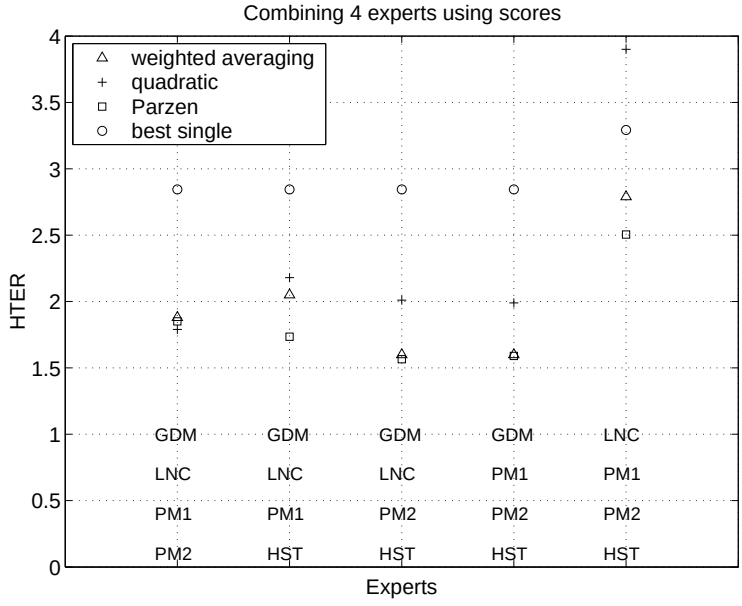


(b)

Figure 5.5: *Combination of three face verification experts: (a) a posteriori probability combination rules; (b) trainable combiners.*



(a)



(b)

Figure 5.6: *Combination of four face verification experts: (a) a posteriori probability combination rules; (b) trainable combiners.*

rates. However, the first pair of expert leads to an improvement while the second not.

When a third expert is added (see Figure 5.5) 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 [113], 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 [99], is required to gain the maximum from the fusion. The vote combination is not shown on Figure 5.6 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. Figure 5.7 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

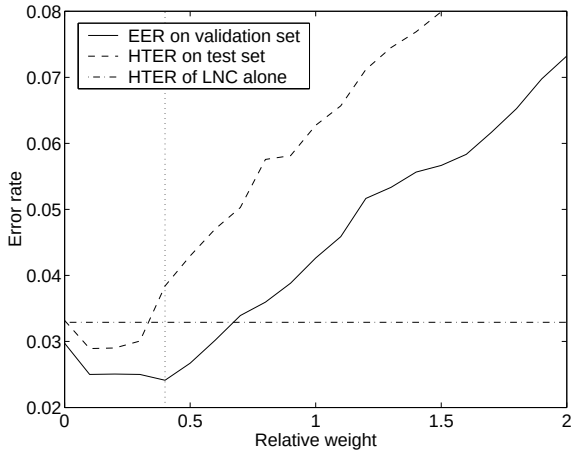


Figure 5.7: *Validation data EER and test data HTER vs. relative weight when combining HST and LNC. The HTER of single expert LNC is also drawn. The vertical line shows the optimal weight found on the evaluation set. Obviously it is not optimal on the test set.*

curves are quite different, which means that the distributions are different for validation and test data. The weight leading to the EER minimum is unfavorable for the HTER on the test data. In fact this difference comes mainly from the HST expert which is slighly overtrained: its false rejection rate on the test data 2% higher than on the validation set. This example confirms the ideas of Duin [32]: 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 ex-

perts as it is determined by taking the minimum HTER on the test set.

5.4 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 of 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.

Chapter 6

Multimodal fusion of Face and Speaker verification experts

We have seen in Chapter 5 that correlation between the classifier scores reduces significantly the improvement of the fusion system. In intramodal fusion, where all classifiers or experts make their decision based on the signal coming from same sensor (same image in our case), this correlation is unfortunately unavoidable. We could possibly reduce it by diversifying as much as possible the classifier design. For example, split the training set so that each classifier is trained on a separate subset, or resample (with replacement as in the bootstrap) the training set to train the different experts (this is the idea of *bagging* [13]). Another approach to reduce correlation between experts is input decimation: some parts of each input image is purposefully removed for the training of a given expert.

An alternative consists of looking for features in an other measurement space, i.e. utilise a sensor measuring another physical phenomenon (but still related to the same state of nature, that is to the same person in the identity verification context). If the two phenomena are independent, extracted features are also independent as well as the two matching scores. We know from Chapter 5 that the product rule combination rule is optimal in this case.

Following this idea, it has been suggested to combine or *fuse* several easily accepted biometric traits, in order to achieve an acceptable level of performance and user friendliness at the same time. This technique is known as *multimodal biometrics*. The aptitude of multimodal biometrics for increasing correct verification rates over monomodal biometrics has been demonstrated in several previous studies, (see for example [14, 29, 60, 48,

105, 97, 117, 9]).

Here, we take a slightly different point of view and look at how multimodality brings independent features which helps in reducing the size of the biometric template. This is important in applications where storage space is of concern. A promising application consists in combining biometric efficiency and smart card (SC) security, by storing the user template on a SC [108]. SC's allow the template to be securely protected and avoid storing biometric data in a central server (central storage is not well accepted by users). However storage space of SC's and transmission speed between server and SC's limit the user template size. It is therefore important to (i) evaluate performance as a function of the template size. (ii) propose strategies to decrease the template sizes.

In this chapter, we study the fusion of the face image and text independent speech data of a user. We analyse its scalability by evaluating the performance at different user template sizes.

The system presented is modular: each monomodal algorithm (face and speech) outputs a matching score reflecting its confidence in the presumed identity. The matching scores are then conciliated using a fusion algorithm which outputs the final authentication decision. The face verification algorithm is based on Linear Discriminant analysis and has been described in Chapter 2. The speech algorithm is based on Gaussian Mixture Modelling and has been developed at IDIAP (Institut Dalle-Molle d'Intelligence Artificielle Perceptive, Martigny, Switzerland).

When the identity of a user has to be verified, speech and face are recorded and compared to previously created user template. A fusion algorithm is used to conciliate the confidences coming from the comparisons and output the verification decision.

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

6.1 Fusion of speaker and face verification algorithm

When the identity of a user has to be verified, speech and face are recorded and compared to a previously created user template. A score reflecting the quality of the matching between the template and the data to verify is

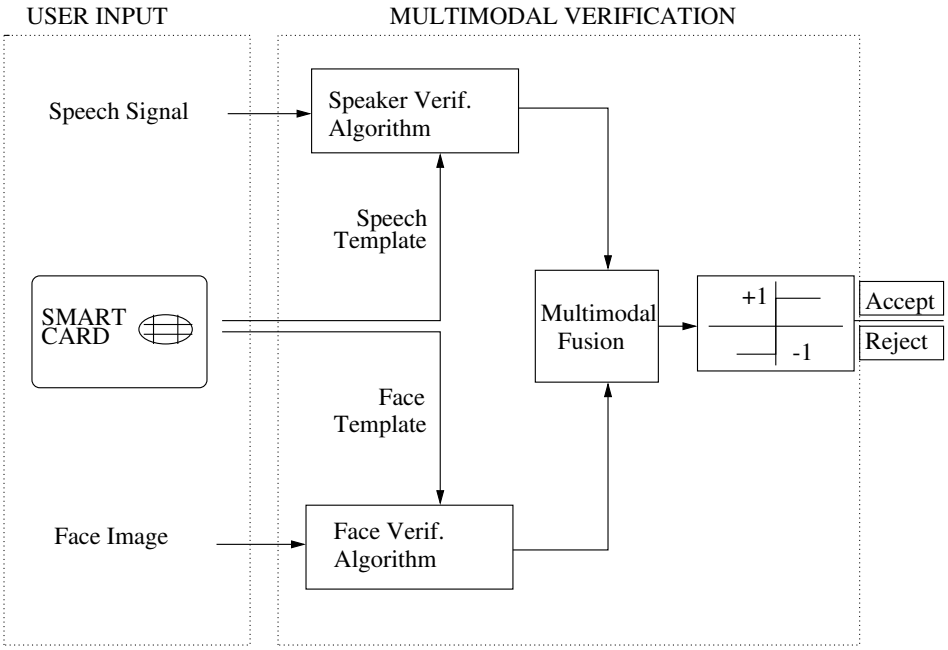


Figure 6.1: *The fusion system adopted*

computed for each modality. The fusion of the two scores resulting from the speech and face algorithms leads to the final decision. This process is depicted in Figure 6.1. Hereafter we describe briefly the speaker and face verification algorithms and the fusion techniques.

6.1.1 Speaker verification algorithm

The speaker verification algorithm used to compute the speech score is text independent and based on Gaussian Mixture Models (GMM) [103]. A parameterisation of the raw voice is performed, creating a vector of Linear Frequency Cepstral Coefficients (LFCC) for each section of 10ms of speech. On top of these coefficients, their first derivatives, as well as the log of the energy, are kept. Finally, a cepstral mean subtraction is performed in order to normalise the data. The user template is represented by a GMM that was adapted using a Maximum A Posteriori method from a general World Model GMM trained on a separate population of speakers. The speech score s_s is computed by estimating the log-likelihood ratio of a speech sequence of LFCC features $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ pronounced by

the speaker i versus the world of all speakers Ω (world model)

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

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

$$p(\mathbf{X}) = \prod_{t=1}^{T} p(\mathbf{x}_t) = \prod_{t=1}^{T} \sum_{n=1}^{N} w_n \cdot \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n), \tag{6.1}$$

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

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{6.2}$$

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

6.1.2 Face verification algorithm

The face verification algorithm outputs a face score s_f based on normalised correlation in the LDA subspace. In Chapter 2, it appeared that this algorithm performed the best on this dataset.

6.1.3 Fusion algorithms

The fusion of the face and speech scores is performed using a second level classifier. That is, the two scores are considered as input features for a classifier which is trained on genuine and impostor score examples. In our experiments, we have opted for a weighted average fusion. This very simple fusion technique appears to performs better than a MLP (multi-layered perceptron) based fusion [20]. In this technique, a new score s is computed by averaging the weighted scores $s = w_s s_s + w_f s_f$. The fusion score s is then thresholded to obtain the final decision. The weights w_s and w_f and the threshold are found so as to minimise the EER on the training set.

