IDIAP RESEARCH REPORT

FACE AUTHENTICATION BASED ON LOCAL FEATURES AND GENERATIVE MODELS

Fabien Cardinaux [a]

IDIAP–RR 05-85

JANUARY 2006

(a) cardinau @ idiap.ch

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 05-85

FACE AUTHENTICATION BASED ON LOCAL FEATURES AND GENERATIVE MODELS

Fabien Cardinaux

JANUARY 2006

Abstract

The principal objective of this thesis is to investigate approaches toward a robust automatic face authentication (AFA) system in weakly constrained environments. In this context, we develop new algorithms based on local features and generative models. In addition, particular attention is given to face localization which is a necessary step of a fully automatic system.

In an authentication scenario, a person claims an identity and, using one or several face images to support this claim, the system classifies the person as either a true claimant (called client) or as an impostor. Unlike face identification, the face authentication task aims to assign a given face image into one of two classes. This task is particularly difficult since any person can be encountered; *ie.* the impostors have usually not been seen before. One of the other major challenges of AFA is the lack of reference images. Indeed, it is not realistic to have a huge amount of images for each identity. Usually, only one or a few images are available and they can not cover all the possible variabilities due to different expression, lighting, background, head pose, hair cut, etc.

Generative models such as Gaussian mixture models (GMMs), one-dimensional hidden Markov models (1D-HMMs) and pseudo two-dimensional hidden Markov models (P2D-HMMs) have proved to be efficient for face identification. In this thesis, we propose to train generative models using maximum *a posteriori* (MAP) training instead of the traditionally used maximum likelihood (ML) criterion. We experimentally demonstrate the superiority of this approach over other training schemes. The main motivation for the use of MAP training is the ability of this algorithm to estimate robust model parameters when there is only a few training images available. Using P2D-HMM trained with MAP, we obtain better performance than state-of-the-art face authentication approaches.

In a second part of this thesis, we proposed some improvements of the baseline systems in order to increase performances with minimal effects in computation time. The first proposition is to extend the feature vectors for the GMM approach in order to embed positional information. This new system improves slightly the performances comparing to the baseline GMM approach. The second proposed approach is an alternative 1D-HMM topology which allows the use of observation vectors representing image blocks instead a whole line for standard 1D-HMM implementation. The experiments demonstrate that this model is significantly more robust than the standard 1D-HMM. Due to is low complexity, it is also eight times faster than a P2D-HMM with the cost of a lower accuracy.

Finally, in the last part of the thesis, we propose a new methodology to evaluate face localization algorithms in the context of face authentication. We first show the influence of localization errors on face authentication systems and then empirically demonstrate the problems of current localization performance measures when applied to this task. In order to properly evaluate the performance of a face localization algorithm, we then propose to embed the final application (the authentication system) into the performance measuring process. We show that our proposed method to evaluate localization algorithms better matches the final authentication performance.

Version Abrégée

Le principal objectif de cette thèse est d'explorer des approches pour un système d'authentification de visages automatique (AFA) dans un environnement faiblement controllé. Dans ce contexte, nous developpons de nouveaux algorithmes basés sur des caractéristiques locales et des modèles génératifs. De plus, une attention particulière est portée sur la localisation des visages qui est une étape nécessaire pour un système entièrement automatique.

Dans un scénario d'authentification, une personne revendique une identité et, en utilisant une ou plusieurs images de visage de cette personne, le système classifie l'accès comme un accès client si la personne est effectivement la personne revendiquée ou comme un accès imposteur. Au contraire de l'identification, l'authentification de visages a pour but d'assigner une image de visage a une classe parmi deux (le client revendiqué ou un imposteur). Cette tache est particulièrement difficile car on ne connait pas à l'avance les personnes que le système peut rencontrer; *ie.* les imposteurs n'ont généralement jamais été vus auparavant. Un autre des défis importants d'AFA est le manque d'images de références. En effet, il n'est pas réaliste d'avoir un grand nombre d'images pour chaque identité à disposition. Généralement, seule quelques images, parfois une seule, sont disponibles et elles ne peuvent pas couvrir toutes les variabilités possibles dues aux changements d'expression, d'éclairage, d'arrière plan, de position de la tête, etc.

Les modèles génératifs comme les mixtures de Gaussiennes (GMMs), les modèles de Markov cachés à une dimension (1D-HMMs) et les modèles de Markov cachés pseudo bi-dimensionels (P2D-HMMs) ont prouvé leur efficacité pour l'identification de visages. Dans cette thèse, nous proposons d'entraîner des modèles génératifs en utilisant un entraînement par maximisation *a posteriori* (MAP) à la place de la maximisation de la vraisemblance (ML) habituellement utilisée. Nous montrons expérimentalement la supériorité de cette approche par rapport aux autres méthodes d'entraînements. La motivation principale pour l'utilisation de l'entraînement par MAP est la capacité de cet algorithme à estimer des modèles robustes lorsque seules quelques images d'entraînements sont disponibles. En utilisant les P2D-HMM entraînés avec MAP, on obtient de meilleures performances que les approches états de l'art en authentification de visages.

Dans une deuxième partie de cette thèse, nous proposons des améliorations des systèmes de bases de façon à augmenter les performances avec un effet minimal sur le temps de calcul. La première proposition est d'étendre les vecteurs caractéristiques de façon à inclure des informations de positions. Ce nouveau système améliore légèrement les performances en comparaison avec l'approche de base utilisant les GMMs. La deuxième approche proposée est une topologie de 1D-HMM différente qui permet l'utilisation de vecteurs d'observations représentant des blocs de l'image à la place de lignes complètes pour des 1D-HMM standards. Les expériences montrent que ce modèle est significativement plus robuste que les 1D-HMMs standards. Ce nouveau système est aussi huit fois plus rapide que le système à base de P2D-HMM mais avec des performances inférieures.

Finalement, dans la dernière partie de la thèse, nous proposons une nouvelle méthodologie pour évaluer les algorithmes de localisation de visages dans le contexte d'une authentification de visages. Nous montrons d'abord l'influence des erreurs de localisation sur le système d'authentification de visages et nous montrons ensuite empiriquement les problèmes des mesures de performances actuellement utilisées en localisation lorsque la tâche finale est l'authentification de visages. De façon à proprement évaluer la

performance des algorithmes de localisation de visages, nous proposons d'inclure l'application finale (le système d'authentification) dans la mesure de la performance. Nous montrons que la méthode proposée pour évaluer les algorithmes de localisation est plus en adéquation avec la performance finale du système d'authentification.

Contents

1 Introduction **13**
1.1 Biometric Recognition 13
1.2 Face Recognition 14
1.3 Challenge 14
1.4 Objective and Contributions of this work 15
1.5 Organization of the thesis 16

2 State of the Art in Face Recognition **17**
2.1 Face Recognition 17
2.1.1 Face Detection and Localization 17
2.1.2 Lighting Normalization 18
2.1.3 Feature Extraction for Face Recognition 19
2.1.4 Classification Task 23
2.2 The Face Authentication Task 24
2.2.1 Evaluation of Face Authentication Systems 25
2.2.2 Client-specific Threshold and Score Normalization 26
2.3 Databases and Experimental Protocols 26
2.3.1 The Extended M2VTS Database (XM2VTS) 27
2.3.2 The BANCA database 28

3 Adapted Generative Models for Face Authentication **31**
3.1 Generative Model Based Classifiers 32
3.1.1 Gaussian Mixture Model 32
3.1.2 1D Hidden Markov Model 32
3.1.3 Pseudo-2D HMM 33
3.2 Maximum *a Posteriori* (MAP) Training 34
3.3 Preliminary Steps 36
3.3.1 Face Localization 36
3.3.2 Pre-processing and Feature Extraction 37
3.4 Experiments and Discussion 37
3.4.1 Manual Face Localization 40
3.4.2 Imperfect and Automatic Localization 40
3.4.3 Number of Training Images 42
3.4.4 Complexity of Models 43
3.5 Conclusions 45

4 Toward Fast and Robust Face Authentication **47**
4.1 Embedding Positional Information for GMM approach 47
4.1.1 Proposed Feature Vectors 47
4.1.2 Results 48
4.2 Local Features based 1D-HMM 48
4.2.1 Proposed Model 48
4.2.2 Feature extraction 50
4.2.3 Results and Discussions 50
4.2.4 Vertical and Horizontal HMM 51
4.2.5 Conclusion 51

5 On the Choice of a Good Face Localization System **53**
5.1 Performance Measures for Face Localization 54
5.1.1 Lack of Uniformity 54
5.1.2 A Relative Error Measure 54
5.1.3 A More Parametric Measure 55
5.1.4 Application-Dependent Measure 55
5.2 Robustness of Current Measures 55
5.2.1 Effect of Face Localization Errors 56
5.2.2 Indetermination of current measure 57
5.3 Approximate Face Authentication Performance 59
5.4 Experiments and Results 61
5.4.1 Training Data 61
5.4.2 Face Localization Performance Measure 61
5.4.3 Evaluation 62
5.5 Conclusion 63

6 Conclusion and Perspectives **65**
6.1 General Summary 65
6.2 Possible Future Directions 66

A Acronyms **67**

List of Figures

1.1 Terminology of Biometric Recognition 14

2.1 Typical Face Recognition System 18
2.2 Holistic and Local Feature Extraction 20
2.3 Graph Matching Example 24
2.4 Adapted Graphs for Faces in Different Views 24
2.5 Examples of Curves Representing the Performance of Face Authentication Systems 26
2.6 Example of images from the XM2VTS 27
2.7 The Lausanne Protocol 28
2.8 Examples of Images from the BANCA Database 30

3.1 Sampling Window and 1D-HMM Topology 34
3.2 P2D-HMM Representation 35
3.3 EPC of GMM Based System Trained via MAP Adaptation 38
3.4 EPCs for Manual Face Localization 41
3.5 Performance for an Increasing Amount of Error in Eye Locations 43
3.6 Performance as a Function of the Number of Original Training Images 44

4.1 1D-HMMs Interpreted as P2D-HMM 49

5.1 Summary of some Basic Measurements Made in Face Localization 54
5.2 Conceptual Representations of the Two Face Authentication Systems 56
5.3 Face Authentication Performance as a Function of Face Localization Errors 58
5.4 Face Bounding Boxes 60
5.5 Face Localization Scanning Parameters: Step x, Step y and Scale Factor 62

List of Tables

2.1 Number of Images Used for the Two Lausanne Protocols 29
2.2 Usage of the Seven BANCA Protocols 29

3.1 Optimum Parameters for Systems Based on GMM, 1D-HMM and P2D-HMM 39
3.2 HTER performance for **manual face localization**, and **automatic face localization**, using GMM, 1D-HMM and P2D-HMM 42
3.3 Complexity of the Models 45

4.1 HTER Performance of GMM with Extended Features for **Manual Face Localization** 48
4.2 HTER Performance of GMM with Extended Features for **Automatic Face Localization** . . . 48
4.3 HTER Performance for **Manual Face Localization** 50
4.4 HTER Performance for **Automatic Face Localization** 51
4.5 HTER Performance for Vertical and Horizontal 1D-HMM 51

5.1 Delta Values and Corresponding HTERs 57
5.2 Comparison of Two Face Localization Performance Measures 63

Chapter 1

Introduction

1.1 Biometric Recognition

Biometric person recognition involves the use of known information about her/his intrinsic characteristics. In other words, the recognition is accomplished by showing ***something you are***; this is in contrast with traditional methods where you show ***something you have***, such as a licence or a passport or methods involving ***something you know*** such as a password or a Personal Identification Number (PIN).

The most popular biometrics used for recognition include: fingerprint, face, voice, iris, DNA, etc.

In a wide range of application, biometric recognition is preferred over traditional methods for various reasons. Indeed, biometric characteristics are difficult to imitate, some of them are unique, the person has in general to be physically present and it obviates the need to remember a password or to carry a token.

Applications include surveillance, forensics, transaction authentication, and various forms of access control, such as border checkpoints and access to digital information [2, 49, 61, 96].

There are three distinct configurations of how a biometric recognition system can be used: the closed set identification task, the open set identification task, and the authentication task (also known as verification). In the **closed set identification** task, the job is to forcefully classify a given biometric sample as belonging to one of K persons (here K is the number of *known* persons). In **open set identification**, the task is to assign the given sample into one of $K + 1$ classes, where the extra class represents an "unknown" or "previously unseen" person. Finally, in the **authentication or verification** task the classifier assigns a given sample into one of two classes: either the sample belongs to a specific person, or it doesn't. In an access control scenario this translates to a person claiming an identity and providing a biometric sample to support this claim; the authentication system then classifies the person as either a true claimant or as an impostor.

The authentication task represents operation where *any* person/pattern could be encountered [43]. This is in contrast to the closed set identification task, where it is assumed that all the persons that are going to be encountered are already known.

Note that in the literature some authors use the term *recognition* as a synonym of *identification* [18, 97, 23], in this thesis, *recognition* is used as a generic term to refer to three different configurations defined before. Figure 1.1 summarizes the terminology used in this thesis.

Further introductory and review material about the biometrics field can be found in the following papers: [23, 61, 82, 93, 96].

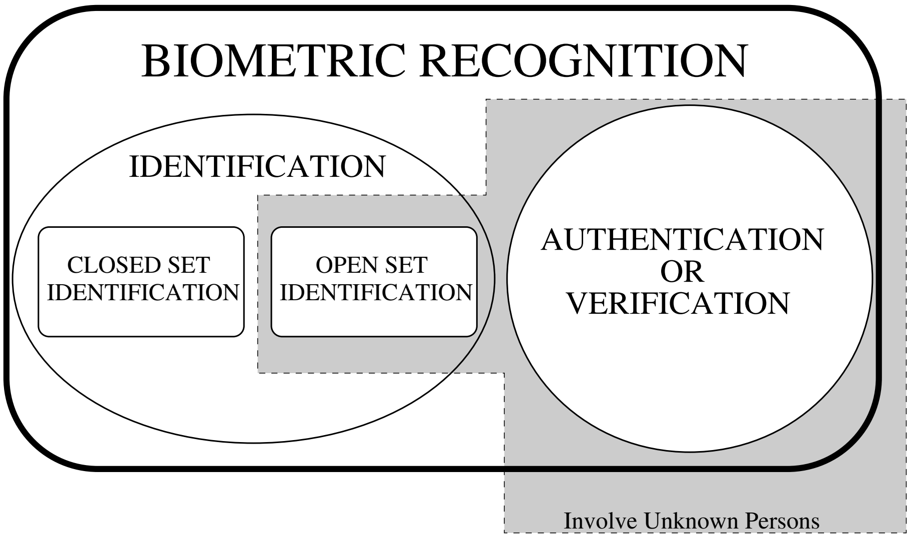


Figure 1.1: Terminology of Biometric Recognition. Biometric recognition is used as a generic term to refer to three different configurations: Open Set Identification, Closed Set Identification and Authentication (also called Verification).

1.2 Face Recognition

In this thesis we exclusively focus on recognition based on face images. The use of the face as a biometric is particularly attractive, as it can involve little or no interaction with the person to be recognized [61]. However, while humans seem to recognize faces with relative ease, automatic face recognition performed by a machine proved to be very difficult. The primary difficulty in recognizing faces arises because of the large variations occurring in a single face. Indeed, it is well admitted that variations between different faces are smaller than variations that occur in a single face in different conditions. For example, two different faces in the same lighting conditions can be more similar than the same face in various lighting conditions. This makes face recognition very challenging.