6.2 Scalability Analysis

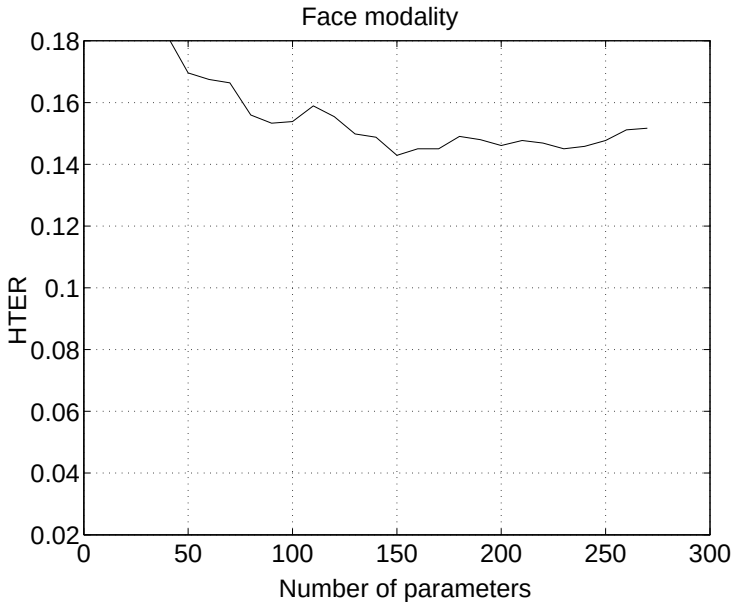
As stated in the introduction, we are interested in evaluating the performance of the monomodal and the multimodal algorithms at different user template sizes. While this template is relatively small for the face modality, it can be very large for the speech modality, mainly because of the large number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

For the speech modality, the number of parameters of the user template depends on the number of Gaussians in the GMM and the feature vector size, i.e. the number of LFCC's. The optimal number of Gaussians is normally determined by Maximum Likelihood on the world model. Here, we have trained the GMM on the world model with the number of Gaussians selected in the following set of values: 10, 25, 50, 100, 200 and 300. For each number of Gaussians, the performance of the speaker verification algorithm is assessed. We also studied the performance variation when k, the number of LFCC used to parameterise the raw voice varied between 4, 8, 12 and 16. Note that the derivatives of the LFCC and the signal energy are added to the feature vector. The feature vector size is therefore $2k + 1$.

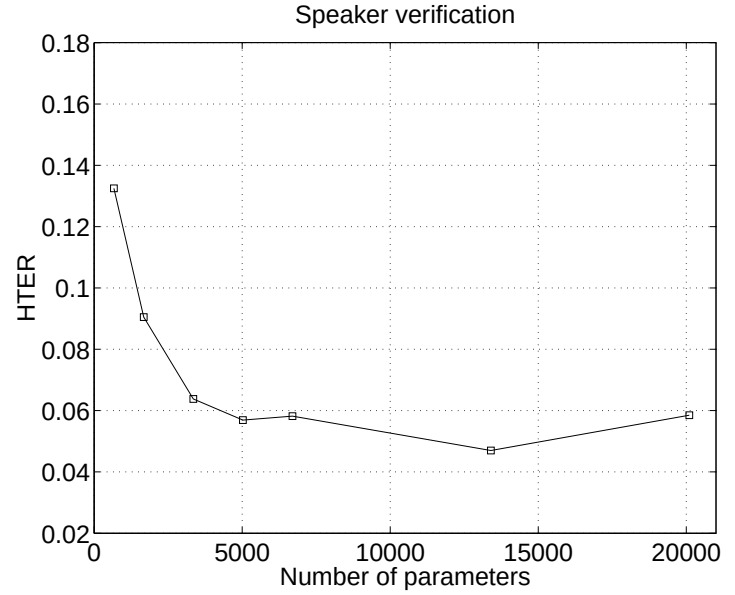
The number of parameters of the multimodal template is simply the sum of the face and speech templates. We studied the performance variation of the multimodal system at different sizes of the speech and face templates.

6.3 Experimental Results and Discussion

According to the protocol described in the previous section, we studied the variation of the HTER as a function of the user template size. Figure 6.2(a) shows the variation of the HTER versus the user template size (expressed in number of parameters needed to store it) for the face verification algorithm. From the figure, the HTER decreases significantly with the first 100 basis vectors. The minimum HTER is reached at 150 vectors, and increases above 150 vectors. This means that the last features extracted



(a)



(b)

Figure 6.2: *(a) HTER versus user template size for face modality. (b) HTER versus number of Gaussians needed to represent user template for speech modality.*

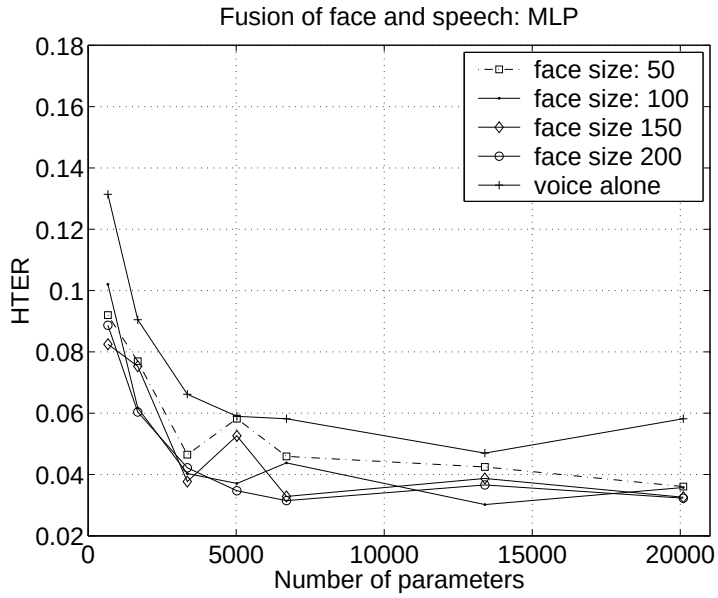
(from 150 to 294) slightly degrades the classification and should not be included in the user templates. The minimum HTER obtained is 14.3%. This high value can be partially explained by the fact that only one enrolment session is available for creating the user template.

Scalability results for the speech modality are shown on Figure 6.2(b). The curve on this figure corresponds to the variation of the HTER with the number of Gaussians and 16 LFCC coefficients. Since the other LFCC parameterisations offer higher error rates at equal number of parameters of the user template, we only present results with 16 LFCC coefficients. The best HTER for the speech modality is equal to 4.7% and is obtained with 200 Gaussians. It requires 13400 parameters to be stored.

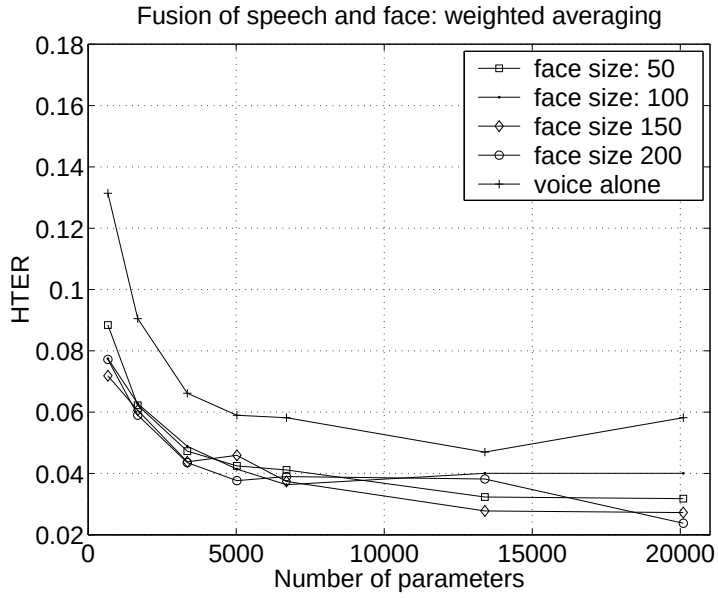
The results of the fusion experiments are presented on Figure 6.3(a) and 6.3(b) for MLP and weighted averaging fusion respectively. From these figures, it appears that the multimodal fusion always outperforms the best single modality (speech in our case). The lowest fusion HTER obtained is equal to 2.38%. The improvement thus reaches almost 50% in spite of the weakness of the face algorithm. Furthermore, the MLP fusion achieves an HTER of 3.77%, i.e. better than the speech modality alone, with only 50 Gaussians instead of 300. In this case, the number of parameters to be stored for the user template is 3500 (3350 for speech and 150 for face), which is almost 4 times less than what is needed for the system using the speech modality only. The speaker verification algorithm with 50 Gaussians for user template achieves an HTER of 6.62%. This result may be of practical interest when storage space is of concern, for example in a biometric system coupled with smart cards [108]. The limited storage and transmission speed of a smart card require user templates as small as possible. The fusion of modalities is therefore a way of improving the performance and reducing the number of parameters needed to be stored and transmitted, without decreasing performance.

6.4 Conclusions

A multimodal identity verification system using the speech and the face image of a user is presented. The experiments were conducted on a realistic database and according to a test protocol that allows only one enrolment session. The results show that the text independent speaker verification algorithm is robust and provides good results in spite of the uncontrolled nature of the data. In comparison, the face verification algorithm appears to be weak. A substantial improvement is gained when the out-



(a)



(b)

Figure 6.3: *(a) HTER for MLP fusion vs. number of Gaussians and face template size. (b) HTER for weighted averaging fusion vs. number of Gaussians and face template size.*

puts of the two monomodal algorithms are fused using simple techniques, the performance getting close to real world application requirements. An empirical analysis of the algorithm scalability with respect to the user template size is presented. This analysis shows that fusion may help in reducing the number of parameters needed to be stored while keeping a satisfactory level of performance [20].

Chapter 7

Dynamic aspects or sequential fusion of scores

7.1 Introduction

Typically during an access attempt the user will be interacting with the system over an extended period of time. Over this period many video frames will be available for face verification and can be used to enhance the system performance in a number of possible ways. For example, the probe image can be selected based on content, or based on the best score.

Multiple frames have been used extensively in face detection tasks for tracking the face through successive frames. In [58, 80] skin colour clusters are used to track the face and a motion model is introduced to estimate image motion and to predict search window. Kalman filtering and its extension to non-Gaussian and non-linear case (CONDENSATION [50]) have also been used to track the face (see for example [70]).

In the verification case, that is when a decision must be taken once the face has already been located, multiple frames can also help in enhancing the accuracy, by accumulating evidences over time. In [62], several techniques are introduced to integrate the information coming from multiple biometric measurements in order to improve the decision. These techniques can be applied for processing successive video frames of the user's face interacting with the system.

In this chapter we present the experimental results obtained using several score integration techniques. Both automatic and manual localisation were considered for two leading face verification methods, namely LDA- and SVM- based algorithms. All the experiments presented in this chapter

were obtained using the BANCA database.

In the next section, we discuss the techniques that can be implemented to integrate the information coming from several frames. Simulation results are presented to investigate the merits of the different integration rules in different situations.

The last section presents the face verification algorithms and the experimental setup. Results of our experiments are also given and commented.

7.2 Multiple frame integration

We are interested in answering the question : how to use several video frames to improve the decision of a face verification system. Although the dynamic aspects of video are very useful for face detection and face tracking, we are interested in the usefulness of multiple frames in verification. Consider that we have aquired a set of N_f frames and the face part has been already localised and cropped, forming a set of N_f random face vectors $\mathbf{x}_j$. The idea is therefore to use the set of face vectors $\mathbf{x}_j$ ($j = 1, 2..., N_f$) instead of a single frame, for instance $\mathbf{x}_1$, to improve the reliability of the verification.

7.2.1 Bayesian approach

The optimal approach consists in taking the decision through the log-likelihood ratio $l(\mathbf{x}_1, ..., \mathbf{x}_{N_f})$ [37]

$$l(\mathbf{x}_1, ..., \mathbf{x}_{N_f}) = \log \frac{p_a(\mathbf{x}_1, ..., \mathbf{x}_{N_f})}{p_b(\mathbf{x}_1, ..., \mathbf{x}_{N_f})},$$

where $p_a(\mathbf{x}_1, ..., \mathbf{x}_{N_f})$ and $p_b(\mathbf{x}_1, ..., \mathbf{x}_{N_f})$ are the joint probability density functions for class ω_a (genuine) and ω_b (impostor) respectively. This approach requires the knowledge of the joint pdf's and, due to the high dimensionality, it is is untractable in practice. Moreover, we cannot assume that the $\mathbf{x}_j$'s are mutually independent, because in fact the conditions (illumination, pose, etc.) may not vary that much from frame to frame (ex. the lighting, an important source of noise, may be approximately constant accross frames.) As a result, two subsequent frames $\mathbf{x}_j$ and $\mathbf{x}_{j+1}$ may show some correlation.

7.3 Score-based integration

A more tractable approach consists in extending the single frame technique to the multiple frame case. Recall from Chapter 1 that in the single frame technique, a *score* is computed from the observation $\mathbf{x}_1$, $s = s(\mathbf{x}_1)$ and compared to a threshold η

$$s(\mathbf{x}_1) \underset{\omega_b}{\overset{\omega_a}{\lessgtr}} \eta.$$

The class ω_a (genuine match) or ω_b (impostor match) is chosen whether the score is greater or smaller than the threshold. The score function is generally a distance measure between $\mathbf{x}_1$ and a template

When multiple frames are available, a score s_j can be obtained from each frame. The problem becomes now: find a function f so that the decision rule

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

leads to higher verification performance. Let us emphasise the difference between what should $f(.)$ be in this case and in the fusion (intramodal or multimodal) case. In fusion, we combine scores coming from different experts. Naturally, an expert should have a contribution (a weight) to the decision related to its accuracy: the more accurate expert, the bigger the expert weight. In multiple frame combination, all scores are emitted by the same expert. For this reason, the combination should either give the same importance to all score either have a mechanism of selection of the "best" frame or best score.

While complex image processing techniques could be deployed for selecting the best frame, a simple solution consists in using a template matching method: select the frame that matches the template the best in the sense of a distance measure. In the case of a dissimilarity score (similarity), this results in taking the minimum (maximum) score for making the decision, i.e.

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

This may favour both genuine accesses and impostor accesses. As we will see below, the merit of this combination rule depends essentially on the score probability function. One could also select the more dissimilar frame, i.e. select the frame that matches the impostor template. This

amounts to taking the maximum score for a dissimilarity based score

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

7.3.1 Score averaging

Alternatively scores corresponding to multiple frames can be seen as multiple outcomes of the score random variable S. The scores can therefore be averaged, so that each s_i has the same importance in the decision. The scores s_i are drawn from the random variable S with score probability distributions $p(s|\omega_b)$ and $p(s|\omega_a)$ in case of impostors or clients. It is well known that the samples average

$$\bar{S} = \frac{1}{N_f} \sum_{j=1}^{N_f} S_j$$

of N_f samples S_j (considered here as random variables) drawn from a given distribution has the same mean than the distribution, i.e.

$$\mathrm{E}[\bar{S}] = \mathrm{E}[S] = \mu.$$

Also, if the samples are drawn independently, the variance of the average is

$$\mathrm{Var}[\bar{S}] = \frac{\sigma^2}{N_f},$$

where σ^2 is the variance of the score distribution S. Therefore $p(\bar{s}|\omega_c)$ ($c \in \{a, b\}$) has the same mean than $p(s|\omega_c)$ but a variance divided by N_f. Because the error rate depends directly on the overlap between the impostor and genuine sample mean densities, if the decision is taken using $\bar{s}$ rather than s, the error rate decreases as N_f increases.

The effectiveness of the average rule for reducing the error is significantly diminished when X_j's and subsequently S_j's are correlated. In this case the variance of $\bar{S}$ can be computed as

$$\mathrm{Var}[\bar{S}] = \frac{\sigma^2}{N_f} + \frac{1}{N_f} \sum_{i=1}^{N_f} \sum_{\substack{j=1 \\ j \neq i}}^{N_f} \mathrm{E}[(S_i - \mu)(S_j - \mu)].$$

The second term containing the correlation between the S_j's contributes to increase the variance of $\bar{S}$.

7.3.2 Minimum rule

The effect of the minimum rule (or maximum rule in case of a similarity score) is more difficult to analyse than the average rule. Let us call S_m the random variable that corresponds to the minimum of the N_f samples drawn from $p(s|\omega_c)$:

$$S_m = \min(S_1, S_2, ..., S_{N_f}).$$

The merit of the minimum rule will essentially depend on the statistics of S_m. The cumulative distribution $F_{S_m}(x)$ of S_m can be found as follows.

$$\begin{aligned}
F_{S_m}(x) &= \mathrm{P}(S_m < x), \\
&= \mathrm{P}(\min(S_1, S_2, ..., S_{N_f}) < x), \\
&= 1 - \mathrm{P}(\min(S_1, S_2, ..., S_{N_f}) > x), \\
&= 1 - \mathrm{P}(S_1 > x, S_2 > x, ..., S_{N_f} > x), \\
&= 1 - \mathrm{P}(S_1 > x)\mathrm{P}(S_2 > x)...\mathrm{P}(S_{N_f} > x), \\
&= 1 - (1 - F_S(x))(1 - F_S(x))...(1 - F_S(x)), \\
&= 1 - (1 - F_S(x))^{N_f},
\end{aligned}$$

where F_S is the cumulative distribution of S and where we have assumed that the samples are drawn independently. The probability density of the sample minimum is therefore

$$p(s_m|\omega_c) = \frac{dF_{S_m}(x)}{dx}\Big|_{s_m} = N_f(1 - F_S(s))^{N_f - 1} p(s|\omega_c)\Big|_{s_m}. \qquad (7.1)$$

7.3.3 Gaussian scores

Suppose that the impostor and genuine score distribution are Gaussian with equal variance σ^2 but different means

$$p(s|\omega_c) \sim \mathcal{N}(s; \mu_c, \sigma^2),$$

with $c \in \{a, b\}$. We draw independently N_f samples s_j from the genuine or the impostor distribution and base our decision on the average of s_j or the minimum s_j. The classification error when the decision is based on the

sample average is equal to

$$\begin{aligned} e &= P(\omega_b)\mathrm{P}(\bar{S}|\omega_b < \eta) + P(\omega_a)\mathrm{P}(\bar{S}|\omega_a > \eta), \\ &= P(\omega_b)\mathrm{P}(Z < \frac{\eta - \mu_b}{\sqrt{N_f}\sigma}) + P(\omega_b)\mathrm{P}(Z > \frac{\eta - \mu_a}{\sqrt{N_f}\sigma}), \\ &= \frac{1}{2}\phi(\frac{\eta - \mu_b}{\sqrt{N_f}\sigma}) + \frac{1}{2}\phi(\frac{\mu_a - \eta}{\sqrt{N_f}\sigma}), \end{aligned}$$

where Z is a Gaussian variable with zero mean and unit variance, $\phi(.)$ is the cumulative distribution function of Z, η is the threshold and both classes are equiprobable.

In the second case, we must study the properties of the sample minimum distribution $p(s_m|\omega_c)$. Using Equation (7.1), $p(s_m|\omega_c)$ can be written

$$p(s_m|\omega_c) = \frac{N_f}{\sqrt{2\pi}\sigma}(1 - \phi(s_m))^{N_f - 1} e^{-(s_m - \mu_c)^2/2\sigma^2}.$$

This probability density must be integrated in order to find the classification error. Unfortunately, the integration is difficult to do analytically and numerical integration is easier. Figure 7.1 shows the classification error versus the number of samples N_f for Gaussian distributed scores for sample average based and minimum based decision. Note that although the original distributions lead to an error of 15%, the average rule reaches almost zero error rate with $N_f = 10$. However, in practice such an improvement is unlikely because the samples are not drawn independently. A saturation is likely to occur when N_f increases. Both integration methods improve the decision over the one sample score case, but clearly the average rule outperforms the minimum rule. This behaviour can be qualitatively explained as follows. In fact, the sample minimum density S_m has a mean that is smaller than S, and this is true for both the impostor and genuine densities. The big advantage of the average rule is that the variance of the sample average density $\bar{S}$ decreases in $\frac{1}{\sqrt{N_f}}$. The variance of $p(s_m|\omega_c)$ does not decrease as fast, and therefore the overlap between the impostor and client densities is larger for the minimum than for the average rule.

7.3.4 Non-Gaussian score distribution

Although natural, Gaussian hypothesis for scores may not be satisfied in practice. In particular the genuine score density has some properties that

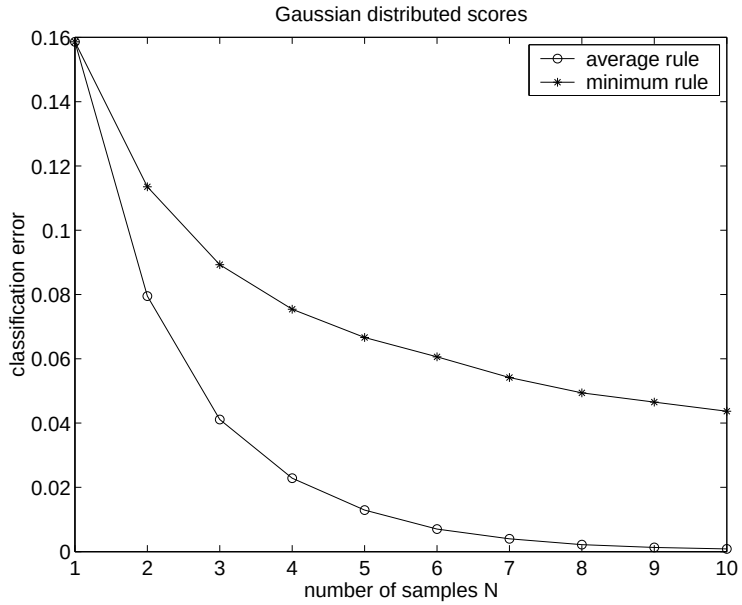


Figure 7.1: *Classification error versus the number of samples N_f for Gaussian distributed scores.*

are common to all biometric system: they are usually asymmetric with a heavy tail or even bimodal [118, 119]. The secondary mode is due to users who consistently return large distance measures when samples are compared to stored templates, called "goats" in [28].

Another contributing factor to the heavy genuine density tail comes from failures during pre-processing. For example, faces wrongly detected for face-based biometrics or failing silence removal for speech-based biometrics lead to large distance measures and false rejection.