1.3 Challenge

In spite of the expanding research in the field of face recognition, a lot of problems are still unsolved. Today, several systems that achieve high recognition rates have been developed, however, such systems work in controlled environments; for most of them, face images must be frontal or profile, background must be uniform and lighting must be constant. Furthermore, lot of published systems are evaluated using manually located faces and the ones which have been evaluated using a fully automatic system showed a big degradation in performances [66]. In most real life applications, the environment is not known *a-priori* and the system should be fully automatic. A weakly constrained Face Recognition system has to deal with the following problems:

- Lighting Variation
- Head Pose changes
- Non-Perfect Detection
- Occlusion

- Aging

In this thesis, we will focus on face recognition approaches toward robust face recognition in weakly constrained environments (WCE). Here, WCE means that the illumination of the face, the head position and the background are not known *a priori* and can change from the reference images to the probe images. In [33], Gross *et al* point out the problems of face recognition in a WCE.

In this context, the face recognition task presents new difficulties such as large variability in the images for the same identity, the lack of reference images and face alignment problems. In fact, to our knowledge, no existing face recognition system can combine speed, accuracy and robustness in unconstrained environment.

The main problem of face authentication in WCE is the large variability between face images. We can define two kinds of variabilities [58]:

- *extra-personal* variabilities: variations in appearance between different identities.
- *intra-personal* variabilities: variations in appearance of the same identity, due to different expression, lighting, background, head pose, hair cut, etc.

For purposes of face recognition, the useful information is *extra-personal* variability and we can consider *intra-personal* variability to be noise. An other important difficulty is the lack of reference images. Indeed, it is not realistic to have a huge amount of images for each identity. Usually, only a few reference images are available and they can not cover all the possible *intra-personal* variabilities. This is the main difficulty of using statistical models for face recognition task.

1.4 Objective and Contributions of this work

This thesis aims to develop a robust automatic face authentication (AFA) in condition as close as possible to real life. In addition we have tried to limit as much as possible the constraints imposed to the user and to the system. In this thesis we focus our attention to model based approaches and more precisely to generative models. This choice is motivated by good performances obtained in previous works [24, 33, 59, 62]. Generative models methods include Gaussian Mixture Model (GMM) [15], one dimensional hidden Markov model (1D-HMM) [77] and Pseudo two-dimentional hidden Markov model (P2D-HMM) [24, 59]. While in the literature, many papers provide biased performances (e.g. using *a posteriori* threshold) or use face localized by hand, we evaluate performances using automatic localization as well as fair and unbiased protocols for different levels of constraints, from more controlled to weakly controlled conditions. In the following, we briefly discuss what are the main contributions resulting from the present thesis:

- **Maximum *a Posteriori* (MAP) training:** We proposed to train generative models using MAP training instead of the traditionally used Maximum Likelihood (ML) criterion [15, 13, 14]. We experimentally demonstrated the superiority of this approach over other training schemes. The main motivation for the use of MAP training is the ability of this algorithm to estimate robust model parameters when there is only a few training images available. Using pseudo two-dimensional HMM (P2D-HMM) [24, 59, 77] trained with MAP, we obtained better performance than state-of-the-art face authentication approaches. Moreover, the proposed system obtained the best performance in an international face authentication competition [56] for the fully automatic systems category.

- **Extended local features with embedded positional information:** We improved Gaussian Mixture Model (GMM) approach through the use of local features with embedded positional information [13, 14]. Using traditional local features and GMMs results in the loss of spatial information. As the spatial relations can carry discriminatory information, we proposed to increase the performance of the GMM approach (without sacrificing its simplicity) by restoring a degree of spatial relations via embedding positional information into each feature vector.

- **Local Features and one-dimensional HMM (1D-HMM):** We proposed a 1D HMM approach [11] which allows the use of local features and, in consequence, is more robust to imperfect localizations.
- **A unified evaluation of the state-of-the-art generative models (GMs) based approaches:** Recently, several variations of GMs have been proposed for face recognition [62, 78, 24, 33, 59, 77]. Compared to the evaluations presented in previous works, this thesis presents several additional aspects:
  - The evaluations are performed for a face authentication task, while in [62, 24, 33, 59, 77] results are presented for a closed set face identification task.
  - In previous works some model parameters are decided arbitrarily; e.g. the number of states of the Hidden Markov Models (HMMs) presented in [77, 59] is five since the face is composed with five horizontal main regions, namely the forehead, the eyes, the nose, the mouth and the chin. This could lead to sub-optimal parameters. In our experiments, the parameters have been chosen in a systematic way using an validation set.
  - Experiments have been performed with images recorded in weakly constrained environment which reveals some drawbacks of these approaches.
  - Evaluation of the effect of different factors such as the quality of the detection and the number of training examples available.
  - A comparison with a Principal Component Analysis (PCA) based system has been performed.
- **A measure to evaluate face localization systems:** We empirically demonstrated the problems of current localization performance measures when applied to the task of face authentication. We then proposed to *embed* the face authentication into the performance measuring process [73, 74].

1.5 Organization of the thesis

The thesis is organized as follows.

Chapter 1 introduces the problems of face authentication and related issues. The objective and contributions of the thesis are also presented.

Chapter 2 gives a survey of state of the art in the domain of face recognition. Most popular approaches at each step of a face recognition system are presented. A short introduction of face detection will be presented. In addition, the evaluation strategy and a description of the face databases used along the thesis are given.

Chapter 3 proposes the use of maximum *a posteriori* to train generative models instead of the traditionally used maximum likelihood training. It presents also an extensive comparison of three different types of generative models in terms of performance, robustness and complexity.

In Chapter 4 we propose two different approches to achieve fast and robust face authentication systems. In a fist time the use of extended feature vectors with embedded positional information is investigated. In a second part, an alternative generative 1D HMM is presented and evaluated.

Chapter 5 demonstrates the inaccuracy of current face localization quality measures and propose a methodology to evaluate the face localization algorithms for the specific task of face authentication.

Chapter 6 summarizes the results obtained so far and outlines promising directions for future research.

Chapter 2

State of the Art in Face Recognition

The goal of this chapter is to provide background information about automatic face recognition (AFR) in general and also about specifics of the automatic face authentication (AFA) task. In a first part we describe most popular approaches for each step of a face recognition system (Face Detection, Face Normalization, Feature Extraction and Classification). Then, the particularities of the AFA task are presented and finally we will give a description of the face databases and evaluation protocols used along the thesis.

2.1 Face Recognition

As it is often done in complex engineering problems, the face recognition system is divided into a set of more tractable sub-tasks. The basic structure of an AFR system is shown in Figure 2.1. The main parts are typically face detection and face recognition which can itself be decomposed in normalization, feature extraction and classification steps. In this section we briefly present the detection and normalization modules, then, we introduce the most popular techniques for the feature extraction and classification steps which are more in the scope of this thesis.

2.1.1 Face Detection and Localization

Most of fully AFR systems presented in the literature involve a separate face detection or localization step. Exceptions include algorithms based on Elastic Graph Matching algorithms [45, 95] where the localization is involved in the classification process (see Section 2.1.4 for more details). Numerous approaches have been proposed to tackle the problems of face detection and localization. See [97] for a recent survey. We can differentiate *face detection* from *face localization*. *Face detection* aims to determine whether or not there are any faces in the image and, if present, return the face location while the goal of *face localization* is to estimate the position of a single face. The localization is defined as a simplified detection problem with the assumption that an input image contains only one face [97, 57]. On the other hand, localization usually needs to provide a more precise position of the face than detection and can be more difficult in this case.

An AFR system can either use face detection or localization depending of the final application. Note that in many face recognition studies, it is often assumed that the detection step has been performed perfectly, however this assumption is not realistic. A non-perfect face detection step can highly affect the face recognition performances and this problem should be taken into account.

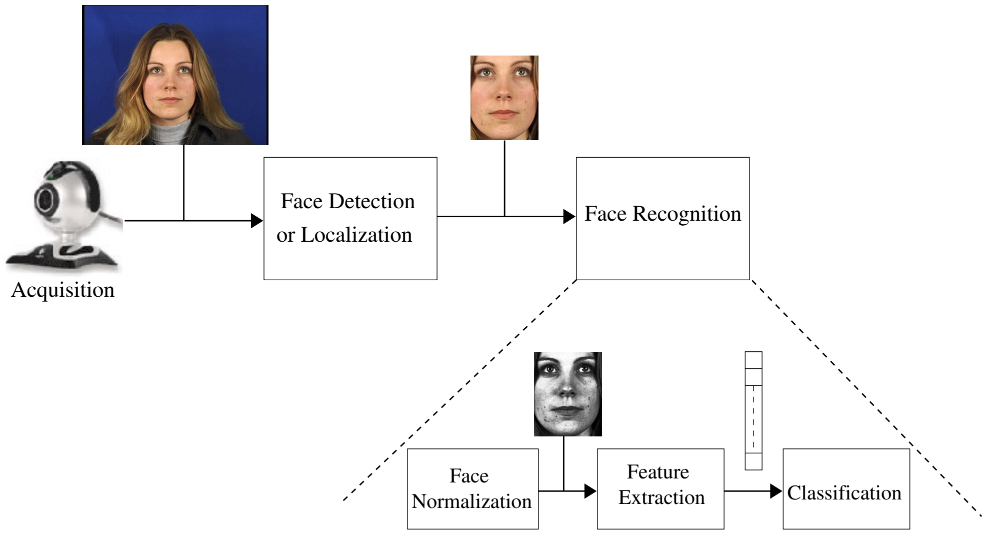


Figure 2.1: Block Diagram of a Typical Face Recognition System

2.1.2 Lighting Normalization

One of the most challenging problem to develop a robust AFR system in a non- or weakly-constrained environment is intra-personal variability due to illumination changes. Even if many papers propose to make the system independent of intra-personal variabilities at the feature extraction [5, 1, 80, 36] or classification [65, 10, 29] levels, most of robust FR systems in the literature also involve preprocessing step to deal with this problem. At the feature extraction stage, the goal is to find an invariant representation of the face image. However, by comparing edge maps, derivatives of the gray level or 2D Gabor features, Adini *et al* [1] empirically demonstrated that none of the representations considered is sufficient by itself to overcome lighting variations. This observation was later formally proved in [16], where the authors showed that there is no function of an image that is discriminative and illumination invariant. At classification step, the algorithm usually need to have examples of face images of the same person with the same pose but under different illuminations. Using preprocessing to deal with illumination variations is very popular since these algorithms can be develop independently of the methods used in the following stages (feature extraction, classification).

Preprocessing algorithms for lighting normalization include general image processing such as histogram equalization, gain/offset correction, non linear transforms (eg. logarithm transform) of the image intensity and Homomorphic Filtering. A brief description of these algorithms is presented in [71].

A second category of lighting normalization approaches are methods based on Land's "retinex" theory [46]. In this theory, the image $I(x, y)$ is regarded as product of the reflectance $R(x, y)$ and the illuminance $L(x, y)$ at each point (x, y):

$$I(x, y) = R(x, y)L(x, y) \tag{2.1}$$

Where the illuminance $L(x, y)$ is assumed to be a low pass version of the image $I(x, y)$. The retinex algorithm thus consists in estimating the reflectance as the ratio of the image and its low pass version. Note that here, the terms *reflectance* and *illumination* are used by analogy with biology phenomena, even if it may not be strictly correct in physical sense. Several implementations have been proposed in the literature in order to find an estimate of the reflectance.

Center/Surround Retinex In [41], Jobson *et al.* propose a single scale retinex algorithm where the reflectance value is given by the ratio of the treated pixel with a weighted average of the intensity in the surrounding area. The derived reflectance value could be expressed as:

$$R(x, y) = \log I(x, y) - \log[I(x, y) * G_s(x, y)] \tag{2.2}$$

where $G_s(x, y)$ denotes a gaussian filter with a variance s. This model was then extended to the multiscale case, which consists basically in performing the same operation at different scales [40]:

$$R(x, y) = \sum_{s=1}^{S} \Bigg(\log[I(x, y)] - \log[I(x, y) * G_s(x, y)] \Bigg) \tag{2.3}$$

Gross and Brajovic's algorithm Based on the common assumption that the luminance $L(x, y)$ can be estimated as a blurred version of the original image, Gross and Brajovic [32] propose to recover $L(x, y)$ by minimizing the following energy-based model:

$$E(L) = \iint_{\Omega} \rho(x, y)(L(x, y) - I(x, y))^2 dxdy + \lambda \iint_{\Omega} (L_x^2 + L_y^2) dxdy \tag{2.4}$$

where the first term forces the luminance function to be close to the image and the second term adds a smoothness constraint on $L(x, y)$. Ω refers to the image, $\rho(x, y)$ is the anisotropic diffusion coefficients and the parameter λ controls the relative importance of the smoothness constraint. The Euler-Lagrange equation for this calculus yields:

$$L(x, y) + \frac{\lambda}{\rho(x, y)} \Bigg(\frac{\partial^2 L(x, y)}{\partial x^2} + \frac{\partial^2 L(x, y)}{\partial y^2} \Bigg) = I(x, y) \tag{2.5}$$

In the discrete case, Euler-Lagrange equation 2.5 becomes:

$$L_{i,j} + \lambda \Bigg[\frac{1}{\rho_{i,j-}}(L_{i,j} - L_{i,j-1}) + \frac{1}{\rho_{i,j+}}(L_{i,j} - L_{i,j+1}) + \frac{1}{\rho_{i-,j}}(L_{i,j} - L_{i-1,j}) + \frac{1}{\rho_{i+,j}}(L_{i,j} - L_{i+1,j}) \Bigg] = I_{i,j}$$

where the anisotropic coefficient is defined as the Weber's contrast:

$$\rho_{a,b} = \frac{|I_a - I_b|}{\min(I_a, I_b)} \tag{2.6}$$

Regarding comparative studies [36, 83] of preprocessing algorithms for lighting normalization, Gross and Brajovic's approach performs the best for AFR and results are significantly improved comparing to traditional methods such as histogram normalization.

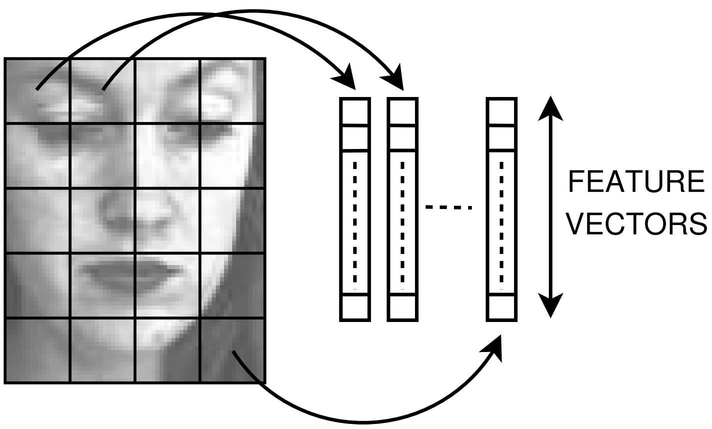
2.1.3 Feature Extraction for Face Recognition

The goal of feature extraction is to find a specific representation of the data that can highlight relevant information. This representation can be found by maximizing a criterion or can be a pre-defined representation. Usually, an image is represented by a high dimensional vector containing pixel values (holistic representation) or a set of vectors where each vector contains gray levels of a sub-image (local representation). Figure 2.2 represents examples of local and holistic feature extraction.