A more realistic choice is to represent the genuine score density by a bimodal mixture of Gaussians

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

where $\pi_1 + \pi_2 = 1$, $\pi_1 \geq 0$ and $\pi_2 \geq 0$. The impostor score density is kept Gaussian $p(s|\omega_b) \sim \mathcal{N}(s; \mu_b, \sigma_b^2)$. In this case, the error rates are obtained through simulations and results are presented on Figure 7.2. In this figure, the classification error is computed for various value of N_f. Genuine samples are randomly drawn from a mixture with parameters $\pi_1 = 0.9$, $\pi_2 = 0.1$, $\mu_{a1} = 0$, $\mu_{a2} = 4$ and $\sigma_{a1} = \sigma_{a2} = 1$. The impostor density

has parameters $\mu_b = 4$ and $\sigma_b = 1$. Note that the "sample average" curve is very similar to the pure Gaussian case. This is because the secondary mode is relatively small (0.1 in probability), and its effect on the average is weak. In contrast, the secondary mode changes drastically the "sample minimum" curve, which now outperforms the average rule.

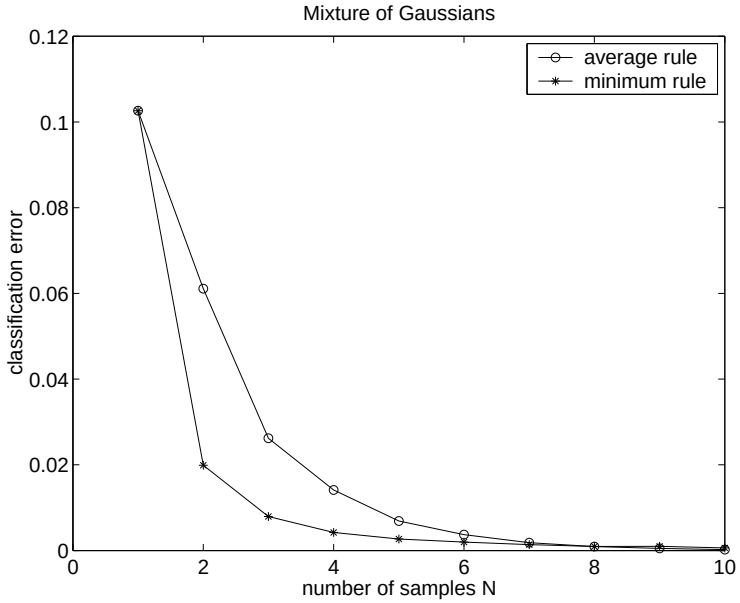


Figure 7.2: *Classification error versus the number of samples N for Gaussian distributed scores.*

The merit of the average and minimum rule depends essentially on the score densities. We see that average rule is advantageous in Gaussian case, while minimum rule starts to outperform the average rule when the genuine density has a heavy tail.

7.4 Face detection and verification algorithms

The first step involves localisation and registration of the face part in the input image. In the results presented below, we have used two different face detectors provided by university of Surrey as well as manually located eye coordinates. Once the face and eyes are located, the face is registered and photometrically normalised. The normalised face image is then used to generate the accept/reject decision.

7.4.1 SVM-based face and eye detection

The whole image is exhaustively scanned at different scales using a small window (20x20 pixels). The content of each window is classified by binary classifiers into face or non-face classes. An SVM classifier was used for this purpose. In order to obtain non-face images for training the SVM, we used a bootstrapping method controlled manually to eliminate face images from the training set belonging to the non-face class. The localisation of the eye centers tests each pixel for being the center of an eye. A rectangular windows scans the face sub-image at several scales. Again the content of the window is classified as eye or non-eye. The best pair of eyes is chosen according to the best score and best geometrical configuration.

7.4.2 Hypotheses-driven Affine Invariant face localisation (HAI)

In contrast to the previous method, local feature detectors are used to detect facial features such as corners of eyes, nostrils, etc. in the input image. Feature configurations that correspond possibly to a face sub-image are transformed into a normalised face space. In the face space, natural and geometric variability of faces is highly reduced because faces are registered. Full affine invariance is thus achieved *before* classification in face/non-face classes. This allows a much more accurate face/non-face classifier to be trained and implemented in the face space. Additionally, the search through subsets of features which invoke face hypotheses is reduced using geometric feature configuration constraints learned from the training set (geometric pruning).

Local features are detected using Gabor filters and the face/non-face classifier is an SVM classifier. The algorithm produces the location of the face in the image and also the positions on the facial features. Eye coordinates can be found from the detected features and used to register the face. More details about this method can be found in [45] and [44].

7.4.3 SVM-based face verification

This method is described in [107]. To label the face vector $\mathbf{x}$ as genuine or impostor, the SVM classifier evaluates the quantity

$$\hat{y} = \text{sign}\left(\sum_{i=1}^{l} y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b\right),$$

where $\hat{y}$ is the decision of the model (accept if $\hat{y}$ is positive, reject otherwise), $\mathbf{x}_i$ is the input vector of the ith training example, y_i its corresponding label ($y = 1$ for genuine $y = -1$ for impostor match), l is the number of training examples, the α_i and b are the parameters of the model, and $K(\mathbf{x}, \mathbf{x}_i)$ is a kernel function that can have different forms. In our case, the kernel is simply linear:

$$K(\mathbf{x}, \mathbf{x}_i) = \mathbf{x}^T \mathbf{x}_i.$$

Note that in contrast to the LDA-based method, the decision rule must be learned individually for each user.

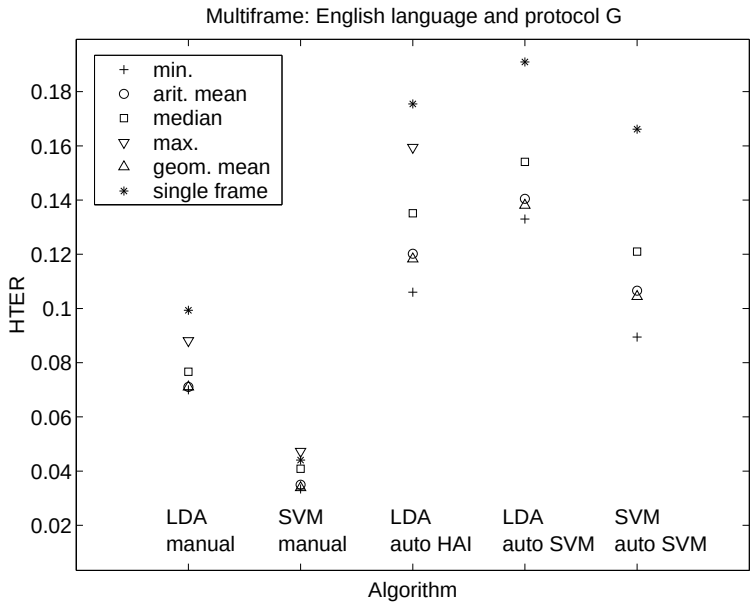


Figure 7.3: *HTER obtained with different integration methods for protocol G on the english part of the BANCA database.*

7.5 Experimental integration results

Multiple frame integration results were obtained on the BANCA database because it contains 5 frames per access.

During an access, five frames are randomly selected from the video stream. For each frame, the face is located (either manually or automatically) and a score is computed. Following Section 7.3, we compare the dif-

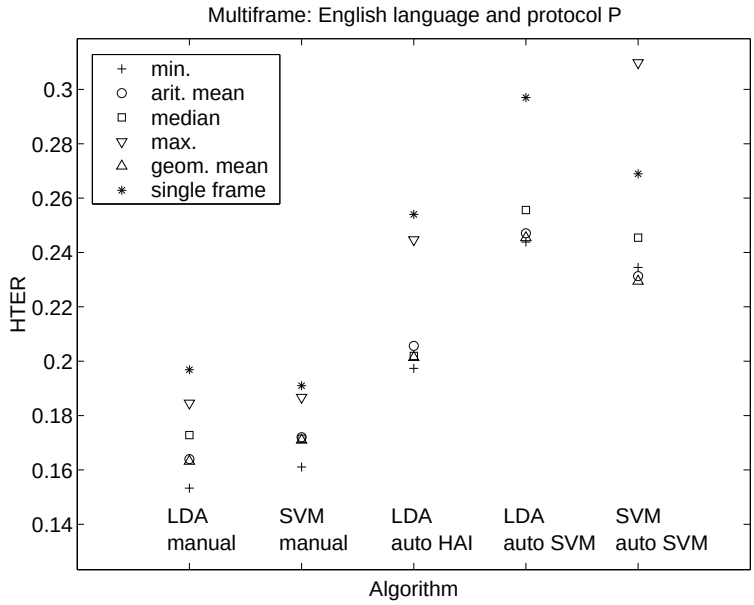


Figure 7.4: *HTER obtained with different integration methods for protocol P on the english part of the BANCA database.*

ferent ways of integrating the five scores obtained during an access. The following score integration techniques are used to obtain the final score s:

- Arithmetic mean : $s = \frac{1}{N}\sum_{i=1}^{N} s_i$,
- Geometric mean : $(\prod_{i=1}^{N} s_i)^{1/N}$,
- Minimum : $s = \min(s_1, ..., s_N)$,
- Maximum : $s = \max(s_1, ..., s_N)$,
- Median : $s = \text{median}(s_1, ..., s_N)$.

The final accept/reject decision is based on s as in the single frame method.

Figure 7.3 shows the average HTER obtained on the english set of the BANCA database using the different integration rules above for protocol G. Both manual and automatic results are presented for the LDA- and the SVM-based face verification algorithm. From the figure, it appears that except for the maximum integration rule, all the rules improve the HTER with respect to single frame based decision. In all cases, the minimum rule

gives the best improvement, up to several percents, over the single frame technique. Arithmetic and geometric average rules give results slightly worse while the median and the maximum rules are much worse. Detailed results gathered in the figure can be found in Appendix B

Figure 7.4 shows the average HTER obtained on the english set of the BANCA database using the different integration rules above for protocol P this time. Obviously, the HTER is much higher than for the G protocol, because only one session is available for training in this protocol. Again, except for the maximum integration rule, all the rules improve the HTER with respect to single frame based decision. Note that this time the minimum rule does not give the best improvement in the case of SVM expert with automatic eye localisation. Again the arithmetic and geometric average, give slightly worse results and median and maximum are much worse. Surprisingly, although the LDA-based expert outperforms the SVM-based expert when eyes are manually located, the SVM-based expert show better performance for automatically detected eyes when an SVM eye detector is used.

7.6 Integration results as a function of number of test frames

In this set of experiments, we investigate the relationship between the improvement brought by integration and the number of test frames integrated. For this, we evaluate the HTER as function of test frames that were integrated. Starting with one frame, we add successively a frame to the set and base the system decision on the set of frames. The number of test frames is thus varied from one to five, five being the number of frame available in the BANCA database. As it performed the best in the previous experiment, the minimum rule was chosen for conducting the experiment. The HTER obtained with varying number of test frames are reported in Figure 7.5 for protocol G and 7.6 for protocol P. As expected, when the number of test frame increases, the HTER decreases. It seems that in the case of LDA algorithm with HAI face detection, the HTER could go down further if more frames were used. In other words, no saturation of performance is observed when increasing the number of test frames for this algorithm. In contrast, the improvement obtained with manual SVM algorithm seem to saturate after the first three frames showing that no useful information is added when more observations are gathered.

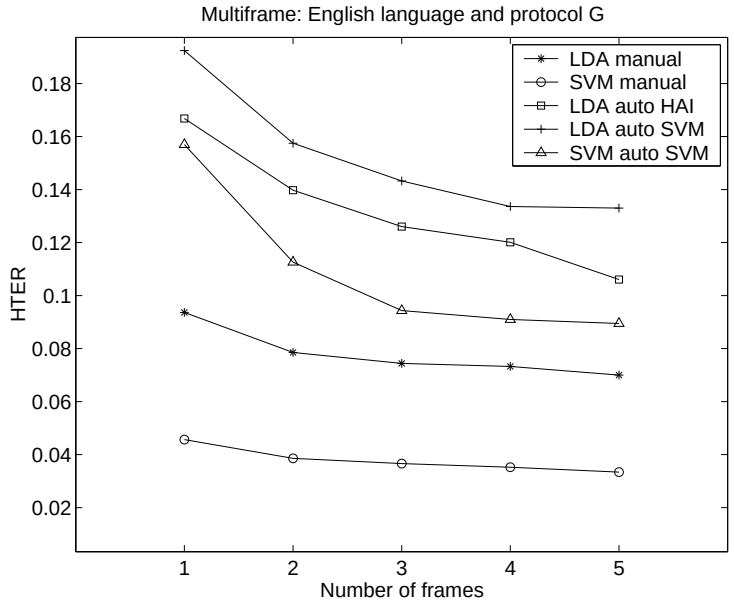


Figure 7.5: *HTER obtained as a function of number of test frames for protocol G on the english part of the BANCA database and minimum integration rule.*

7.7 Conclusions

In this chapter, we have evaluated different techniques for integrating the information extracted from the multiple frames available during a user access. Our approach is based on an extension of the static verification method, that is, the final decision is based a combination of the static score obtained for each frame. It is shown analytically that the average score combination rule lead always to an improvement for Gaussian distributed scores and outperfoms the minimum rule. However, for bimodal genuine score densities which are likely to occur in practice, the minimum rule may outperform the average rule.

Different combination rules were tested and compared on images from the BANCA database. Different face localisation and face verification algorithm were considered for these experiments, namely manual localisation, SVM based localisation, Hypothesis driven Affine Invariant face localisation, LDA face verification and SVM face verification. From the results, it appears that the average and minimum rules always improve the error rates in comparison with static verification. In the manual locali-

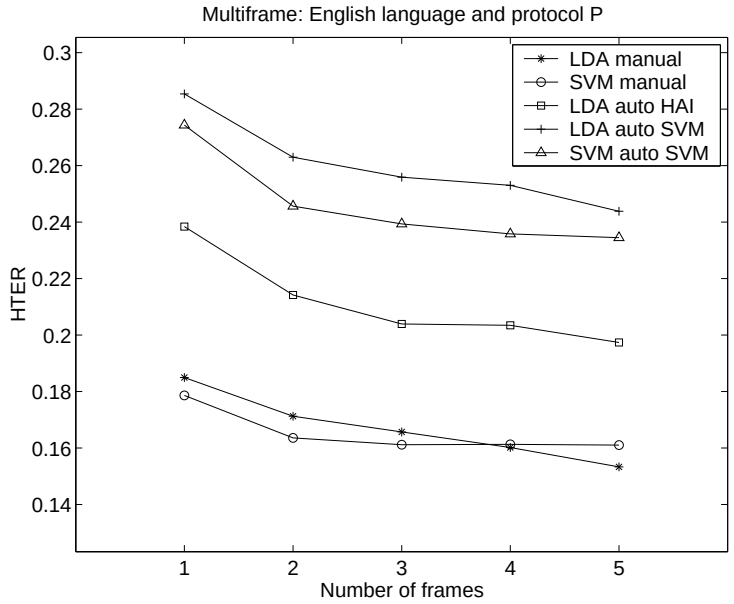


Figure 7.6: *HTER obtained as a function of number of test frames for protocol P on the english part of the BANCA database.*

sation case and when three training images are available (protocol G), the average rule may outperform the minimum rule. In the automatic localisation case, the minimum rule is clearly superior. This is in accordance with the analytical results. The automatic localisation may fail or output an inaccurate face position. Since both LDA and SVM are known to be sensitive to mis-registrations, genuine accesses are likely to be rejected (high score) while impostor accesses are certainly rejected. As a result the genuine score distribution has a heavy tail and the minimum rule is advantageous in this case.

Integration results as a function of number of test frames have also been presented. It is shown experimentally that in most cases the error rate is monotonically decreasing with the number of test frame. This is an important result as it shows that the decision should be based on as many test frames as possible, the only limiting factor being the available hardware resources, e.g. the time required to process the multiple frames. It is however likely that above a certain number of test frames, the performance improvement will saturate.

Chapter 8

Conclusions and perspectives

We have studied the problem of personal identity verification using facial images. Several appearance based face verification algorithms have been developed and their properties thoroughly studied, first analytically, and then empirically on two different standard face databases using well defined testing protocols.

The studied algorithms include standard template matching, "eigenfaces" or Principal Component Analysis (PCA) based face verification, Linear Discriminant Analysis (LDA) based verification and the Probabilistic Matching. The choice of these specific algorithms was motivated by their reputation, ease of implementation and known performance (LDA based and Probabilistic Matching based algorithms showed the best performance at the FERET face recognition competition [95]). Influence of factors such as the matching scheme or the image size on the verification performance have been studied.

The verification algorithm in which facial images are matched in the Linear Discriminant Analysis (LDA) subspace has proven to lead to the best verification results on both databases. Although optimal in theory, the Euclidean distance is outperformed by the normalised correlation (NOC). Note that both the Euclidean distance and NOC suppose a uniform impostor probability density function. Trainable matching techniques, which explicitely model the impostor class, appear to be inaccurate with respect to NOC. The main difference between the Euclidean distance and NOC is that the latter does not depend on the norm of the two vectors to compare. Hence, it seems that identity is more related to the *orientation* (with respect to the average face) in the face representation space, rather than to exact position in the space. However, the method remains seriously affected

by large pose and illumination condition differences or, in other words, by the inadequacy between recording conditions of the enrolment and testing phases.

Interestingly, all tested algorithms seem to be unaffected by low-pass filtering and down-sampling of the images. This suggests that only low frequency information is used to make the decision. How to take advantage of the high frequencies to increase verification accuracy (small details in facial features, edges) remains an open problem.

To improve verification performance we suggested that decisions coming from different sources of information about the identity should be combined or fused. This idea was illustrated by combining at the score level (i) a speaker and a face verification algorithm (multi-modal fusion) (ii) several face verification algorithms (intramodal fusion) (iii) several face scores corresponding to several video frames during the same access. In all cases, if the fusion method is designed carefully, the fused system achieved better verification results than the face verification algorithm alone.

The fusion can be performed by combining the class *a posteriori* probability estimates given by each classifier as it is advocated in classifier combination theory, or by training a classifier directly in the score space – trainable combiner approach, which is used in practice. A simple method for obtaining a class *a posteriori* probability estimate from the algorithm score has been proposed. When the classifiers are independent (i.e. the scores given by each classifier are class conditionally independent), the product rule is the optimal. We studied the behaviour of different *a posteriori* probability combination rules when classifiers are no more independent but positively correlated. It appears that these rules are relatively robust to correlation between the classifiers and achieve similar performance as trainable combiners while being more modular. Trainable combiners and probability combination rules were tested on the face verification problem by combining five different face verification algorithms. As the five classifiers output a score based on the same image, the scores are very correlated. Moreover classifiers have different accuracies. 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 that the multiple intramodal expert combination can increase the robustness of identity verification based on facial images. Interestingly, the simplest trainable combiner, which is the weighted averaging, seems

to give comparable, or even better results than more complex classifiers in the score space.

If the classifiers use different modalities to calculate their respective score, we can consider that the scores are very little correlated. In that case the fusion is very effective and a substantial improvement can be obtained even if the individual expert accuracies differ by a factor three. We showed empirically that the combination of a face verification algorithm (HTER = 15%) with a text independent speaker verification algorithm (HTER = 5%) achieves better performance (HTER = 3%) than the speaker verification algorithm alone in spite of the difficulty of the data. Again, it appeared that, on the multimodal data, the weighted averaging combiner gives results comparable to those using a fusion based on Multi-Linear Perceptron. Using a combination of independent sources of information, such as the speech and face modalities in our case, provides independent features. The size global face-speech biometric template is in fact much smaller, at same performance level, than the speech biometric alone. This shows that looking for independent sources of information introduces efficient features and may help reducing the feature vector size. It is important in applications where storage space is of concern and the feature vector must be as small as possible, e.g. the implementation of biometrics on smart cards.

The third aspect of decision fusion considered in the thesis concerns the combination of scores related to the same sensor which measures sequentially over time several instances of the same physical phenomenon. In the identity verification context, this is illustrated by the fusion of scores corresponding to successive frames recorded during the user access. In contrast to the two preceding cases, the scores are coming from the same verification algorithm and all of them have therefore the same importance. This constatation leads to two simple combination rules, namely the score average rule and the score minimum rule or best matching rule. It is shown analytically that the average rule always leads to an improvement for Gaussian distributed scores and outperforms the minimum rule. However, for genuine score density which have a heavy tail, the minimum rule may outperform the average rule. Experiments on real data confirm the analytical results. Using the BANCA database, five video frames are used to make the decision. In the manual face localisation case, the average rule may outperform the minimum rule. In the automatic localisation case, where the localisation may fail or output an inaccurate face position, the minimum rule is clearly superior. Since both LDA and SVM based

face verification algorithms are known to be sensitive to mis-registrations, genuine accesses have higher scores resulting in a heavy tail genuine score distribution.

8.1 Perspectives

Identity verification using facial images is still in its infancy. While the algorithms presented in this thesis achieve accurate results on controlled data, the uncontrolled enviroment case is still problematic. Several extensions in both face verification and fusion are worth investigating.

Component-based face verification

We noticed that the global LDA approach ignores high frequency information. In our opinion the high frequencies useful for classification can be incorporated in the algorithm by using a hybrid appearance-feature based approached. Such approaches include component-based face recognition [47, 49]: patches containing facial features are automatically extracted from the image using several SVM classifier, each trained to locate a given facial feature. Instead of matching images, patches are matched. Changings in head pose mainly lead to changes in the relative position of the facial components which does not disturb the verification as patches are matched separately in constrast to the global image which is entirely affected by changes in head pose.

Synthetic training face images

Verification algorithms are severly affected by differences in pose and illumination conditions between template and test images. Increasing the number of training images per person used for enrolment (as in the protocol G of the BANCA database) allows to reduce the error rate significantly. An alternative is to generate synthetic images of a person under various illumination conditions and poses from one single frontal view. This can be done by using a 3D model of the face that is "morphed" to the current image as in [10] and related work. An SVM classifier can be trained using the new genuine examples and impostor examples. The advantage over [10] is that the generation of synthetic image can be done offline in the verification scenario. Also the model image is aquired during enrolment where pose, illumination and synthetic image generation can be controlled by a human operator.