Typically, the vectors are projected into a new space (the feature space), then, the least relevant features can be removed to reduce the dimension of the feature vector according to a criterion (such as lowest amount of variance).

In this section we present an overview of the most relevant feature extraction techniques for face recognition.

(a) Holistic Representation: one vector represents the whole image



(b) Local Feature Extraction: each vector represents a part of the face image

Figure 2.2: Examples of holistic (a) and local (b) feature extraction

Holistic Representation

One feature extraction technique, based on *Principal Component Analysis* (PCA), was first used for face recognition by Turk and Pentland [89]. The aim of PCA is to find a representation of the data minimizing the reconstruction error. The PCA finds the orthogonal directions that account for the highest amount of variance. The data is then projected into the subspace spanned by these directions. In practice, the principal component axes are the eigenvectors of the covariance matrix of the data. The corresponding eigenvalues indicate the proportion of variance of the data projections along each direction.

Another feature extraction method used in face recognition is based on *Linear Discriminant Analysis* (LDA, also known as Fisher Discriminant Analysis) [47, 99]. The LDA subspace holds more discriminant features than the PCA subspace. LDA finds a subspace in which the variability is maximized between different class data, and at the same time where variability in the same class data (face images of the same identity) is minimum.

We define the *within-class scatter matrix* as :

$$S_w = \sum_{j=1}^{c}\sum_{i=1}^{N_j}(\mathbf{x}_i^j - \mu_j)(\mathbf{x}_i^j - \mu_j)^T \quad (2.7)$$

where $\mathbf{x}_i^j$ is the ith sample of class$j$ and μ_j is the mean of class j. And we define the *between-class scatter matrix* as:

$$S_b = \sum_{j=1}^{c}(\mu_j - \mu)(\mu_j - \mu)^T \quad (2.8)$$

where μ is the mean of all classes. The goal is to maximize the between-class measure while minimizing the within-class measure. One way to do this, is to maximize the ratio $\frac{det|S_b|}{det|S_w|}$. Intuitively, for face recognition, LDA should outperform PCA because it inherently deals with class discrimination. However, Martinez and Kak [55] have shown that PCA might outperform LDA when the number of samples per class is small.

Local Representations

In the approaches described in the previous section, the representations were found using the statistics of the entire image. A number of researchers have argued that local filters are more robust than global representation for face image analysis [64, 31]. In this section we will present three kinds of local filters used in face recognition: Local PCA, two-dimensional (2D) Gabor filters and 2D discrete cosine transform (DCT).

Local PCA: Padgett and Cottrell [63] found in 1997 that local PCA based feature extraction was more effective than the global PCA approach. A set of sub-windows are taken from random locations in the training dataset images. Then, the principal component of these small sub-windows are found. The first p principal components were then used to filter the full images.

2D Gabor Wavelets: An alternative to local features such as local PCA is pre-defined local filters such as families of Gabor filters. Gabor functions were extended to two-dimension by Daugman in 1985 [19]. Gabor filters are known as good feature detectors and such filters remove most of the variability in images that is due to variations in lighting. Representations based on Gabor wavelets have been used successfully to recognize facial identity in images [45, 66]. The feature vector at a given point (x_0, y_0) of an image F is typically composed with outputs of Gabor filters $\Psi(y_0, x_0, \omega, \theta)$ at multiple spacial scales ω and orientations θ. In practice, a set of image locations is selected; these locations can be, for example, the nodes of a grid placed over a given face image F. At each location, the inner product of F with each member of the family is computed:

$$P_{j,k} = \int_y \int_x \Psi(y_0 - y, x_0 - x, \omega_j, \theta_k) F(y, x)\, dx\, dy \quad (2.9)$$

for $j = 1, 2, ..., N_\omega$ and $k = 1, 2, ..., N_\theta$. Here, the node is located at (y_0, x_0). A feature vector of dimension $N_\omega N_\theta$ for the location (y_0, x_0) is then constructed using the modulus of each inner product:

$$\overrightarrow{x} = [\, |P_{1,1}|\, |P_{1,2}| \cdots |P_{1,N_\omega}| \cdots |P_{2,1}|\, |P_{2,2}| \cdots |P_{2,N_\omega}| \cdots |P_{N_\theta,N_\omega}|\,]^{\,T} \quad (2.10)$$

2D Discrete Cosine Transform: The face images are analyzed on a block by block basis. Each block is decomposed in terms of 2D Discrete Cosine Transform (DCT) basis functions [30]. A feature vector for each block is then constructed with the DCT coefficients. An extension of 2D DCT, referred to as DCT-mod2, has been proposed in [81]. Compared to traditional DCT feature extraction, the first three DCT coefficients are replaced by their respective horizontal and vertical deltas in order to reduce the effects of illumination direction changes. Further description of DCT-mod2 feature extraction can be found in Chapter 3.

2.1.4 Classification Task

The classification step consists of attributing a label to the input data. This step differs according to the specific task (closed or open set identification, authentication). For an identification task, the label is the identity and eventually an *unknown identity* for an open set identification, while for authentication the label is either true (client) or false (impostor). In all cases, the recognition system typically provides a score $\Lambda_I(X)$ corresponding to an opinion on the probe face pattern X to be the identity I. For closed set face identification purpose, we can recognize identity I^* corresponding to the probe face pattern X as follows:

$$I^* = \arg\max_I \Lambda_I(X) \tag{2.11}$$

For face authentication purpose, the decision is reached as follows: given a threshold τ, the claim is accepted when $\Lambda_I(X) \geq \tau$ and rejected when $\Lambda_I(X) < \tau$.

Open set identification can be considered as a combination of the two previously described tasks. The recognized identity I^* corresponding to the probe face is found as follows:

$$I^* = \begin{cases} \text{unknown} & \text{if } \Lambda_I(X) < \tau \; \forall \; I \\ \arg\max_I \Lambda_I(X) & \text{otherwise} \end{cases} \tag{2.12}$$

The different classification methods are simply different ways to estimate $\Lambda_I(X)$. The different approaches can thus generally be used independently of the specific task. In this section we will explore representative approaches to compute the score $\Lambda_I(X)$.

Similarity Measure

Current approaches to image matching for object recognition often make use of simple image similarity metrics such as the Euclidean distance between the reference image (or the mean of the reference images) and the test image. Because of curse of dimensionality problem, the distance metrics are not computed in the image space but in an appropriate subspace such as PCA or LDA (see Section 2.1.3). The matching score between the probe face image and the identity I can be set to be inversely proportional to the distance between the feature vector of the probe image and the feature vector of the reference images of the identity I. More appropriate metrics have been proposed in the literature such as Mahalanobis distance [8] or Normalized correlation [44]. Note that the metric named *Angle* in [8] is similar to the normalized correlation since the two functions actually compute the cosine of the angle between feature vectors of the probe and the test images.

Elastic Graph Matching

Lades et al. [45] proposed an approach for face recognition using Gabor filters called *Elastic Graph Matching*. In this approach a face is represented by a labeled graph. The graph is a rectangular grid placed on the image (Fig. 2.3) where nodes are labeled with responses of Gabor filters in several orientations and several spatial frequencies called *jets*. The edges are labeled with distances $\Delta\vec{x}_e = \vec{x}_n - \vec{x}_{n'}$ where edge e connects node n with n'.

Comparing two faces is accomplished by adapting and matching the graph R of a reference image to the graph G of the test image.

The adaptation is performed by minimizing the cost function:

$$C(G, R) = \lambda \cdot \sum_{i \epsilon N} d_n(G_{n_i}, R_{n_i}) + \sum_{j \epsilon E} d_e(G_{e_j}, R_{e_j})$$

Where $d_n(G_{n_i}, R_{n_i})$ is the difference between the *jets* for the *ith* node G_{n_i} of the probe grid G and the *jets* for the *ith* node R_{n_i} of the reference grid R; $d_e(G_{e_j}, R_{e_j})$ denotes the difference between the edge j of the grid

G and the edge j of the grid R; N and E are respectively the number of nodes and edges. The coefficient λ controls the rigidity of the image graph (large values penalizing distortion). First, the face location is found by positioning the reference grid on the test image to minimize the cost function (this operation is performed using a rigid model corresponding to $\lambda \to \infty$). In a second step, the position of each individual node is varied to minimize the cost function.

Then, the quality of a match is evaluated by the final result of the cost function. The matching score $\Lambda_R(G)$ can be set to be inversely proportional to the final cost function $C(G, R)$.

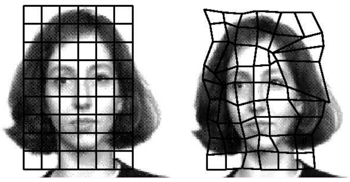


Figure 2.3: Example of grid matching. (a) reference grid, (b) matched grid

In 1997, Wiskott et al. introduced a new data structure called *bunch graph* [95]. They used a face adapted grid (Fig. 2.4) where the nodes refer to specific facial landmarks, called *fiducial points* (eyes, mouth, nose, ...). This new system is able to handle larger galleries and larger variations in pose and to improve recognition rate as compared to the previous graph structure (Fig. 2.3).

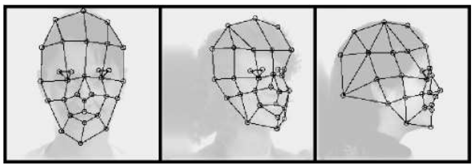


Figure 2.4: Adapted graphs for faces in different views

Statistical Model Based Approaches

Model-driven algorithms are usually more robust than classical approaches however they require a training process. Basically, a model is trained from a set of reference images for each identity, and the score is then computed given a probe image and the parameters of the model corresponding to an identity.

The score can be computed using **discriminant models** [42, 52, 12] (such as Multi-Layer Perceptrons or Support Vector Machines). We will assume that we have access to a training dataset of l pairs (X_i, y_i) where X_i is a vector containing the pattern, while y_i is the class of the corresponding pattern. For face recognition task, we train one model per identity, y_i being coded as $+1$ for patterns corresponding to this identity and as -1 for patterns corresponding to an other identity. The main drawback of using discriminant models for face recognition is the difficulty to train them with a small training dataset.

An alternative to discriminant models are **generative models**. This approach estimates the likelihood of the face image being a specific identity using models representing identities. This approach is presented in details in Chapter 3.

2.2 The Face Authentication Task

While the previous section presents techniques that can be applied for any particular face recognition application, this section presents some specifics of the AFA (also valid for any biometric authentication task).

Authentication is a binary classification task which either accept or reject an identity claim. Since a system is generally able to deal with many clients, it is an aggregation of binary classifications. The evaluation of the system and the strategy used to choose the decision thresholds are specific to the authentication task and are presented in this section.

2.2.1 Evaluation of Face Authentication Systems

Performance measure

Authentication systems make two types of errors: a False Acceptance (FA), which occurs when the system accepts an impostor, or a False Rejection (FR), which occurs when the system refuses a true claimant. The performance is generally measured in terms of False Acceptance Rate (FAR) and False Rejection Rate (FRR), defined as:

$$\text{FAR} = \frac{\text{number of FAs}}{\text{number of impostor accesses}} \tag{2.13}$$

$$\text{FRR} = \frac{\text{number of FRs}}{\text{number of true claimant accesses}} \tag{2.14}$$

The FAR and FRR are usually related, meaning that decreasing one increases the other. To aid the interpretation of performance, FAR and FRR are often combined using the Half Total Error Rate (HTER), defined as:

$$\text{HTER} = \frac{\text{FAR} + \text{FRR}}{2} \tag{2.15}$$

A particular case of the HTER, known as the Equal Error Rate (EER), occurs when the system is adjusted (e.g. via tuning the threshold) so that FAR=FRR.

Trade-off between FAR and FRR

Since FAR and FRR are related and decreasing one has the consequence to increase the other, the setting of an authentication system will depend on the situation. In some situations it may be more important to have a system with a very small FAR, while in other situations a small FRR might be more important. The decision threshold (introduced in Section 2.1.4) is the major parameter that influences the trade-off between FAR and FRR. In order to be able to evaluate the system for a specific situation a Weighted Error Rate (WER) is defined as:

$$\text{WER}(\tau^*) = \omega \text{FAR}(\tau^*) + (1-\omega)\,\text{FRR}\,(\tau^*) \tag{2.16}$$

where $\omega \in [0,1]$ is set for a specific situation and τ^* is the threshold that minimizes the WER for a given ω. For fair performance evaluation, the theshold that minimizes the WER should be chosen on a separate dataset (not the one used to evaluate the system); this threshold is then named *a priori* threshold. In the literature, we can often observe that threshold is set on the same dataset than the one used for performance evaluation; in this case, it is an *a posteriori* threshold. Note that *a posteriori* threshold leads to biased results; along this thesis we always provide results with *a priori* threshold.

Graphical Performance Representation

In order to see performance with respect to the trade-off between the FAR and FRR, researchers often use Receiver Operating Characteristics (ROC) curve [87], which represents the FRR as a function of the FAR. An other version of the plot is the Detection Error Tradeoff (DET) curve [54], which is a non-linear transformation of the ROC curve in order to make results easier to compare. If the scores of client accesses and impostor accesses follow a Gaussian distribution, the DET curve is a line. However, it has been recently observed that these curves can be misleading [6] as they do not take into account that, in real life, the threshold has to be selected *a priori*. We prefer the use of the *Expected Performance Curve* (EPC) [6], that can be interpreted as an unbiased version of the ROC curve. For each value of ω in Equation 2.16, the threshold τ^* is first found on the validation set; the HTER is then found on the test set and is plotted as a function of ω. Figure 2.5 shows examples of ROC, DET and EPC curves.

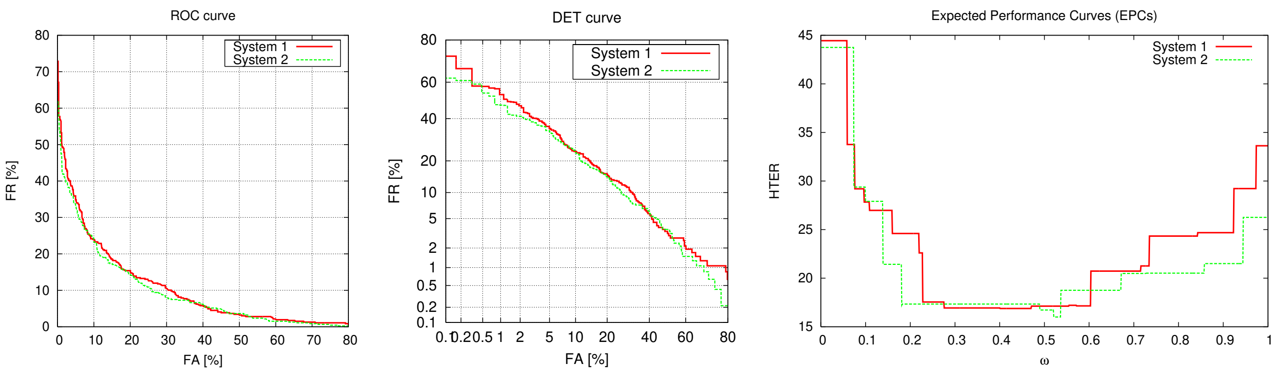


Figure 2.5: Examples of curves representing the performance of face authentication systems. From left to right: Receiver Operating Characteristics (ROC), Detection Error Tradeoff (DET) curves and Expected Performance Curves (EPCs)

.