Intramodal fusion: creating diversity in classifier opinions

It is well known that classifier combination is more efficient when the classifiers do not agree on patterns. In our approach, the face classifiers were optimised independently, without taking into account the error diversity. Method such as AdaBoost, that were proposed to generate diversity in classifier opinions, could be deployed in conjunction with component-based approach to enhance the verification.

Joint tracking and verification

The approach taken in this thesis is modular: the general problem of face verification has been decomposed in different modules, each one implementing a sub-task, executed sequentially: face localisation, face normalisation (geometric - photometric), feature extraction, matching with template and final decision. The localisation and verification steps are explicitely separated. Recently, it has been proposed to integrate the localisation/tracking and verification step in one single task [11, 127]. The advantage is that the temporal identity information is intrinsically incorporated into the method. In this method, video is modelled by a time series state space model, the state vector being the tracking affine parameters of the face to track, and the observations being the video frames or some features (for example Gabor jets) extracted from it. The goal (tracking/verification) is to compute the posterior distribution of the state vector given the observations up to current time. This is solved by Bayesian recursive estimation and the Sequential Importance Sampling (SIS) technique [50] is invoked to generate a numerical solution. The posterior distribution of identity, which is needed to make the decision, is estimated by marginalizing the posterior distribution of the state vector over a proper affine region around the posterior mean. This method is computationally more efficient than ours because a detection algorithm is only needed to give an approximate position of the face in the first frame.

Multimodal fusion: incorporating speech into the dynamical model

Our multi-modal fusion is done at the score level. However, we could take advantage of the fact that both speech and video are dynamical processes. The time series state space model described in the previous paragraph could be extended to incorporate the speech information. The observation vector would consists of facial features extracted from the video stream

and speech features extracted from the speech stream.

Appendix A

Optimality of the product rule

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 5.2.2. 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 $Z \sim \mathcal{N}(0, 1)$, i.e.

$$\phi(x) = P(Z < x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du.$$

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 Equation (5.1), which can be written in this case: choose class ω_a if

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

where $\mathbf{s} = (s_1, s_2, ..., s_R)^T$. Because of (A.1), we have

$$\boldsymbol{\mu}_a = \boldsymbol{\mu}_b + K\mathbf{v}$$

with $\mathbf{v} = (1, 1, ..., 1)^T$ and the optimal rule becomes

$$K\mathbf{s}^T\Sigma^{-1}\mathbf{v} > K(\boldsymbol{\mu}_a + \boldsymbol{\mu}_b)^T\Sigma^{-1}\mathbf{v}.$$

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

$$\Sigma = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \dots & \dots & \dots & \dots \\ \rho & \dots & \rho & 1 \end{pmatrix}.$$

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 is the dimension of Σ or equivalently the number of experts. The vector $\mathbf{v}$ is also an eigenvector of matrix Σ^{-1} with eigenvalue $C(\rho)^{-1}$ because

$$\begin{aligned} \Sigma\mathbf{v} &= C(\rho)\mathbf{v} \\ \Sigma^{-1}\Sigma\mathbf{v} &= C(\rho)\Sigma^{-1}\mathbf{v} \\ C(\rho)^{-1} &= \Sigma^{-1}\mathbf{v} \end{aligned}$$

Therefore the optimal rule becomes simply

$$\mathbf{s}^T\mathbf{v} > (\boldsymbol{\mu}_a + \boldsymbol{\mu}_b)^T\mathbf{v}.$$

In this particular case, the optimal decision rule does not depend on the correlation between the experts. The product rule, which is optimal if $\rho = 0$ is therefore also optimal for any value of correlation.

Appendix B

Detailed verification results of multiframe score integration

In this appendix, we present the detailed results of the multiframe score integration presented in Chapter 7. The verification experiments were performed on the BANCA database according to the BANCA protocol.

Prot.	Algorithm	Fusion	HTER g1	HTER g2	average HTER
G	LDA manual	min	7.16	6.84	**7.00**
G		arit.mean	7.37	6.84	7.10
G		median	8.76	6.57	7.67
G		max	8.23	9.40	8.81
G		geom.mean	7.37	6.84	7.10
G		single frame	10.84	9.03	9.94
G	SVM manual	min	2.83	3.85	**3.34**
G		arit.mean	3.26	3.74	3.50
G		median	3.69	4.49	4.09
G		max	5.50	3.95	4.73
G		geom.mean	3.04	3.74	3.39
G		single frame	4.11	4.70	4.41

Table B.1: *Protocol G for manually detected faces*

Prot.	Algorithm	Fusion	HTER g1	HTER g2	average HTER
P	LDA manual	min	16.45	14.21	**15.33**
P		arit.mean	17.63	15.17	16.40
P		median	18.38	16.19	17.28
P		max	19.39	17.52	18.46
P		geom.mean	17.47	15.17	16.32
P		single frame	19.87	19.50	19.68
P	SVM manual	min	14.53	17.68	**16.11**
P		arit.mean	15.12	19.28	17.20
P		median	15.60	18.70	17.15
P		max	17.15	20.19	18.67
P		geom.mean	15.12	19.07	17.09
P		single frame	17.57	20.62	19.10

Table B.2: *Protocol P for manually detected faces*

Prot.	Algorithm	Fusion	HTER g1	HTER g2	average HTER
G		min	11.59	9.62	**10.60**
G		arit.mean	13.09	10.95	12.02
G	LDA with	median	14.69	12.34	13.51
G	HAI based	max	16.99	14.90	15.95
G	local.	geom.mean	13.19	10.47	11.83
G		single frame	19.23	15.87	17.55
G		min	15.71	10.90	**13.30**
G		arit.mean	15.81	12.29	14.05
G	LDA with	median	18.91	11.91	15.41
G	SVM based	max	23.93	19.93	21.93
G	local.	geom.mean	15.97	11.65	13.81
G		single frame	23.56	14.64	19.10
G		Min	9.13	8.76	**8.95**
G		Ari.mean	9.78	11.54	10.66
G	SVM with	Median	12.02	12.18	12.10
G	SVM based	Max	20.03	19.87	19.95
G	local.	Geo.mean	9.78	11.11	10.44
G		Single	16.45	16.77	16.61

Table B.3: *Protocol G for automatically detected faces*

Prot.	Algorithm	Fusion	HTER g1	HTER g2	average HTER
P		min	20.14	19.34	**19.74**
P		arit.mean	19.98	21.15	20.57
P	LDA with	median	20.25	20.14	20.19
P	HAI based	max	24.63	24.31	24.47
P	local.	geom.mean	19.44	20.83	20.14
P		single	25.69	25.11	25.40
P		min	26.28	22.49	**24.39**
P		arit.mean	25.00	24.41	24.71
P	LDA with	median	26.01	25.11	25.56
P	SVM based	max	35.10	32.64	33.87
P	local.	geom.mean	25.27	23.82	24.55
P		single	30.02	29.38	29.70
P		Min	23.40	23.50	23.45
P		Ari.mean	24.04	22.22	23.13
P	SVM with	Median	26.44	22.65	24.55
P	SVM based	Max	32.69	29.27	30.98
P	SVM local.	Geo.mean	23.88	22.01	**22.94**
P		Single	28.74	25.05	26.90

Table B.4: *Protocol P for automatically detected faces*

Appendix C

Detailed results on BANCA database

In this appendix, we present the detailed results split in sub-protocols of the face verification experiments presented in Chapter 4. The experiments were performed on the BANCA database according to the BANCA protocol. In all cases, histogram equalision was used for photometric normalisation. Also the score on which the decision is based, is chosen to be the minimum score (see Section 4.3.1).

Matching in the image space

Prot.	Metric	g2		g1		
		FAR	FRR	FAR	FRR	HTER
Mc	NOC	13.46	3.85	6.73	6.73	**10.50**
	EUD	13.46	10.26	13.46	13.46	12.82
	NMM	10.58	8.97	14.42	14.42	12.66
Ud	NOC	13.46	33.33	30.77	30.77	**24.84**
	EUD	14.42	52.56	36.54	36.54	29.41
	NMM	20.19	38.46	35.58	35.58	27.40
Ua	NOC	23.08	38.46	37.50	37.50	30.85
	EUD	14.42	46.15	41.35	41.35	**28.04**
	NMM	22.12	39.74	39.42	39.42	31.41
Md	NOC	16.35	21.79	19.23	19.23	**16.91**
	EUD	13.46	32.05	30.77	30.77	22.28
	NMM	12.50	32.05	31.73	31.73	21.63
Ma	NOC	8.65	34.62	27.88	27.88	**19.71**
	EUD	8.65	41.03	35.58	35.58	24.52
	NMM	14.42	38.46	31.73	31.73	24.68
G	NOC	13.46	21.37	19.87	19.87	**16.35**
	EUD	13.46	35.47	30.13	30.13	22.33
	NMM	14.10	31.20	29.17	29.17	20.75
P	NOC	17.95	31.62	30.13	30.13	**25.48**
	EUD	21.47	34.62	30.45	30.45	26.66
	NMM	28.85	36.75	33.97	33.97	32.05

Table C.1: *BANCA database (English subset): error rates for the different configuration and for normalised correlation (NOC), Euclidean distance (EUD) and nearest mean metric (NMM).*

Matching in the image space with symmetrised images

Prot.	Metric	g2		g1		
		FAR	FRR	FAR	FRR	HTER
Mc	NOC	13.46	10.26	8.65	8.65	10.98
	EUD	12.50	7.69	9.62	9.62	11.30
	NMM	16.35	7.69	11.54	11.54	12.74
Ud	NOC	9.62	34.62	30.77	30.77	21.96
	EUD	10.58	34.62	25.96	25.96	20.67
	NMM	15.38	32.05	28.85	28.85	23.24
Ua	NOC	17.31	43.59	38.46	38.46	29.33
	EUD	19.23	44.87	43.27	43.27	30.69
	NMM	20.19	33.33	31.73	31.73	**25.80**
Md	NOC	10.58	20.51	15.38	15.38	14.18
	EUD	11.54	28.21	28.85	28.85	18.11
	NMM	11.54	29.49	26.92	26.92	19.55
Ma	NOC	7.69	34.62	25.00	25.00	17.79
	EUD	10.58	30.77	25.96	25.96	18.75
	NMM	20.19	48.72	35.58	35.58	28.69
G	NOC	10.90	17.95	15.06	15.06	13.22
	EUD	10.58	26.07	20.83	20.83	16.40
	NMM	16.67	26.92	24.68	24.68	20.70
P	NOC	15.38	29.91	26.92	26.92	22.44
	EUD	18.91	28.21	25.64	25.64	23.32
	NMM	19.23	30.34	25.96	25.96	23.90

Table C.2: *BANCA database (English subset): error rates for the different configuration and for normalised correlation (NOC), Euclidean distance (EUD) and nearest mean metric (NMM).*