2.2.2 Client-specific Threshold and Score Normalization

In theory we can estimate a specific threshold for each client in the database, however in practice it requires many client access scores for each single identity which is usually not the case. As a consequence, most of face authentication papers found in the literature use a global threshold (common for all the clients). Nevertheless, several techniques exist to handle client-specific threshold. These techniques are surveyed in [68]. Moreover, this paper shows that client specific threshold normalization is strongly related to score normalization; ie. manipulating the threshold or the scores has the same effect.

On one hand, it is usually not possible to obtain a large number of client access scores for each identity since it would require a lot of client data, on the other hand an impostor can be any identity (except the claimed identity) and it is easy to obtain a large number of impostor access scores. Therefore, one popular approach is to derive the client specific threshold τ_I from the distribution of impostor access scores. The client specific threshold can be computed as:

$$\tau_I = \mu_I + \triangle \cdot \sigma_I \tag{2.17}$$

where μ_I and σ_I are respectively the mean and the standard deviation of the impostor access scores distribution and $\triangle$ is a global parameter selected to minimize WER. This approach has been used for face authentication in [42]. The dual function in terms of score normalization is the Z-norm [68].

2.3 Databases and Experimental Protocols

Many databases of face images are available and the choice of the ones used in order to evaluate the performances of an AFA system is not trivial. Actually, There is no ideal database and the test data should represent as closely as possible the data encountered in real life and thus depends on the final application. A large variability occurs in the existing face databases in terms of camera quality, time lapse between the different

Figure 2.6: Example of images from the XM2VTS. Each row is a subject and each column correspond to a different session recorded over a period of 5 months.

images and level of control of the environment. The level of control of the environment includes illumination conditions, background (uniform or noisy), head pose variability and if there is a possibility of occlusions or not. In [18], the author claims that *experimental results are only valid for a given dataset* and that the comparison of two algorithms should be done on the same data using the same testing protocol. This motivated the choice of using the XM2VTS and the BANCA databases to perform experiments along this thesis. Indeed, these two databases are publicly available, popular and are associated with a well defined protocol which allows easy comparison with other algorithms. Furthermore, while the XM2VTS data are recorded with controlled conditions, the BANCA database contains images from different conditions, from controlled to degraded.

2.3.1 The Extended M2VTS Database (XM2VTS)

The XM2VTS database contains synchronized video and speech data from 295 subjects, recorded during four sessions taken at one month intervals. Two images have been extracted from the video sequence of each session, conducting to a total of 2360 none compressed color images of size 720×576 pixels coded on 24 bits in RGB. Figure 2.6 shows examples of the XM2VTS database. The set of images was recorded in controlled conditions with blue uniform background and controlled lighting. The main intra-personal variabilities come from expression changes and time lapse between sessions.

The Lausanne Protocols (LPs). Associated with the XM2VTS database, the LPs [50, 51] were proposed to evaluate multi-modal (face and speaker) authentication systems. The database is divided into three sets: a

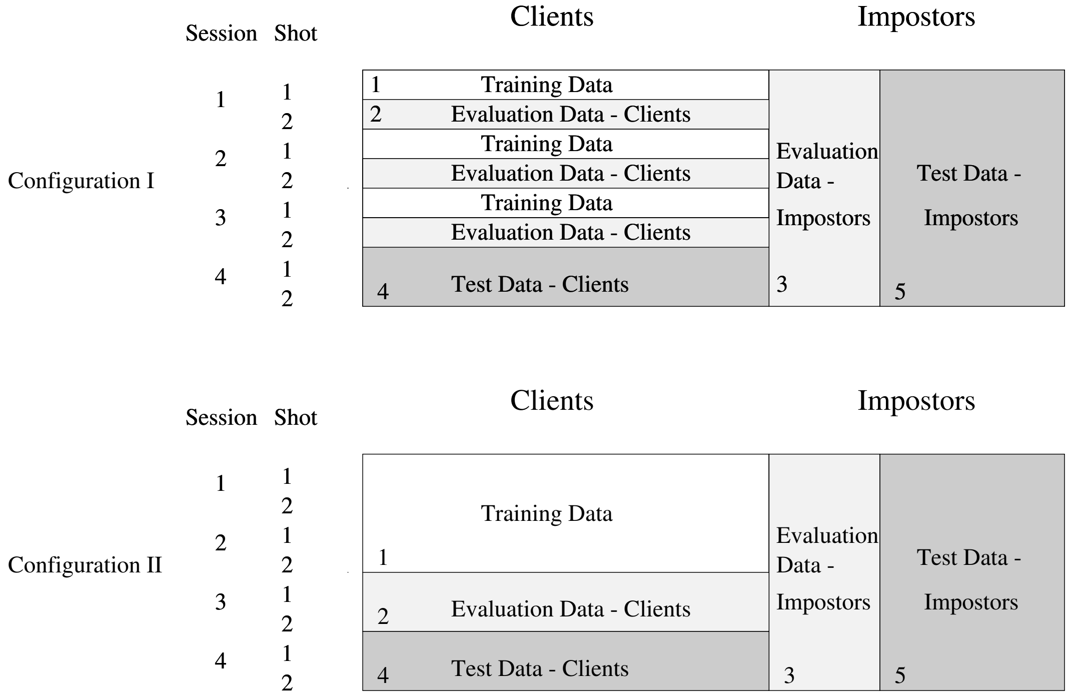


Figure 2.7: Partitioning of the XM2VTS database according to the Lausanne protocol Configuration I (top table) and II (bottom table)

training set, an validation set and a test set [1]. The training set is used to build client models, while the validation set is used to compute the decision thresholds (as well as other hyper-parameters). Finally, the performances are evaluated on the test set.

The 295 subjects were divided into a set of 200 clients, 95 impostors (25 for the validation set and 70 for the test set). Two configurations specify client images used for training, validation and test; the distribution of the impostors in validation and test sets is kept constant for the two protocols. Figure 2.7 reports the repartition of the images in the different sets and Table 2.1 is the summary of the number of data used for each step of the evaluation protocols.

In Configuration I (LP1), the first image of the three first sessions compose the training set, the second image of the same sessions are used for validation and images from the fourth session are used to test the system.

In Configuration II (LP2), all images of sessions one and two are used for training, the third session constitutes the validation set and the last session is used to test the system.

2.3.2 The BANCA database

The more recent BANCA Database [3] follows the main idea of the XM2VTS database of providing multi-modal data associated with a well defined protocol for Biometric Authentication. However, the main motivation of the European project BANCA was to record data in realistic scenarios. The considered scenarios could be for example a network transaction or an Automatic Teller Machine (ATM) where the conditions of

[1] The terminology is not consistent in the evaluation protocols associated with the XM2VTS [50] and the BANCA [3] databases. To be consistant, along this thesis we use "*training set*", "*validation set*" and "*test set*" which correspond respectively to "*training set*", "*evaluation set*" and "*test set*" in the original Lausanne protocol description [50] and to "*training set*", "*development set*" and "*evaluation set*" in the original BANCA protocol description [3].

Table 2.1: Number of images used for each dataset for the two Lausanne protocols. LP1 and LP2 refers respectively to the configurations I and II.

Datasets	Lausanne Protocols	
	LP1	LP2
Training client images	3	4
Validation client accesses	600 (3×200)	400 (2×200)
Validation impostor accesses	$40,000$ ($25 \times 8 \times 200$)	
Test client accesses	400 (200×2)	
Test impostor accesses	$112,000$ ($70 \times 8 \times 200$)	

illumination, background and the quality of the camera can not be controlled. The database is composed with five separate corpora, each containing 52 subjects; the corpora are named after their country of origin (English, French, German, Italian and Spanish). For each corpus, the dataset is splitted in two separate groups (g1 and g2) of 13 males and 13 females. Note that an additional set of 30 other persons (15 males and 15 females) have been recorded in order to make up the *world data* set. Each subject participated in 12 recording sessions over several months, in different conditions and with different cameras. Each of these sessions contains two video recordings: one true claimant access and one impostor attack. Five "frontal" (not necessarily directly frontal) face images have been extracted from each video recording. Sessions 1-4 contain images for the *controlled* condition, while sessions 5-8 and 9-12 respectively contain *degraded* and *adverse* conditions. The latter two conditions differ from the *controlled* condition in terms of image quality, lighting, background and pose. See Figure 2.8 for an example of the differences. Unlike the XM2VTS database, the BANCA database contains high variability in illumination, pose, resolution, background and quality of the camera.

The BANCA Protocols. According to the original experiment protocols [3], there are seven distinct configurations that specify which images can be used for training and testing: Matched Controlled (Mc), Matched Degraded (Md), Matched Adverse (Ma), Unmatched Degraded (Ud), Unmatched Adverse (Ua), Pooled test (P) and Grand test (G). Table 2.2 describes the usage of different sessions in each configuration. Unlike the LP associated with the XM2VTS database, two different sets of identity are used for the validation

Table 2.2: Usage of the seven BANCA protocols (C: client, I: impostor). The numbers refer to the ID of each session.

Test Sessions	Train Sessions			
	1	5	9	1,5,9
C: 2-4 I: 1-4	Mc			
C: 6-8 I: 5-8	Ud	Md		
C: 10-12 I: 9-12	Ua		Ma	
C: 2-4,6-8,10-12 I: 1-12	P			G

and test sets, which guaranty that the hyper parameters are not tuned for the specific identities present in the database. The two groups g1 and g2 are thus used alternatively as validation and test set.

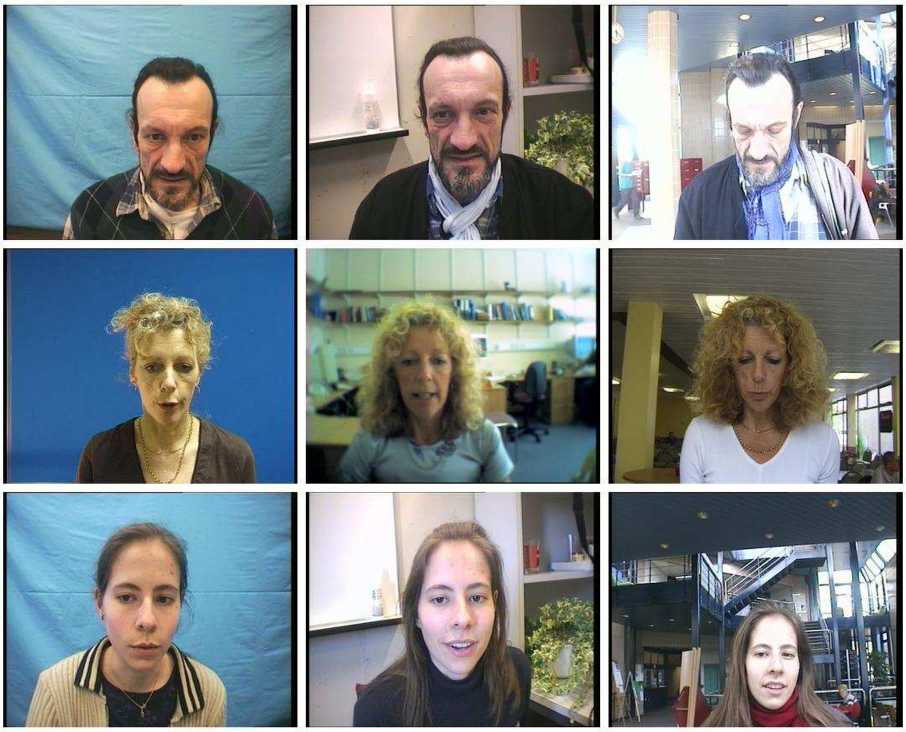


Figure 2.8: Examples of images from the BANCA database. The left column represents images from the controlled condition, the middle column corresponds to degraded condition and the right column corresponds to adverse condition

Chapter 3

Adapted Generative Models for Face Authentication

This chapter proposes a framework for face authentication using generative models. In particular, we propose to train the models with Maximum *a Posteriori* (MAP) [28] adaptation instead of the traditional Maximum Likelihood (ML). We also present an extensive comparison of three different types of generative models. The contributions presented in this chapter have been published in [15, 13, 14].

The previous chapter described the common structure of an AFR system and gave an overview of the solutions offered in the literature for each step of the processing. In this thesis we focus our attention on model based approaches and more precisely on generative models. This choice is motivated by the good performance obtained in previous works [24, 33, 59, 62]. Generative models methods include Gaussian Mixture Model (GMM) [15], One-dimensional Hidden Markov Model (1D-HMM) [77], Pseudo two-dimensional Hidden Markov Model (P2D-HMM) [24, 59].

In the approaches presented in [24, 59, 77, 80, 62], generative models are trained using the Maximum Likelihood (ML) criterion via the Expectation Maximization (EM) algorithm [20]. It is generally known that one of the drawbacks of training via this paradigm is that a lot of data is required to properly estimate model parameters; this can be a problem when there are only a few training images available. In an attempt to tackle this problem, Eickeler *et al.* [24] proposed to use a well trained generic (non-person specific) model as the starting point for ML training. While the results in [24] were promising, they were obtained on the rather easy Olivetti Research Ltd. (ORL) database [77]. Through experiments on the much harder BANCA database [3], we will show that even with the generic model as the starting point, ML training still produces poor models. We propose to replace ML training with Maximum *a Posteriori* (MAP) training [28] (also called *MAP adaptation* since this approach derive a client specific model from a generic model), which effectively can circumvent the lack of data problem. A first attempt to use MAP to train GMMs for face authentication was presented in [78]; here we go further since we show that this approach can also be successfully applied for more complex models such as 1D-HMM and P2D-HMM.

Furthermore, we demonstrate that the performance of the overall face authentication system can be *highly dependent* on the performance of the face localization (or detection) algorithm (i.e. the algorithm's ability to accurately locate a face, with no clipping or scaling problems). In other words, face recognition techniques which obtain good performance on manually located faces do not necessarily obtain good performance on automatically located faces. We make the claim that the face recognition technique *must be designed* from the ground up to handle imperfectly located faces.

Finally we show that complexity of a face recognition system is an important consideration in a practical implementation. By "complexity" we mean the number of parameters to store for each person as well as the time required to make an authentication. If a face model is to be stored on an electronic card (e.g. an access card), the size of the model becomes an important issue. Moreover, the time needed to authenticate a person should not be cumbersome, implying the need to use techniques which are computationally simple.

3.1 Generative Model Based Classifiers

While in most previous works in the literature, generative models were used for face identification, we propose here a framework to use this kind of model for AFA.

Let us denote the parameter set for client C as λ_C, and the parameter set describing a generic face (non-client specific) as $\lambda_{\overline{C}}$. Given a claim for client C's identity and a set of T feature vectors $X = \{\mathbf{x}_t\}_{t=1}^{T}$ supporting the claim (extracted from the given face), we find an opinion on the claim using:

$$\Lambda(X) = \log P(X|\lambda_C) - \log P(X|\lambda_{\overline{C}}) \tag{3.1}$$

where $P(X|\lambda_C)$ is the likelihood of the claim coming from the true claimant and $P(X|\lambda_{\overline{C}})$ is the likelihood of the claim coming from an impostor. The generic face model is also known as a *world model* and a *Universal Background Model* [53, 21]; it is typically trained with data from many people. The authentication decision is then reached as described in Section 2.1.4: given a threshold τ, the claim is accepted when $\Lambda(X) \geq \tau$ and rejected when $\Lambda(X) < \tau$.