Matching in the LDA subspace with symmetrised images

Prot.	Metric	g2		g1			PCA	LDA
		FAR	FRR	FAR	FRR	HTER	dim.	dim.
Mc	EUD	11.54	8.97	14.42	15.38	12.58	330	140
	NOC	4.81	2.56	1.92	7.69	4.25	330	150
	NMM	8.65	10.26	12.50	8.97	10.10	330	140
	GDM	12.50	7.69	11.54	14.10	11.46	330	140
Ud	EUD	20.19	39.74	35.58	16.67	28.04	330	140
	NOC	14.42	19.23	22.12	15.38	17.79	330	150
	NMM	19.23	30.77	32.69	19.23	25.48	330	140
	GDM	25.00	30.77	28.85	17.95	25.64	330	140
Ua	EUD	20.19	37.18	41.35	19.23	29.49	330	140
	NOC	15.38	33.33	35.58	19.23	25.88	330	150
	NMM	22.12	25.64	28.85	23.08	24.92	330	140
	GDM	22.12	35.90	37.50	20.51	29.01	330	140
Md	EUD	14.42	24.36	29.81	8.97	19.39	330	140
	NOC	9.62	15.38	12.50	8.97	11.62	330	150
	NMM	18.27	21.79	25.00	14.10	19.79	330	140
	GDM	18.27	24.36	25.00	21.79	22.36	330	140
Ma	EUD	10.58	32.05	30.77	5.13	19.63	330	140
	NOC	4.81	16.67	12.50	2.56	9.13	330	150
	NMM	7.69	41.03	30.77	7.69	21.79	330	140
	GDM	11.54	39.74	28.85	6.41	21.63	330	140
G	EUD	11.86	29.49	25.64	12.82	19.95	330	140
	NOC	4.81	8.55	8.01	4.70	6.52	330	150
	NMM	10.90	27.35	22.12	9.40	17.44	330	140
	GDM	13.46	26.07	20.19	9.83	17.39	330	140
P	EUD	21.15	31.20	32.37	20.09	26.20	330	140
	NOC	16.99	18.38	17.95	14.10	16.85	360	230
	NMM	21.79	27.35	25.64	20.51	23.82	330	140
	GDM	23.08	28.21	28.21	19.66	24.79	330	140

Table C.3: *BANCA database (English subset): rates for the different configuration and for normalised correlation (NOC), Euclidean distance (EUD) and nearest mean metric (NMM). The images were symmetrised before matching.*

Appendix D

Publications

European Commission Deliverables for the IST BANCA project

We have contributed to the following deliverables:

1. BANCA D2.5 Robustification face verification algorithms,
2. BANCA D2.6a Intramodal combination,
3. BANCA D2.6b Dynamic aspects,
4. BANCA D3.3 Scalability,
5. BANCA D3.2 Classifier Combination.

Journal and Reviewed Conference Publications

1. J. Czyz, J. Kittler and L. Vandendorpe: Multiple classifier combination for face-based identity verification. *Submitted to Pattern Recognition*. April 2003.
2. F. Lefebvre, J. Czyz and B. Macq: A Robust soft hash algorithm for digital image signature. *ICIP 2003 - International Conference on Image Processing*. September 2003.
3. J. Czyz, S. Bengio, C. Marcel and L. Vandendorpe: Scalability of audio-visual person identity verification. *AVBPA 2003 - International conference on Audio- and Video-based Biometric Person Authentication*. June 2003.

4. K. Messer, J. Kittler, M. Sadeghi, S. Marcel, C. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, J. Czyz, L. Vandendorpe, S. Srisuk, R. Paredes, B. Kepenekci, F. Tek, G. Akar, F. Deravi and N. Mavity: Face verification competition on the XM2VTS database *Proc. of Int. Conf. on Audio- and Video-based Person Authentication*. June 2003.

5. C. Havran, L. Hupet, Jacek Czyz, J. Lee, L. Vandendorpe and M. Verleysen: Independent component analysis for face authentication, *KES 2002 - International Conference on Knowledge-based Intelligent Information and Engineering System*. September 2002.

6. J. Czyz, J. Kittler amd Luc Vandendorpe: Combining face verification experts. *ICPR 2002 - International Conference on Pattern Recognition*. August 2002.

7. J. Kittler, M. Ballette, J. Czyz, F. Roli and L. Vandendorpe: Enhancing the performance of personal identity authentication systems by fusion of face verification experts. *ICME 2002 - International Conference on Multimedia and Expo.* August 2002.

8. J. Kittler, M. Ballette, J. Czyz, F. Roli and L. Vandendorpe: Decision level fusion of intramodal personal identity verification experts. *MCS 2002 - 3rd International Workshop on Multiple Classifier*. June 2002.

9. J. Czyz and L. Vandendorpe: Evaluation of LDA-based face verification with respect to available computational resources. *PRIS 2002 - International Workshop on Pattern Recognition in Information Systems*. April 2002.

Bibliography

[1] Y. Adini, Y. Moses, and S. Ullman. Face recognition: the problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):721–732, 1997.

[2] K. Ali and 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.

[3] F. Alkoot and J. Kittler. Experimental evaluation of expert fusion strategies. *Pattern Recognition Letters*, 20, 1999.

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

[5] R. Baron. Mechanisms of human recognition. *Int. Journal of Man Machine Studies*, 15(2), 1981.

[6] M. Bartlett, H. Lades, and T. Sejnowski. Independent component representations for face recognition. In *Proceedings of the SPIE, Vol 3299: Conference on Human Vision and Electronic Imaging III*, pages 528–539, 1998.

[7] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, 19(7):711–720, 1997.

[8] S. Bengio, F. Bimbot, J. Mariéthoz, V. Popovici, F. Porée, E. Bailly-Baillière, G. Matas, and B. Ruiz. Experimental protocol on the BANCA database. IDIAP-RR 05, IDIAP, 2002.

[9] S. Bengio, C. Marcel, S. Marcel, and J. Mariéthoz. Confidence measures for multimodal identity verification. *Information Fusion*, 3(4), 2002.

[10] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH'99 Conference Proceedings*, 1999.

[11] B.Li and R.Chellappa. A generic approach to simultaneous tracking and verification. *IEEE trans. on image processing*, 11(5), May 2002.

[12] R. Bolle, S. Pankanti, and N. Ratha. Evaluation techniques for biometrics-based authentication systems. In *Proceeding of the International Conference on Pattern Recognition (ICPR)*, 2000.

[13] L. Breiman. Bagging predictors. *Machine Learning*, 24, 1996.

[14] R. Brunelli and D. Falavigna. Person identification using multiple cues. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(10):955–966, 1995.

[15] R. Brunelli and T. Poggio. Face recognition: features versus templates. *IEEE Trans. on Pattern Analysis and Machine intelligence*, 15(10), 1993.

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

[17] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *Proceedings of the European Conference on Computer Vision*, 1998.

[18] T. Cootes, C. Taylor, A. Lanitis, D. Cooper, and J. Graham. Building and using flexible models incorporating grey level information. In *Proceedings of the International Conference on Computer Vision*, pages 242–246, 1993.

[19] I. Cox, J. Ghoshn, and P. Yianilos. Feature-based face recognition using mixture distance. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 1996.

[20] J. Czyz, S. Bengio, C. Marcel, and L. Vandendorpe. Scalability analysis of audio-visual person indentity verification. In *Proc. of Int. Conf. on Audio- and Video-based Person Authentication*, 2003.

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

[22] J. Czyz, J. Kittler, and L. Vandendorpe. Multiple classifier combination for face-based identity verification. *Submitted to Pattern Recognition*, 2003.

[23] J. Czyz and L. Vandendorpe. Analysis of lda-based face verification with respect to computational resources. In *Workshop on Pattern Recognition in Information Systems*, 2002.

[24] J. Daugman. Face and gesture recognition: Overview. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):675–676, July 1997.

[25] J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency and orientation optimised by two dimensional cortical filters. *J. Opt. Soc. Amer.*, 2(7):1160–1169, 1985.

[26] P. Devijver and J. Kittler. *Pattern recognition: a statistical approach*. Prentice-Hall, 1982.

[27] T. Dietterich. Ensemble methods in machine learning. In *First International Workshop on Multiple Classifier System, Lecture Notes in Computer Science, J. Kittler and F. Roli (Ed.)*, 2000.

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

[29] B. Duc, E. Bigun, J. Bigun, G. Maitre, and S. Fischer. Fusion of audio and video information for multimodal person authentication. *Pattern Recognition Letters*, 18(9), 1997.

[30] R. Duda, P. Hart, and D. Stork. *Pattern classification, second edition*. Wiley-Interscience, 2001.

[31] R. Duin and D. Tax. Classifier conditional posterior probabilites. In *Advance in Pattern Recognition*, volume LNCS 1451. Springer, 1998.

[32] R. P. W. Duin. The combining classifier: to train or not to train. In *Proc. of Int. Conf. on Pattern Recognition, Quebec, Canada*, August 2002.

[33] K. Etemad and R. Chellappa. Discriminant analysis for recognition of human face images. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 2148–2151, 1994.

[34] J. Fitzpatrick, D. Hill, and C. Maurer. Image registration. In M. Sonka and J. Fitzpatrick, editors, *Handbook of medical imaging, vol. 2*, pages 449–513. SPIE press, 2000.

[35] P. Flynn, K. Bowyer, and P. J. Phillips. Assessment of time dependency in face recognition: a initial study. In *Proc. of Int. Conf. on Audio- and Video-based Person Authentication*, 2003.

[36] B. Frey, A. Colmanarez, and T. Huang. Mixture of local linear subspaces for face recognition. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, June 1998.

[37] K. Fukunaga. *An Introduction to Statistical Pattern Recognition, second edition*. Academic Press, 1990.

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

[39] R. Gonzalez and E. Woods. *Digital image processing*. Addison Wesley, 1992.

[40] D. Gorodnichy. Facial recognition from video. In *Proc. of Int. Conf. on Audio- and Video-based Person Authentication*, 2003.

[41] R. Gross and V. Brajovic. An image preprocessing algorithm for illumination invariant face recognition. In *Proceedings of the International Conference on Audio- and Video-based Biometric Person Authentication*, 2003.

[42] B. W. Group. Best practices in testing and reporting performance of biometric devices, January 2000.

[43] G. Hachez, F. Koeune, and J.-J. Quisquater. Biometrics, access control, smart cards: a not so simple combination. In *Conference on Smart Card Research and Advanced Applications (CARDIS 2000)*, pages 273–288, September 2000.

[44] M. Hamouz, J. Kittler, J. Kamarainen, and H. Kalviainen. Face detection by learned affine correspondences." sspr/spr 2002: pp. 566-575. In *Joint IAPR international workshops on Structural, Syntactic, and Statistical Pattern Recognition, SSPR/SPR. Proceedings, Lecture Notes in Computer Science 2396 Springer*, 2002.

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

[46] C. Havran, L. Hupet, J. Czyz, J. Lee, L. Vandendorpe, and M. Verleysen. Independent component analysis for face authentication. In *KES2002, International Conference on Knowledge-Based Intelligent Information and Engineering Systems*, September 2002.

[47] B. Heisele, T. Serre, M. Pontil, T. Vetter, and T. Poggio. Categorization by learning and combining object parts. In *Advances in Neural Information Processing Systems*, 2002.