In this chapter, we explore three different types of model that have been proposed for AFR in the literature. From the simplest to the most complex, these models are GMM, 1D-HMM and P2D-HMM.

3.1.1 Gaussian Mixture Model

In the GMM approach, the likelihood of a set of feature vectors is found with

$$P(X|\lambda) = \prod_{t=1}^{T} P(\mathbf{x}_t|\lambda) \tag{3.2}$$

where

$$P(\mathbf{x}|\lambda) = \sum_{k=1}^{N_G} m_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \tag{3.3}$$

$$\lambda = \{m_k, \mu_k, \Sigma_k\}_{k=1}^{N_G} \tag{3.4}$$

Here, $\mathcal{N}(\mathbf{x}|\mu, \Sigma)$ is a D-dimensional gaussian density function with mean μ and diagonal covariance matrix Σ. N_G is the number of gaussians and m_k is the weight for gaussian k (with constraints $\sum_{k=1}^{N_G} m_k = 1$ and $\forall\, k : m_k \geq 0$).

Each feature vector X describes a different part of the face (corresponding to a local approach described in Section 2.1.3). We note that the spatial relations between face parts are lost (see Equation 3.2). In other words, the position of each part does not matter in the likelihood estimation. This should lead to a robustness to imperfect localization of the face, however, in return, discriminatory information carried by spatial relations is lost. A possible way to restore a degree of spatial relations is proposed in the next chapter.

3.1.2 1D Hidden Markov Model

The one-dimensional HMM (1D-HMM) is a particular HMM topology where only self transitions or transitions to the next state are allowed. This type of HMM is also known as a top-bottom HMM [77] or left-right HMM in the context of speech recognition [70]. Here the face is represented as a sequence of overlapping *rectangular* blocks from top to bottom of the face (see Figure 3.1 for an example). The model is characterized by the following:

1. N, the number of states in the model; each state corresponds to a region of the face; $S = \{S_1, S_2, \ldots, S_N\}$ is the set of states. The state of the model at row t is given by $q_t \in S,\ 1 \leq t \leq T$, where T is the length of the observation sequence (number of rectangular blocks).

2. The state transition matrix $A = \{a_{ij}\}$. The topology of the 1D-HMM allows only self transitions or transitions to the next state:

$$a_{ij} = \begin{cases} P(q_t = S_j | q_{t-1} = S_i) & \text{for } j = i,\ j = i+1 \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

3. The state probability distribution $B = \{b_j(\mathbf{x}_t)\}$, where

$$b_j(\mathbf{x}_t) = P(\mathbf{x}_t | q_t = S_j) \tag{3.6}$$

   The features are expected to follow a continuous distribution and are modeled with mixtures of gaussians.

In compact notation, the parameter set of the 1D-HMM is:

$$\lambda = (A, B) \tag{3.7}$$

If we let Q to be a state sequence $q_1, q_2, \cdots, q_T$, then the likelihood of an observation sequence X is:

$$P(X|\lambda) = \sum_{\forall\ Q} P(X, Q|\lambda) \tag{3.8}$$

$$= \sum_{\forall\ Q} \prod_{t=1}^{T} b_{q_t}(\mathbf{x}_t) \prod_{t=2}^{T} a_{q_{t-1}, q_t} \tag{3.9}$$

The calculation of this likelihood according to the direct definition in Equation (3.9) involves an exponential number of computations; in practice the Forward-Backward procedure is used [70]; it is mathematically equivalent, but considerably more efficient.

Compared to the GMM approach described in Section 3.1.1, the spatial constraints are much more strict, mainly due to the rigid preservation of horizontal spatial relations (e.g. distance between the eyes). The vertical constraints are more relaxed, though they still enforce the top-to-bottom segmentation (e.g. the eyes have to be above the mouth). The relaxation of constraints allows for a degree of vertical translation and some vertical stretching (caused, for example, by an imperfect face localization).

3.1.3 Pseudo-2D HMM

Emission probabilities of 1D HMMs are typically represented using mixtures of gaussians. For the case of P2D-HMM, the emission probabilities of the HMM (now referred to as the “main HMM”) are estimated through a secondary HMM (referred to as an “embedded HMM”). The states of the embedded HMMs are in turn modeled by a mixture of gaussians. This approach was used for the face identification task in [24, 77] and the training process is described in detail in [60]. As shown in Figure 3.2, we chose to perform the vertical segmentation of the face image by the main HMM and horizontal segmentation by embedded HMMs. We made this choice because the main decomposition of the face is instinctively from top to the bottom (forehead, eyes, nose, mouth). Note that the opposite choice has been made in [24, 77].

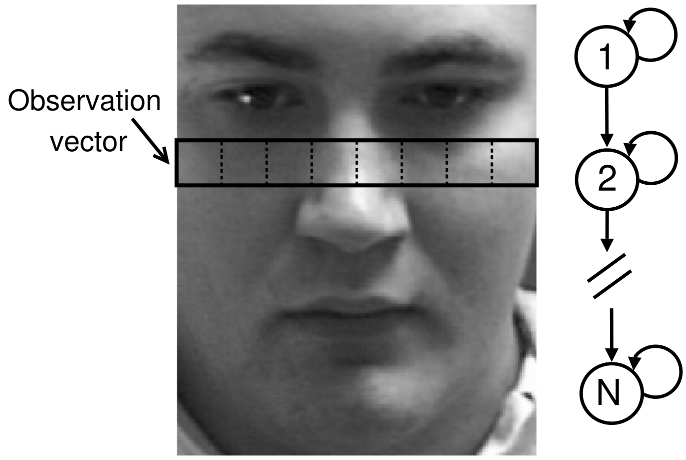


Figure 3.1: Sampling window and 1D-HMM topology.

The degree of spatial constraints present in the P2D-HMM approach can be thought of as being somewhere in between the GMM and the 1D-HMM approaches. While the GMM approach has no spatial constraints and the 1D-HMM has rigid horizontal constraints, the P2D-HMM approach has relaxed constraints in both directions. However, the constraints still enforce the left-to-right segmentation of the embedded HMMs (e.g. the left eye has to be before the right eye), and top-to-bottom segmentation (e.g. like in the 1D-HMM approach, the eyes have to be above the mouth). The relaxed constraints allow for a degree of both vertical and horizontal translations, as well as some vertical and horizontal stretching of the face.

3.2 Maximum *a Posteriori* (MAP) Training

The traditional approach to train generative models is to maximize the likelihood via EM algorithm [20]. While this approach is appropriate to train the world model since we can have a large number of training data, it can be a problem to train the client models because only one or a few images are usually available. We propose instead the use of MAP training which includes prior information and thus should train better models when only a few training images are available.

We compare three different ways to train the client models:

1. Traditional ML training, where k-means initialization is used [20, 22].

2. ML training with a generic (non-client specific) model as the starting point (as in [24]); data from many people is used the find the parameters of the generic model via traditional ML training; this is the same generic model used for calculating $P(X|\lambda_{\overline{C}})$ in Equation (3.1) for all generative approaches.

3. MAP training [28]; here a generic model is used as in point (2) above, but instead of using it merely as a starting point, the model is *adapted* using client data. Given a set of training vectors, S, the probability density function (pdf) $P(S|\lambda)$ and the prior pdf of λ, $P(\lambda)$, the MAP estimate of model parameters, $\lambda_{\texttt{MAP}}$, is defined as:

$$\lambda_{\texttt{MAP}} = \arg\max_{\lambda} P(\lambda|S) \tag{3.10}$$

$$= \arg\max_{\lambda} P(S|\lambda)P(\lambda) \tag{3.11}$$

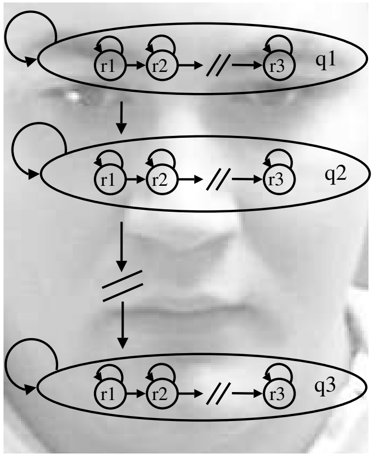


Figure 3.2: P2D-HMM: the emission distributions of the vertical HMM are estimated by horizontal HMMs. q_i represent the states of the main HMM and r_j represent the embedded HMMs states.

Assuming λ to be uniform is equivalent to having a non-informative $P(\lambda)$, reducing the solution of $\lambda_{\texttt{MAP}}$ to the standard ML solution. Thus, the difference between ML and MAP training is in the definition of the prior distribution for the model parameters to be estimated.

An implementation of MAP training for client model adaptation consists of using a global parameter to tune the relative importance of the prior. In this case, for a GMM, the equations for adaptation of the parameters are [28, 53, 21]:

$$\hat{w}_k = \left[\alpha w_k + (1-\alpha) \sum_{t=1}^{T} P(k|\mathbf{x}_t) \right] \gamma \tag{3.12}$$

$$\hat{\mu}_k = \alpha \mu_k + (1-\alpha) \frac{\sum_{t=1}^{T} P(k|\mathbf{x}_t)\mathbf{x}_t}{\sum_{t=1}^{T} P(k|\mathbf{x}_t)} \tag{3.13}$$

$$\hat{\Sigma}_k = \alpha \left(\Sigma_k + \mu_k \mu_k^{\prime} \right) + (1-\alpha) \frac{\sum_{t=1}^{T} P(k|\mathbf{x}_t)\mathbf{x}_t \mathbf{x}_t^{\prime}}{\sum_{t=1}^{T} P(k|\mathbf{x}_t)} - \hat{\mu}_k \hat{\mu}_k^{\prime} \tag{3.14}$$

where $\hat{w}_k$, $\hat{\mu}_k$ and $\hat{\Sigma}_k$ are respectively the new weight, mean and covariance matrix of the k-th gaussian, w_k, μ_k and Σ_k are the corresponding parameters in the generic model, $P(k|\mathbf{x}_t)$ is the posterior probability of the k-th gaussian (from the client model from the previous iteration), $\alpha \in [0, 1]$ is the adaptation factor chosen empirically on a separate validation set and finally γ is computed over all adapted weights to ensure they sum to unity. Each $\hat{\Sigma}_k$ is forced to be diagonal by setting the off-diagonal elements to zero. Note that in Equation (3.12) the new mean is simply a weighted sum of the prior mean and new statistics; $(1-\alpha)$ can hence be interpreted as the amount of faith we have in the new statistics.

The above formulation of MAP training makes the assumption of independence between the parameters of the individual mixture components and the set of the mixture weights; furthermore we consider that we can model the prior knowledge about the parameter vector of mixture weights with a Dirichlet density and the prior knowledge about the means and variances with normal-Wishart densities [28].

The adaptation procedure is iterative, thus an initial client model is required; this is accomplished by copying the generic model. It has been observed that it is sometimes preferable to adapt only the means of the gaussians [21]; we will empirically show that this is also valid for our experiments in Section 3.4. When only the means are adapted the other parameters are copied from the generic model.

For the case of the 1D-HMM, MAP adaptation of the means can be written as follows [c.f. Equation (3.13)]:

$$\hat{\mu}_{k,i} = \alpha \mu_{k,i} + (1-\alpha) \frac{\sum_{t=1}^{T} P(q_t = i|\mathbf{x}_t) P(k|\mathbf{x}_t)\mathbf{x}_t}{\sum_{t=1}^{T} P(q_t = i|\mathbf{x}_t) P(k|\mathbf{x}_t)} \tag{3.15}$$

where $P(q_t = i|\mathbf{x}_t)$ is the posterior probability of the state i at row t and $P(k|\mathbf{x}_t)$ is the posterior probability of its k-th gaussian.

For the case of the P2D-HMM, let us denote the sequence of T observation vectors representing the consecutive horizontal strips of an image as $X = \{\mathbf{x}_t\}_{t=1}^{T}$. Each strip can itself be represented as a sequence of B observation vectors $\mathbf{x}_t = \{\mathbf{x}_{t,b}\}_{b=1}^{B}$ representing the consecutive blocks composing the strip. The corresponding equation for MAP adaptation of the means [c.f. Eqns. (3.13) and (3.15)] is:

$$\hat{\mu}_{k,i,j} = \alpha \mu_{k,i,j}^{w} + (1-\alpha)\hat{\mu}_{k,i,j}^{\mathtt{ML}} \tag{3.16}$$

with:

$$\hat{\mu}_{k,i,j}^{\mathtt{ML}} = \frac{\sum_{t=1}^{T} P(q_t = i|\mathbf{x}_t) \sum_{b=1}^{B} P(r_{t,b}^{i} = j|\mathbf{x}_{t,b}) P(m_{t,b}^{i,j} = k|\mathbf{x}_{t,b})\mathbf{x}_{t,b}}{\sum_{t=1}^{T} P(q_t = i|\mathbf{x}_t) \sum_{b=1}^{B} P(r_{t,b}^{i} = j|\mathbf{x}_{t,b}) P(m_{t,b}^{i,j} = k|\mathbf{x}_{t,b})} \tag{3.17}$$

where $P(q_t = i|\mathbf{x}_t)$ is the posterior probability of the state i of the main HMM, $P(r_{t,b}^{i} = j|\mathbf{x}_{t,b})$ is the posterior probability of the state j of its embedded HMM and $P(m_{t,b}^{i,j} = k|\mathbf{x}_{t,b})$ is the posterior probability of its k-th gaussian.

3.3 Preliminary Steps

Before the classification itself, preliminary steps are necessary for an AFA system (see Section 2.1). Face localization, pre-processing and feature extraction steps are briefly presented in this section.

3.3.1 Face Localization

Face recognition results in the literature are usually presented assuming manual face localization (e.g. see [24, 59, 60, 77]); in only relatively few publications performance evaluation is found while using automatic face localization (e.g. [15, 76]). While assuming manual (i.e. perfect) localization makes the results independent of the quality of the face localization system, they are biased when compared to a real life system, where the face needs to be automatically located. There is no guarantee that the automatic face localization system will provide a correctly located face (i.e. the face may be translated and/or at an incorrect scale). In this thesis we present results for both manually and imperfectly located faces.

For "manual face localization" experiments, we use the manually annotated eye center positions. For "automatic face localization" experiments, we use the face detector proposed by Fröba and Ermst in [26]. The detector employs local features based on the "*Modified Census Transform*", which represent each location of the image by a binary pattern computed from a 3×3 pixel neighborhood. Face detection is carried out by analyzing all possible windows in the given image at different scales; each window is classified as either containing a face or the background. The classification is performed by a cascade classifier similar to the approach proposed by Viola and Jones [92]; training of the classifier is accomplished using a version of the boosting algorithm [25]. In our experiments the eye positions are inferred from the position and scale of the detected face. Note that this assumes that at most only one face is present in each image.

If all the windows were classified as containing the background, we consider that the given image does not contain a face and we perform the authentication using, if available, other images supporting the claim. If all given images are deemed not to contain a face, the claim is considered to have come from an impostor.