[48] L. Hong and A. K. Jain. Integrating faces and fingerprints for personal identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1295–1307, 1998.

[49] J. Huang, B. Heisele, and V. Blanz. Component-based face recognition with 3d morphable models. In *Proc. of Int. Conf. on Audio- and Video-based Person Authentication*, pages 27–34, 2003.

[50] M. Isard and A. Blake. CONDENSATION-conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

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

[52] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: a review. *IEEE Trans. on Pattern Analysis and Machine intelligence*, 22(1), 2000.

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

[54] K. Jonsson. *Robust Correlation and Support Vector Machines for Face Identification*. PhD thesis, University of Surrey, 2000.

[55] K. Jonsson, J. Kittler, Y. P. Li, and J. Matas. Support vector machines for face authentication. In *Proceedings of British Machine Vision Conference*, pages 543–553, 1999.

[56] K. Jonsson, J. Kittler, and J. Matas. Learning support vectors for face authentication: sensitivity to mis-registrations. In *Asian Conference on Computer Vision*, 2000.

[57] T. Kanade. *Computer recognition of human faces*. Birkhauser Verlag, Stuttgart, Germany, 1977.

[58] S. M. Kenna, S. Gong, and Y. Raja. Face recognition in dynamic scenes. In *Proceedings of the British Machine Vision Conference*, 1997.

[59] M. Kirby and L. Sirovich. Application of the Karhunen-Loève procedure for characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.

[60] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, 20(3):226–239, 1998.

[61] J. Kittler, Y. P. Li, and J. Matas. On matching scores for LDA-based face verification. In *Proceedings of British Machine Vision Conference*, 2000.

[62] J. Kittler, J. Matas, K. Jonsson, and M. R. Sanchez. Combining evidences in personal identity verification systems. *Pattern Recognition Letters*, 18(9), 1997.

[63] J. Kittler and F. Roli, editors. *Multiple Classifier Systems, Second International Workshop, MCS 2001 Cambridge, UK, July 2-4, 2001, Proceedings*, volume 2096 of *Lecture Notes in Computer Science*. Springer, 2001.

[64] C. Kotropoulos, A. Tefas, and I. Pitas. Frontal face authentication using morphological elastic graph matching. *IEEE Transactions on Image Processing*, 9(4):555–560, Apr. 2000.

[65] L. Kuncheva. A theoretical study on six fusion strategies. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, 24(2):281–286, 2002.

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

[67] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, Mar. 1993.

[68] A. Lanitis, C. Taylor, and T. Cootes. Towards automatic simulation of ageing effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):442–455, 2002.

[69] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997.

[70] H.-S. Lee, D. Kim, and S.-Y. Lee. Robust face tracking using skin color and facial shape. In *Proc. of Int. Conf. on Audio- and Video-based Person Authentication*, 2003.

[71] S. T. Li and J. Lu. Face recognition using the nearest feature line method. *IEEE Transactions on Neural Networks*, 10(2):439–443, 1999.

[72] S. Z. Li, K. L. Chan, and C. Wang. Performance evaluation of the nearest feature line method in image classification and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1335–1349, 2000.

[73] C. Liu and H. Wechsler. Enhanced fisher linear discriminant models for face recognition. In *Proceeding of the International Conference on Pattern Recognition (ICPR)*, 1998.

[74] C. Liu and H. Wechsler. Face recognition using evolutionary pursuit. In *Proceeding of the European Conference on Computer Vision (ECCV)*, 1998.

[75] C. Liu and H. Wechsler. Evolutionary pursuit and its application to face recognition. *IEEE Trans. on Pattern Analysis and Machine intelligence*, 22(6), 2000.

[76] J. Luettin and 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.

[77] A. Martinez and A. Kak. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, 2001.

[78] J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotroupolous, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Bigun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abduljaoued, and E. Majoraz. Comparison of face verification results on the xm2vts database. In *Proceeding of the International Conference on Pattern Recognition (ICPR)*, 2000.

[79] J. Matas, K. Jonsson, and J. Kittler. Fast face localisation and verification. In *Proceedings of British Machine Vision Conference*, 1997.

[80] S. McKenna, S. Gong, and Y. Raja. Modelling facial colour and identity with gaussian mixtures. *Pattern Recognition*, 31(12):1883–1892, Dec. 1998.

[81] K. Messer, J. Kittler, M. Sadeghi, S. Marcel, C. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, J. Czyz, L. Vandendorpe, S. Srisuk, R. Paredes, B. Kepenekci, F. Tek, G. Akar, F. Deravi, and N. Mavity. Face verification competition on the XM2VTS database. In *Proc. of Int. Conf. on Audio- and Video-based Person Authentication*, 2003.

[82] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: the extended M2VTS database. In *Proc. of Int. Conf. on Audio- and Video-based Person Authentication*, pages 72–77, March 1999.

[83] T. P. Minka. Automatic choice of dimensionality of pca. Technical Report TR 514, MIT Media Laboratory Perceptual computing section, 2000.

[84] B. Moghaddam, T. Jebara, and A. Pentland. Bayesian face recognition. *Pattern Recognition*, 33(11):1771–1782, 2000.

[85] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, 19(7):696–710, 1997.

[86] B. Moghaddam, W. Wahid, and A. Pentland. Beyond eigenfaces: probabilistic matching for face recognition. In *Proc. of Int. Conf. on Automatic Face and Gesture Recognition*, April 1998.

[87] H. Moon and P. J. Phillips. Analysis of PCA-based face recognition algorithms. In *Empirical Evaluation Techniques in Computer Vision*, 1998.

[88] H. Moon and P. J. Phillips. Computational and performance aspects of PCA-based face recognition algorithm. *Perception*, 30:303–321, 2001.

[89] E. Osuna, R. Freund, and G. Gorosi. Training support vector machines: a application to face detection. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 1997.

[90] A. Papoulis. *Probability, Random Variables and Stochastics Processes*. McGraw-Hill, 1965.

[91] J. R. Parks. Personal identification - biometric. In D. Lindsay and W. Price, editors, *Information security*, pages 181–191. North Holland: Elsevier Science, 1991.

[92] P. S. Penev and J. J. Atick. Local feature analysis. *Network: Computation in Neural Systems*, 7(3), 1996.

[93] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 1994.

[94] P. J. Phillips. Matching pursuit filters applied to face identification. *IEEE Transactions on Image Processing*, 7(8):1150–1164, Aug. 1998.

[95] P. J. Phillips, H. Moon, P. Rauss, and S. Rizvi. The FERET evaluation methodology for face recognition algorithm. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 1997.

[96] S. Pigeon and L. Vandendorpe. The M2VTS multimodal face database (release 1.00). In *Proc. of Int. Conf. on Audio- and Video-based Person Authentication*, pages 403–409, 1997.

[97] S. Pigeon and L. Vandendorpe. Image-based multimodal face authentication. *Signal Processing*, 69(1), 1998.
[98] D. Politis. Computer-intensive method in statistical analysis. *IEEE Signal Processing Magazine*, 15(1), 1998.
[99] S. Prabhakar and A. K. Jain. Decision-level fusion in fingerprint verification. *Pattern Recognition*, 35:861–874, 2002.
[100] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, Cambridge (UK) and New York, 2nd edition, 1992.
[101] V. Pugachev. *Théorie des probabilités et statistique mathématique*. Editions de Moscou, 1982.
[102] A. Pujol, J. Vitria, F. Lumbresas, and J. Villanueva. Topological principal component analysis for face encoding and recognition. *Pattern recognition letters*, 22(6-7), 2001.
[103] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE trans. on Speech and Audio Processing*, 3(1):72–83, 1995.
[104] F. Roli and J. Kittler, editors. *Multiple Classifier Systems, Third International Workshop, MCS 2002, Cagliari, Italy, June 24-26, 2002, Proceedings*, volume 2364 of *Lecture Notes in Computer Science*. Springer, 2002.
[105] A. Ross, A. K. Jain, and J.-Z. Qian. Information fusion in Biometrics. In *Proc. Int. Conf. on Audio- and Video-based Person Authentication*, pages 355–359, 2001.
[106] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, 20(1):23–38, 1998.
[107] M. Sadeghi, J. Kittler, A. Kostin, and K. Messer. A comparative study of automatic face verification algorithms on the BANCA database. In *Proceedings of the International Conference on Audio- and Video-based Biometric Person Authentication*, 2003.
[108] R. Sanchez-Reillo. Including biometric athentication in a smart card operating system. In *Proc. of Int. Conf. on Audio- and Video-based Person Authentication*, pages 342–347, 2001.
[109] W. Shen, M. Surette, and R. Khanna. Evaluation of automated biometric-based identification and verification systems. *Proceedings of the IEEE*, 85(9), sept 1997.
[110] T. Sim, S. Baker, and M. Bsat. The CMU Pose, Illumination, and Expression (PIE) database. In *Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2002.
[111] K. K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1997.
[112] D. Swets and J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. on Pattern Analysis and Machine intelligence*, 18(8), 1996.
[113] D. Tax, M. van Breukelen, R. Duin, and J. Kittler. Combining multiple classifiers by averaging or by multiplying? *Pattern Recognition*, 33:1475–1485, 2000.

[114] K. Tumer and J. Ghosh. Error correlation and error reduction in ensemble classifiers. *Connection Science*, 8(3/4), 1996.
[115] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1), 1991.
[116] H. VanTrees. *Detection, Estimation and Modulation Theory*. Wiley, 1968.
[117] P. Verlinde, G. Chollet, and M. Acheroy. Multi-modal identity verification using expert fusion. *Journal of Information Fusion*, 1(1), 2000.
[118] J. Wayman. Confidence interval and test size estimation for biometric data. In *Proceedings of IEEE AutoID conference*, 1999.
[119] J. Wayman. Technical testing and evaluation of biometric identification devices. In A. K. Jain, R. Bolle, and R. Pankanti, editors, *Biometrics: personal identification in a networked society*, pages 345–368. Kluwer academic publisher, 1999.
[120] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, July 1997.
[121] W. S. Yambor, B. A. Draper, and J. R. Beveridge. Analyzing PCA-based face recognition algorithms: eigenvectors selection and distance measure. In *Empirical Evaluation Techniques in Computer Vision*, 2000.
[122] M.-H. Yang, D. J. Kriegman, and N. Ahuja. Face detection using multimodal density models. *Computer Vision and Image Understanding*, 84, 2001.
[123] M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Trans. on Pattern Analysis and Machine intelligence*, 24(1), 2002.
[124] W. Zhao. Improving the robustness of subspace face recognition system. In *Proc. of Int. Conf. on Audio- and Video-based Person Authentication*, pages 78–83, March 1999.
[125] W. Zhao and R. Chellappa. Robust face recognition using symmetric shape-from-shading. Technical Report CAR-TR-919, Center for automation research, University of Maryland, 1999.
[126] W. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. In *Proc. of Int. Conf. on Automatic Face and Gesture Recognition*, pages 336–341, 1998.
[127] S. Zhou and R. Chellappa. Probabilistic human recognition from video. In *Proc. of Int. Conf. on Automatic Face and Gesture Recognition*, 2002.