3.3.2 Pre-processing and Feature Extraction

Based on given eye positions, a gray-scale 80×64 (rows$\times$columns) face window is cropped out of each valid image (i.e. an image which is deemed to contain a face). When using manually found eye positions, each face window contains the face area from the eyebrows to the mouth; moreover, the location of the eyes is the same for each face window (via geometric normalization). Fig. 3.1 shows an example face window.

Histogram equalization is used to normalize the face images photometrically. We then extract *DCTmod2* features from each image face [80]. We have found this combination of histogram equalization and feature extraction to provide good results in preliminary experiments. The feature extraction process is summarized as follows. The face window is analyzed on a block by block basis; each block is $N_P \times N_P$ (here we use N_P=8) and overlaps neighboring blocks by a configurable amount of pixels[1]. Each block is decomposed in terms of 2D Discrete Cosine Transform (DCT) basis functions [30]. A feature vector for a block located at row a and column b is then constructed as:

$$\mathbf{x}_{(a,b)} = \left[\; \Delta^h c_0 \; \Delta^v c_0 \; \Delta^h c_1 \; \Delta^v c_1 \; \Delta^h c_2 \; \Delta^v c_2 \;\; c_3 \; c_4 \ldots c_{M-1} \; \right]'$$

where c_n represents the n-th DCT coefficient, while $\Delta^h c_n$ and $\Delta^v c_n$ represent the horizontal and vertical delta coefficients respectively; the deltas are computed using DCT coefficients extracted from neighboring blocks. Compared to traditional DCT feature extraction [24, 59], the first three DCT coefficients are replaced by their respective deltas in order to reduce the effects of illumination direction changes, without losing discriminative information. In this study we use M=15 (based on [80]), resulting in an 18 dimensional feature vector for each block. The degree of overlap has three main effects:

1. As the delta coefficients are computed from neighboring blocks, the larger the overlap between the blocks, the smaller the spatial area used to derive each feature vector.

2. With a large overlap, the DCT coefficients from a set of (horizontally or vertically) consecutive blocks will not vary abruptly.

3. When using a large overlap, the parts of each face are in effect "sampled" at various degrees of translations, resulting in models which should be robust to minor translations of the faces. This is in *addition* to the translation robustness provided by the GMM classifier, where the location of each block has little influence. By itself, GMM's built-in robustness only works when the size of the translation is equivalent to an integral multiple of the block size.

4. A large overlap increases dependence between consecutive blocks.

3.4 Experiments and Discussion

For each client model, the training set was composed of five images extracted from the same video sequence. We artificially increased this to ten images by mirroring each original image. The generic model was trained with 571 face images (extended to 1142 by mirroring) from the Spanish corpus of BANCA (containing faces different from the English and French corpora), thus making the generic model independent of the subjects present in the client database. DCTmod2 features were extracted using either a four or a seven pixel overlap; experiments on the validation set showed that an overlap of four pixels is better for the GMM approaches while an overlap of seven pixels is preferred by the P2D-HMM approach. For the 1D-HMM approach, a seven pixel overlap was also used, but feature vectors from the same row of blocks were concatenated to form a large observation vector. To keep the dimensionality of the resultant vector reasonable, we chose to

[1] A similar overlapping approach is used in processing of speech signals [67, 21, 84].

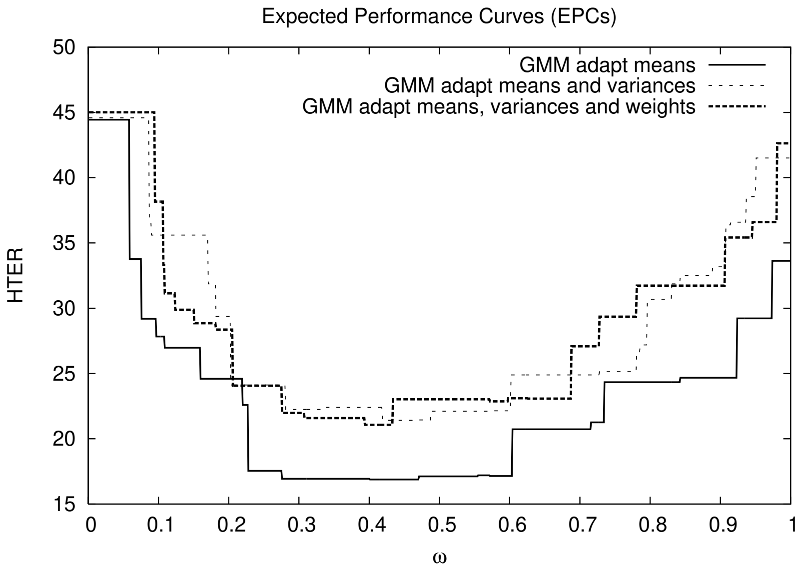


Figure 3.3: EPC performance of GMM based system trained via MAP adaptation. Three configurations of MAP adaptation are shown.

concatenate vectors from every eighth block (thus eliminating horizontally overlapped blocks). This resulted in 126 dimensional feature vectors for each rectangular block.

In order to optimize each model, we used the validation set to select the size of the model (e.g. number of states and gaussians) as well as other hyper-parameters, such as the adaptation coefficient α, and the decision threshold τ; the parameters were chosen to minimize the EER. The final performance of each model was then found on the test set.

It has been observed that in applications such as speaker authentication [53, 21], MAP based training obtains best performance when only the means are adapted (rather than also adapting the covariance matrices and weights). Fig. 3.3 shows EPCs for the GMM based system for three cases: (i) all parameters are adapted, (ii) means and covariance matrices are adapted, (iii) only means are adapted. Database protocol P was employed in this evaluation. As adapting only the means provides the best performance, we have elected to use this strategy for both GMM and HMM approaches. Hence for the rest of this thesis, the MAP training strategy will refer to the adaptation of the means only.

Models trained using the traditional ML criterion have a *ML* suffix; for ML training initialized with a generic model, the suffix is *init*; for MAP training, the suffix is *MAP*.

Table 3.1 shows the optimum parameters empirically determined, for all approaches. The parameters considered are the number of states and gaussians per state for the HMM approaches, the total number of gaussians, the number of iteration for the background model and for the client models, the adaptation factor α, the variance floor factor[2] and the decision threshold. It can be observed that MAP training generally allows the total number of gaussians to be higher (thus modeling the faces more accurately), when compared to the two ML based training paradigms. The P2D-HMM approach utilizes the largest number of gaussians.

For comparison purposes, we also evaluate the performance of a PCA based system, which in effect has rigid constraints between face parts. The classifier used for the PCA system is somewhat similar to the local

[2] For each dimension d, the variance of a gaussian can not be lower than the variance floor σ_d^f. The variance floor is computed to be proportional to the variance σ_d of the training data along the dimension d: $\sigma_d^f = \gamma.\sigma_d$; where γ is the variance floor factor.

Table 3.1: Optimum parameters for systems based on GMM, 1D-HMM and P2D-HMM. *ML*: client models trained using traditional ML criterion; *init*: client models trained using ML initialized with a generic model; *MAP*: client models trained using MAP.

System	Number of states		Gaussians	Total
	main HMM	embedded HMM	per state	gaussians
GMM *ML*	-	-	-	256
GMM *init*	-	-	-	512
GMM *MAP*	-	-	-	512
1D-HMM *ML*	16	-	1	16
1D-HMM *init*	32	-	1	32
1D-HMM *MAP*	32	-	1	32
P2D-HMM *ML*	8	16	4	512
P2D-HMM *init*	16	16	2	512
P2D-HMM *MAP*	16	4	64	4096

System	Number of iterations		Adaptation	Variance floor	Decision threshold
	Background model	Client models	factor (α)	factor	(Protocol P)
GMM *ML*	6	2	-	0.15	-2.44
GMM *init*	6	2	-	0.25	-2.14
GMM *MAP*	6	2	0.45	0.35	-0.06
1D-HMM *ML*	4	4	-	0.45	-1324
1D-HMM *init*	4	4	-	0.45	-1402
1D-HMM *MAP*	4	4	0.45	0.40	34
P2D-HMM *ML*	2	8	-	0.22	-11524
P2D-HMM *init*	2	8	-	0.20	-11643
P2D-HMM *MAP*	2	8	0.40	0.40	-512

feature GMM approach. The main difference is that only two gaussians are utilized: one for the client and one to represent the generic model. Due to the very small amount of client specific training data, and since PCA feature extraction results in one feature vector per face, each client model inherits the covariance matrix from the generic model and the mean of each client model is the mean of the training vectors for that client. A similar system has been used in [79, 82]. Feature vectors with 160 dimensions were found to provide optimal performance on the validation set.

In Section 3.4.1 we present the results for manual face localization, while Section 3.4.2 contains results for imperfect and automatic face localization. In Section 3.4.3 we study the effects of varying the number of training images and finally in Section 3.4.4 we compare the complexity of the local feature approaches.

Note that the result tables presented in Sections 3.4.1 and 3.4.2 also contain performance figures for the two best systems reported in [76]. The first system is based on combination of Linear Discriminant Analysis and Normalized Correlation (LDA/NC), while the second system is based on a Support Vector Machine (SVM) classifier. Like the PCA based system, these LDA/NC and SVM systems are holistic in nature. It must be noted that in [76], only the English corpus was used and a different automatic face localization system was employed. As such the results from [76] are not directly comparable, but are included as an example of the performance degradation that occurs when automatic face localization is utilized (compared to using manually located faces).

3.4.1 Manual Face Localization

Table 3.3(a) shows the results in terms of HTER for manual face localization; Fig. 3.4 shows the corresponding EPCs. When comparing the different training strategies, MAP training provides a clear performance advantage in almost all the cases. The only exception is the 1D-HMM approach for which all training approaches obtain similar performance. ML training with initialization by a generic model generally does not result in better models compared to traditional ML training (where k-means initialization is used).

When comparing performance across different models, it can be seen that the two HMM approaches (1D and P2D-HMM) obtain considerably better performance than the two GMM based approaches.

The 1D-HMM outperforms the P2D-HMM approach when ML training is utilized; this can be explained by the inherently much larger number of parameters used in P2D-HMM (hence requiring more training data). However, when MAP training is used, the lack of data problem is effectively circumvented, resulting in the P2D-HMM approach obtaining the best overall performance. We also perform a statistical significance test to assess if the P2D-HMM significantly better than the other models. Recently, a statistical significance test was proposed for person authentication [7]. Using this test on P protocol results (with manual face localization) for models trained with MAP, we observe that the P2D-HMM system is statistically significantly better than the GMM system with a confidence of more than 96%. However, the P2D-HMM is not significantly better than the 1D-HMM with a confidence of only 53%.

3.4.2 Imperfect and Automatic Localization

Prior to using the automatic face locator described in Section 3.3.1, we first study how each system is affected by an increasing amount of error in the position of the eyes. For this set of experiments we used exactly the same models as in Section 3.4.1 (i.e. trained with manually localized faces). The eye positions were artificially perturbed using:

$$eye_x = eye_x^{\mathtt{gt}} + \xi \tag{3.18}$$
$$eye_y = eye_y^{\mathtt{gt}} + \xi \tag{3.19}$$

where $eye_x^{\mathtt{gt}}$ and $eye_y^{\mathtt{gt}}$ are the ground-truth (original) co-ordinates for an eye. ξ is a random variable and follows a normal distribution such that $\xi \sim \mathcal{N}(0, \sigma^2)$, where $\sigma^2 = V \cdot D_{\mathtt{eyes}}$, with $D_{\mathtt{eyes}}$ being the Euclidean distance between the two eyes. $V \in [0, 1]$ and can be interpreted as the amount of introduced error. When $V = 1$, the largest translation (in one axis) will tend to be about half of the distance between the eyes.

Results in Figure 3.5 show that GMM and P2D-HMM based systems are quite robust to imperfect face localization. In contrast, the PCA and 1D-HMM systems are significantly more sensitive, with their discrimination performance rapidly decreasing as the error is increased. We attribute this performance degradation to the more constrained spatial relation between face parts; while the 1D-HMM system allows for some vertical displacement, it has rigid constraints in the horizontal direction; in the PCA based system the relations are rigidly preserved along both axes.

Table 3.3(b) shows that the observations from perturbation experiments are confirmed when the automatic face locator is utilized. The PCA system is the most affected, followed by the 1D-HMM. In Table 3.3(a) it was shown that when using MAP based training and manual face localization, the 1D-HMM approach outperforms the two GMM based systems; however, for automatic face localization, the GMM approach generally outperforms the 1D-HMM system. The P2D-HMM system again obtains the best overall performance, with minimal degradation in discrimination ability when compared to manually located faces. Furthermore, using the statistical significance test presented in [7], we observe that the P2D-HMM trained with MAP is statistically significantly better than all the other sytems presented in this chapter with a confidence of more than 99.5%.

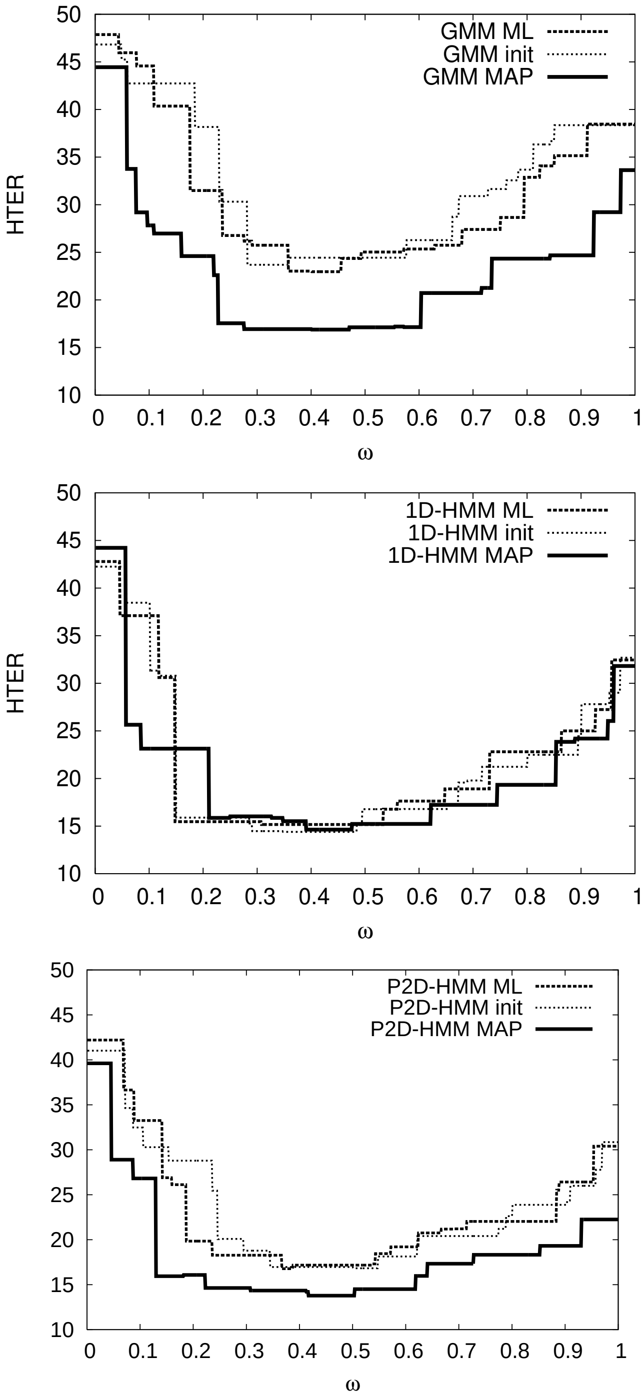


Figure 3.4: EPCs for manual face localization.

Table 3.2: HTER performance for (a) **manual face localization**, and (b) **automatic face localization**, using GMM, 1D-HMM and P2D-HMM. *ML*: client models trained using traditional ML criterion; *init*: client models trained using ML initialized with a generic model; *MAP*: client models trained using MAP. The asterisk indicates the best result for a protocol, while boldface indicates the best result within a model type and protocol.

System	Protocol			
	Mc	Ud	Ua	P
PCA	9.5	20.9	20.8	18.4
LDA/NC (from [76])	4.9	16.0	20.2	14.8
SVM (from [76])	5.4	25.4	30.1	20.3
GMM *ML*	12.9	28.9	26.0	22.9
GMM *init*	12.8	29.7	28.3	23.8
GMM *MAP*	**8.9**	**17.3**	**20.9**	**17.0**
1D-HMM *ML*	9.1	17.8	17.1	15.9
1D-HMM *init*	9.1	**15.6**	17.4	**14.7**
1D-HMM *MAP*	**6.9**	16.3	**17.0**	**14.7**
P2D-HMM *ML*	9.0	19.0	18.0	17.5
P2D-HMM *init*	8.6	16.5	19.2	17.0
P2D-HMM *MAP*	∗ **4.6**	∗ **15.3**	∗ **13.1**	∗ **13.5**

(a) HTER performance for **manual face localization**

System	Protocol			
	Mc	Ud	Ua	P
PCA	22.4	29.7	33.7	29.0
LDA/NC (from [76])	22.6	25.4	27.1	25.2
SVM (from [76])	19.7	30.4	33.2	27.8
GMM *ML*	16.7	33.3	33.3	27.7
GMM *init*	19.8	35.0	35.1	29.7
GMM *MAP*	**9.5**	**21.0**	**24.8**	**19.5**
1D-HMM *ML*	21.0	28.8	29.5	27.0
1D-HMM *init*	21.3	30.1	31.4	28.1
1D-HMM *MAP*	**13.8**	**25.9**	**23.4**	**21.7**
P2D-HMM *ML*	12.1	25.2	26.9	22.3
P2D-HMM *init*	13.5	24.6	26.5	22.5
P2D-HMM *MAP*	∗ **6.5**	∗ **15.9**	∗ **14.7**	∗ **14.7**

(b) HTER performance for **automatic face localization**

3.4.3 Number of Training Images

The relatively small number of face images available to train each client model can be a limiting factor in obtaining precise face models. In some applications, such as surveillance, there may be only one reference image (e.g. a passport photograph). In the experiments reported in Sections 3.4.1 and 3.4.2, five images were available for each client; the number of images was artificially increased to ten by mirroring each original image. In this section we evaluate the effects of decreasing the number of original training images.

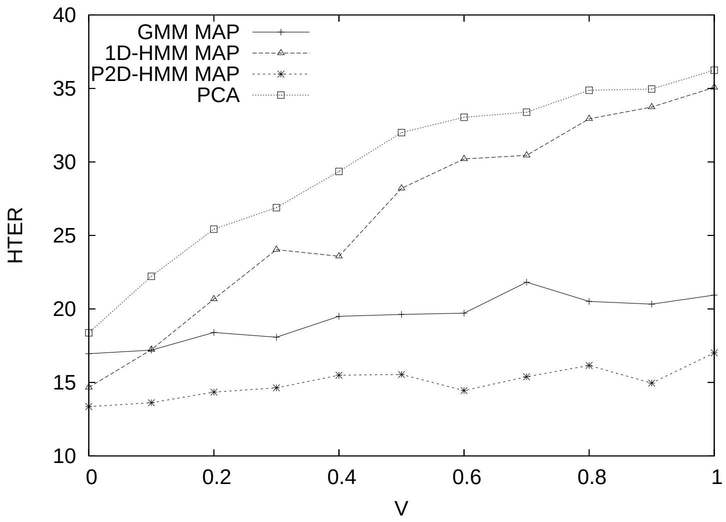


Figure 3.5: Performance for an increasing amount of error in eye locations.

Fig. 3.6 shows the performance as a function of the number of original images (i.e. mirrored versions were also utilized). Database protocol P was employed in this evaluation. Irrespective of the training strategy and model, the greatest improvement generally occurs when two training images are utilized instead of one; moreover, discrimination performance tends to saturate at three images. The exception is the MAP trained P2D-HMM approach, where there is no clear benefit in utilizing more than one image. Overall, MAP training is the least sensitive to the number of training images. Lastly, the GMM system benefits the most from an increase in the number of training images.

3.4.4 Complexity of Models

Apart from the performance, the complexity of a given model is also an important consideration; here, by "complexity" we mean the number of parameters to store for each client as well as the time required for training and authentication. If we wish to store each model on an electronic card (e.g. a credit card), the size of the model becomes an important issue. We are specifically interested in the number of *client specific* parameters, meaning that we count only parameters which are different between the clients.

Table 3.3 shows the complexity of each local feature model used in our experiments (using hyper-parameters tuned for optimal discrimination performance, such as the number of gaussians); specifically, we show the number of client specific parameters, the time taken to train the world model, the client model training time, and the time required to authenticate one claim (comprised of five images). The experiments were done on a Pentium IV 3 GHz running Red Hat Linux 7.3. The times include pre-processing time; the values in brackets indicate the time for authentication or training excluding steps such as face localization, normalization and feature extraction. While the implementation of GMM and HMM based systems was not specifically optimized in terms of speed, we believe the times presented are indicative.

The number of client specific parameters for GMM based approaches is the sum of the parameters for the means, covariance matrices (both dependent on the dimensionality of feature vectors) and weights; for the HMM based approaches transition probabilities are also taken into account. When MAP training is used, only the means need to be counted, since the other parameters are shared by all clients; the shared parameters can be stored only once in the system for all clients (e.g. there is no need to store them in each client's electronic card). This is in contrast to ML based training, where there are no parameters shared between client models. For

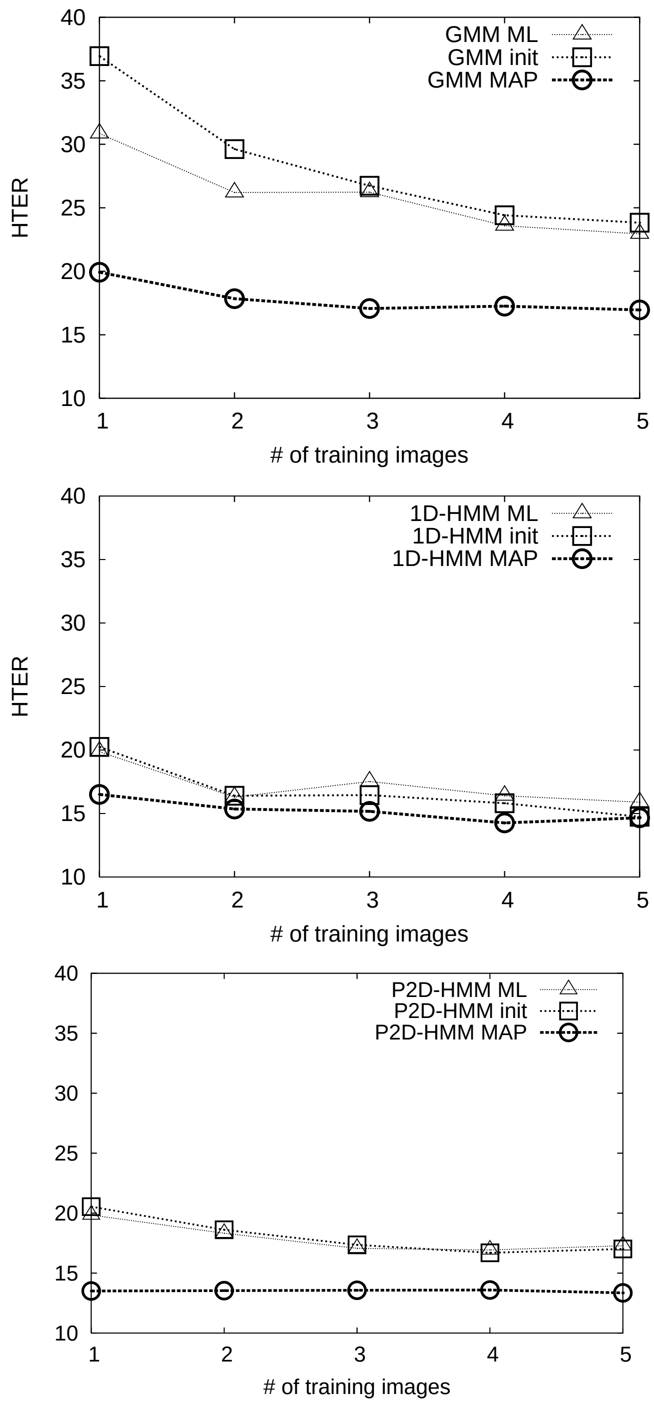


Figure 3.6: Performance as a function of the number of original training images.

Table 3.3: **Complexity of the models**. Times are given in terms of seconds. Values in brackets exclude pre-processing time (e.g. face localization, normalization, feature extraction, etc).

Model type	GMM			1D-HMM			P2D-HMM		
Training type	ML	init	MAP	ML	init	MAP	ML	init	MAP
number of client specific parameters	9,472	18,944	9,216	4,063	8,127	4,032	19,207	19,471	73,728
world model training time	295s (163s)	470s (337s)	470s (355s)	181s (3s)	184s (6s)	192s (14s)	2873s (2695s)	1873s (1695s)	7967s (7789s)
client model training time	2s (0.5s)	2s (1s)	2s (1s)	2s (0.5s)	2s (0.5s)	3s (2s)	65s (64s)	88s (87s)	251s (250s)
time for authentication of one claim (5 images)	0.95s (0.07s)	1.10s (0.22s)	1.12s (0.24s)	1.22s (0.13s)	1.25s (0.16s)	1.31s (0.22s)	5.74s (4.65s)	7.25s (6.16s)	19.89s (18.80s)

example, when using the GMM approach and an equal number of gaussians for both ML and MAP training, the number of client specific parameters for MAP trained models is about half of the number required for ML based training.

Training of the generic model can be done off-line and hence the time required is not of great importance; however, the time taken to train each client model as well as the time for one authentication are quite important. There shouldn't be a long delay between a user enrolling in the system and being able to use the system; most importantly, the authentication time should not be cumbersome, in order to aid the adoption of the authentication system. The GMM and 1D-HMM approaches have short training and authentication times of around three and one seconds, respectively. We note that for these three approaches, the pre-processing steps considerably penalize the speed of the authentication.

When using MAP trained models, the P2D-HMM approach has a considerably higher training and authentication time, at approximately 4 minutes for training each client model and 20 seconds for an authentication. With current computing resources, this authentication time can be considered as being too long for practical deployment purposes. When using ML trained models, the training and authentication time is significantly reduced, which is partly due to the total number of gaussians being smaller. However, ML trained models obtain considerably worse discrimination performance.

3.5 Conclusions

In most of the previous literature related to these models, the experiments were performed with controlled images (perfect face localization, controlled lighting, background, pose, expression, etc.); however, for most secure authentication applications, the system has to be robust to more challenging conditions.

In this chapter we evaluated the performance, robustness and complexity of GMM and HMM based approaches, using both perfect and automatic face localization, on the relatively difficult BANCA database. We evaluated different training techniques for both GMM and HMM based systems; we showed that the traditionally used Maximum Likelihood (ML) training approach has problems estimating robust model parameters when there is only a few training images available; we proposed to tackle this problem through the use of Maximum *a Posteriori* (MAP) training, where the lack of data problem can be effectively circumvented. We showed that models estimated with MAP are significantly more robust and are able to generalize to adverse conditions present in the BANCA database.

A second finding of this chapter is that systems that utilize rigid spatial constraints between face parts (such as PCA and 1D-HMM based systems), are easily affected by face localization errors, which are caused by an automatic face locator. In contrast, systems which have relaxed constraints (such as GMM and P2D-HMM based), are quite robust.

While the 1D-HMM based approach achieves promising performance for manually (i.e. perfectly) located faces and outperforms the GMM approach, for automatically located faces its performance degrades considerably and is worse than the GMM approach. More generally, we claim that good performance on manually located faces does not necessarily reflect good performance in real life conditions, where an automatic localization system must be used. As automatic localization cannot guarantee perfect face localization, this indicates that any new technique must be designed from the ground up to handle imperfectly located faces.

We also showed that while the P2D-HMM approach has overall the best performance, it requires relatively long times for training and authentication. For the GMM and the 1D-HMM based approaches, the preprocessing severely penalizes the speed of the authentication; efforts should thus be made on this part to speed up the overall system.

Chapter 4

Toward Fast and Robust Face Authentication

This chapter proposes two novel approaches performing fast and robust Face Authentication. The first proposition is an extension of the GMM based system presented in the previous chapter. This proposed extension of the GMM approach has been published in [13, 14]. The second approach is based on an alternative 1D-HMM structure with the capacity to deal with local features. We also show that traditional 1D-HMM as well as the proposed 1D-HMM can be seen as particular cases of P2D-HMM. This new 1D-HMM approach is the theme of a technical report [11].

In the previous chapter, generative models trained with MAP adaptation have shown promising performances. However, none of them combines robustness and low computational cost. The P2D-HMM obtains the overall best performance; however, it is computationally intensive, making it inappropriate for most of applications on current hardware. For 1D-HMM based approaches, an observation sequence represents consecutive horizontal strips and conserves rigid horizontal spatial constraints. As a consequence, typical 1D-HMM approaches are sensitive to imperfectly localized faces (due to a non-perfect face detection) and face deformation (due to different face expressions). GMMs are less complex and faster than HMMs with the cost of lower accuracy. In this chapter we investigate two different approaches that present good performance / robustness / complexity Trade-Offs.

4.1 Embedding Positional Information for GMM approach

4.1.1 Proposed Feature Vectors

In the GMM approach presented is the previous chapter, the spatial relation is effectively lost, as each block is treated independently (see Equation 3.2), resulting in good robustness to imperfectly located faces [15] and to out-of-plane rotations [79].

As the spatial relations can carry discriminatory information, we propose to increase the performance of the GMM approach (without sacrificing its simplicity) by restoring a degree of spatial relations via embedding positional information into each feature vector. Doing so should place a weak constraint on the areas that each gaussian in the GMM can model, thus making a face model more specific. By working in the feature domain, the relative low-complexity advantage of the GMM approach is retained. Formally, an extended feature vector

Table 4.1: HTER performance of GMM with extended features for **manual face localization**

Feature Vectors	Models	Protocol				Time for authentication of one claim (5 images)
		Mc	Ud	Ua	P	
Blocks DCTmod2	GMM	8.9	17.3	20.9	17.0	1.1s
Blocks DCTmod2 extended	**GMM**	**8.5**	**17.5**	**20.8**	**16.4**	**1.3s**

Table 4.2: HTER performance of GMM with extended features for **automatic face localization**

Feature Vectors	Models	Protocol			
		Mc	Ud	Ua	P
Blocks DCTmod2	GMM	9.5	21.0	24.8	19.5
Blocks DCTmod2 extended	**GMM**	**8.5**	**18.4**	**22.5**	**19.1**

for position (a, b) is obtained with:

$$\mathbf{x}^{\texttt{extended}}_{(a,b)} = \begin{bmatrix} \mathbf{x}^{\texttt{original}}_{(a,b)} \\ a \\ b \end{bmatrix}$$

where $\mathbf{x}^{\texttt{original}}_{(a,b)}$ is the original feature vector for position (a, b).

We note that in [94] the author proposes a similar approach for speech recognition which include the temporal position, called "time index", in the feature vector.

4.1.2 Results

Table 4.1 shows the results in terms of HTER for manual face localization (we use the manually annotated eye center positions) and the time needed to perform an authentication. We remark that the performances of the system using the extended features (including positional information) generally outperform the baseline GMM system. However, we note that this improvement is not statistically significant (using statistical significant test proposed in [7]). Moreover, this improvement is with the cost of a slightly higher computation time, however this system is still fifteen times faster than the P2D-HMM system presented in previous section.

The performances for experiments using automatic face localization are reported in Table 4.2. Similarly to the previous chapter, for these experiments, we use the face detector proposed by Fröba and Ernst in [26]. These results demonstrate that feature vectors with embedded positional information increases the performance of the GMM approach, with no loss of robustness to errors in face localization. This indicates that spatial relations between face parts carry discriminative information.

4.2 Local Features based 1D-HMM

The second proposed approach is an alternative 1D-HMM structure which deals with observation vectors representing a block of the image instead of a whole strip in traditional 1D-HMM approaches. The blocks of a same line are then treated independently with no spatial constraints, making the model robust to misalignment.

4.2.1 Proposed Model

The observation sequence for traditional 1D-HMM is typically composed of vectors that represent consecutive horizontal strips (see Figure 3.1 for an example) . As a consequence, 1D-HMM system has rigid horizontal constraints and does not allow the system to deal with face deformations, when the expression change and imperfectly aligned faces, when localization has not been perfectly done.

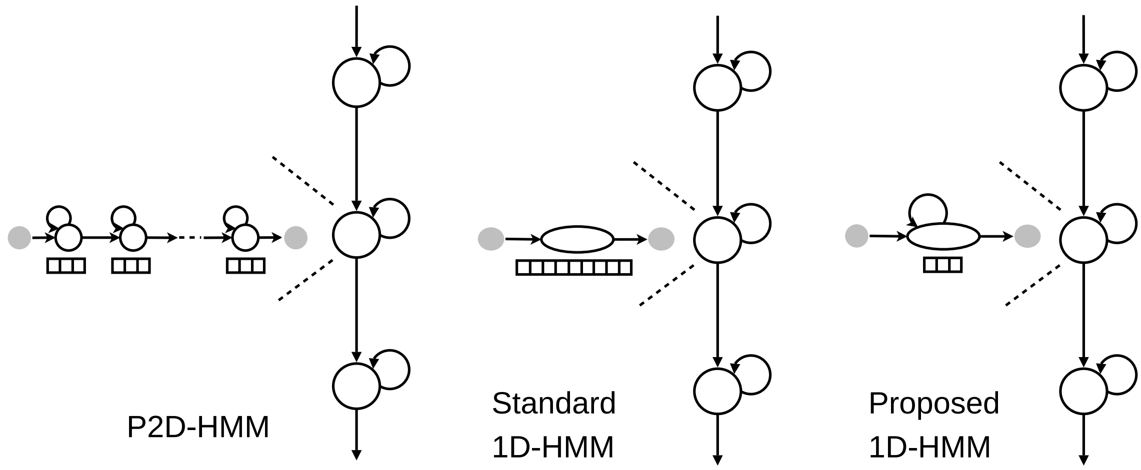


Figure 4.1: 1D-HMMs interpreted as P2D-HMM. The left figure shows the general P2DHMM where the emission probabilities of the main HMM are estimated through an embedded HMM. The figure in the middle shows a particular case of P2D-HMM which corresponds to a standard 1D-HMM where the embedded HMM consists just of one state and only one observation vector representing a whole strip is used. In the left figure the represented P2D-HMM corresponds to the proposed 1D-HMM where the observation vectors represents a block and the transition probabilities are equal.

For the proposed 1D-HMM structure, the observation vectors are extracted from blocks similarly to the GMM and P2D-HMM approaches. Let us denote the sequence of N_S observation vectors representing the consecutive horizontal strips of an image as $X = \{\mathbf{x}_s\}_{s=1}^{N_S}$. Each strip can itself be represented as a sequence of N_B observation vectors $\mathbf{x}_s = \left\{\mathbf{x}_s^b\right\}_{b=1}^{N_B}$ representing the consecutive blocks composing the strip. To model the sequence of horizontal strips X, we use a 1D-HMM with the emission probabilities represented using mixtures of gaussians. If we make the assumption that the feature vectors representing the blocks of a same strip are independent and are generated by the same distribution, the likelihood of the strip s for the state S_j can be estimated with:

$$P(\mathbf{x}_s|q_s = S_j) = \prod_{b=1}^{N_B} P(\mathbf{x}_s^b|\lambda_j) \tag{4.1}$$

Then, the likelihood of an observation sequence X is:

$$P(X|\lambda) = \sum_{\forall Q} \prod_{s=1}^{N_S} \prod_{b=1}^{N_B} P(\mathbf{x}_s^b|\lambda_j) \prod_{s=2}^{N_S} a_{q_{s-1},q_s} \tag{4.2}$$

where a_{q_{s-1},q_s} is the transition probability from state q_{s-1} to state q_s (see Equation 3.5) and $P(\mathbf{x}_b|\lambda_j)$ is estimated with:

$$P(\mathbf{x}_b|\lambda_j) = \sum_{k=1}^{N_G} m_k^s \, \mathcal{N}(\mathbf{x}_s^b|\mu_k^s, \Sigma_k^s) \tag{4.3}$$

As shown in Figure 4.1, this 1D-HMM structure can be seen as a particular case of P2D-HMM where the embedded HMM has only one state and equal transition probabilities. Note that the traditional 1D-HMM structure can also be represented as a P2D-HMM with only one embedded HMM state which emits the entire strip observation vector at once and with null self transition probability [94].

Table 4.3: HTER performance for **manual face localization**

Feature Vectors	Models	Protocol				Time for authentication
		Mc	Ud	Ua	P	of one claim (5 images)
Blocks DCTmod2	P2D-HMM	4.6	15.3	13.1	13.5	19.9s
Blocks DCTmod2	GMM	8.9	17.3	20.9	17.0	1.1s
Blocks DCTmod2	Standard 1D-HMM	6.9	16.3	17.0	14.7	19.9s
Blocks DCT	P2D-HMM	5.6	18.8	13.2	15.1	19.9s
Strips DCT	Standard 1D-HMM	6.6	20.1	20.0	16.6	1.3s
Blocks DCT	**Proposed 1D-HMM**	**5.4**	**16.1**	**17.2**	**15.1**	**2.5s**

4.2.2 Feature extraction

In the previous chapter DCTmod2 feature extraction [80] has been used; compared to traditional 2D DCT, the first three coefficients are replaced by their respective *delta features* in order to reduce the effects of illumination direction changes. DCTmod2 features make the system more robust; however the *delta features* are dependant on the number of overlapped pixels. Since in our system we use different horizontal and vertical overlaps (see Section 4.2.3), the delta features would not be consistant. This motivated the choice of using standard 2D DCT features.

4.2.3 Results and Discussions

We compare the proposed 1D-HMM approach to GMM, standard 1D-HMM and P2D-HMM systems. Similarly to the systems presented in previous chapter, an overlap of four pixels is used for the GMM approach while the strips are overlapped by seven pixels for the standard 1D-HMM and the blocks are overlapped by seven pixels for the P2D-HMM. For the proposed 1D-HMM approach, we choose to use an overlap of seven pixels between consecutive horizontal strips and we don't use horizontal overlap between the blocks. This choice is motivated by the assumption of independence between the blocks made in Equation (4.1) since an overlap between consecutive blocks would increase the dependence.

Manual Face Localization

Table 4.3 shows the results in terms of HTER for manual face localization (we use the manually annotated eye center positions) and the time needed to perform an authentication. The authentication time is given in seconds for a claim which corresponds to five images in the BANCA protocol; the time includes the pre-processing and the experiments were performed on a Pentium IV 3 Ghz. It can be seen that the proposed 1D-HMM performs better than standard 1D-HMM or GMM approaches. The performances are similar to the P2D-HMM when the same features are used, however the computational cost for the proposed 1D-HMM is much less important. We note that better performances are obtained if we use the P2D-HMM with DCTmod2 features.

Automatic Localization

For automatic face localization experiments, similarly to the previous chapter, we use the face detector proposed by Fröba and Ernst in [26]. Results presented in Table 4.4 show that the proposed 1D-HMM is less affected by imperfect localization than the standard 1D-HMM which conserves a rigid spatial constraint in a same strip. Indeed, the proposed 1D-HMM is statistically significantly better than the standard 1D-HMM with a confidence of more than 99.9% using the statistical signifant test proposed in [7]. The P2D-HMM approach which performs an horizontal alignment between the blocks is more robust to imperfect localization with the cost of an authentication time much more important. However, we observe that, using the test of statistical significance,

Table 4.4: HTER performance for **automatic face localization**

Feature Vectors	Models	Protocol			
		Mc	Ud	Ua	P
Blocks DCTmod2	P2D-HMM	6.5	15.9	14.7	14.7
Blocks DCTmod2	GMM	9.5	21.0	24.8	19.5
Blocks DCTmod2	Standard 1D-HMM	13.8	25.9	23.4	21.7
Blocks DCT	P2D-HMM	5.7	18.8	16.5	16.5
Strips DCT	Standard 1D-HMM	20.8	29.3	31.0	27.3
Blocks DCT	**Proposed 1D-HMM**	**9.0**	**20.4**	**22.0**	**18.2**

Table 4.5: HTER performance for Vertical (V) and Horizontal (H) 1D-HMM.

System	Protocol			
	Mc	Ud	Ua	P
V 1D-HMM	5.4	16.1	17.2	15.1
H 1D-HMM	6.7	22.0	23.6	19.4
Combination	5.4	15.7	18.1	14.9

(a) HTER performance for **manual face localization**

System	Protocol			
	Mc	Ud	Ua	P
V 1D-HMM	9.0	20.4	22.0	18.2
H 1D-HMM	12.6	27.4	32.5	25.4
Combination	9.1	21.5	25.1	19.4

(b) HTER performance for **automatic face localization**

the P2D-HMM with DCT features is not statistically significantly better than the proposed 1D-HMM with a confidence of only 70.8%.

4.2.4 Vertical and Horizontal HMM

Previously, we made the choice to perform the main segmentation of the face image vertically. We made this choice since the main decomposition of the face is instinctively from top to the bottom (forehead, eyes, nose, mouth). However, the opposite choice has been made in [24]. It is interesting to see what are the performances if we use the proposed 1D-HMM to perform horizontal segmentation. In this case the image is decomposed in vertical strips, themselves decomposed in blocks.

Since the computation time for this model is relatively low, we can combine Vertical and Horizontal 1D-HMM in order to improve the performances. The combined score for a claim is computed through the weighting sum of the likelihoods of horizontal and vertical models:

$$\Lambda_{comb} = w\Lambda_v + (1-w)\Lambda_h \tag{4.4}$$

where Λ_v and Λ_h are respectively the scores estimated through the vertical and horizontal HMMs, w is the weight factor determined empirically on the validation set and Λ_{comb} is the combined score.

Results are presented in Table 4.5, they demonstrate that the natural decomposition of the face from top to bottom is also the most efficient for automatic face authentication using 1D-HMM. Furthermore, we notice that the combination of horizontal and vertical models does not significantly improve the performance.

4.2.5 Conclusion

In this chapter, two systems have been derived from the baseline systems in order to improve performances with minimal effects in computation time. The first proposition is to extend the feature vectors for the GMM approach in order to embed positional information. This new system improves slightly the performances for an AFA task on the BANCA benchmark database comparing to the baseline GMM approach with the cost of a slight loss in rapidity.

The second proposed approach is an alternative 1D-HMM topology which allows the use of local features (blocks) as observation vectors instead of using a whole strip of the image for standard 1D-HMM implementation. The experiments performed for a face authentication application demonstrate that this model is significantly more robust than the standard 1D-HMM. Due to its low complexity, it is also eight times faster than a P2D-HMM with the cost of a lower accuracy when an automatic localization system is used.

Two implementations of the proposed 1D-HMM are investigated, while the first one performs a vertical segmentation of the face, the second performs an horizontal segmentation. The results clearly demonstrate the superiority of the vertical segmentation.

Since feature extraction approaches including *delta features*, such as DCTmod2 [80], have shown to perform better than standard DCT decomposition, as a future work we plan to investigate a similar feature extraction that could be independent of the degree of overlap, and thus be consistent even if the horizontal and vertical amounts of overlap are not equal.

Chapter 5

On the Choice of a Good Face Localization System

In this chapter, we empirically demonstrate the problems of current localization performance measures when applied to the task of face authentication. We then propose a new performance measuring process which *embed* the face authentication. Preliminary work related to this chapter have been published in [73] and extended experiments are presented in a technical report [11].

As defined in Chapter 2 face localization (FL) is the process of finding the exact position of a face in a given image [98, 91]. It is generally used as an important step in AFR [15, 88] as well as several other applications such as face tracking [37, 17, 85]. We have seen in previous chapters that errors made during this step seriously affect the final performance of AFR. The choice of the localization system is thus crucial. Unfortunately, it is difficult to measure the performance of a face localization algorithm, as no universal criterion has been acknowledged in the literature for this purpose. In fact, we argue in this chapter that such a criterion does not exist and propose instead the use of a criterion that would be specifically tailored for each application for which the localization algorithm would be designed. Here we concentrate on the AFA task.

In that context, the best localization algorithm should be the one that minimizes the number of errors made by the authentication algorithm.

We thus start by analyzing how various kinds of localization errors affect the performance of two different face authentication algorithms. This empirical analysis, presented in Section 5.2, clearly demonstrates that not all localization errors induce the same authentication error.

In a second part, we go one step further: knowing that authentication in itself is not error-free, we propose a new localization measure adapted to the task of authentication. This measure estimates directly the authentication errors as a function of the errors made by the localization algorithm. In this chapter, we estimate this measure using a simple K nearest neighbor (KNN) algorithm. We then show empirically that the localization measure estimated by this simple procedure better reflects the performance of a face localization algorithm when used for a face authentication task.

The chapter is thus organized as follows. Section 5.1 presents an overview of classical measures currently used in the literature in order to evaluate the performance of a face localization algorithm. Section 5.2 then presents two different empirical analyzes that both show that the performance of a localization algorithm can only make sense in the context of the application for which the localization algorithm was built for. This is then followed by Section 5.3, which presents the idea consisting in estimating the error made by the authentication process given the error made by the localization process. Section 5.4 evaluates empirically how this new performance measure behaves on a real benchmark database, and finally Section 3.5 concludes the chapter.

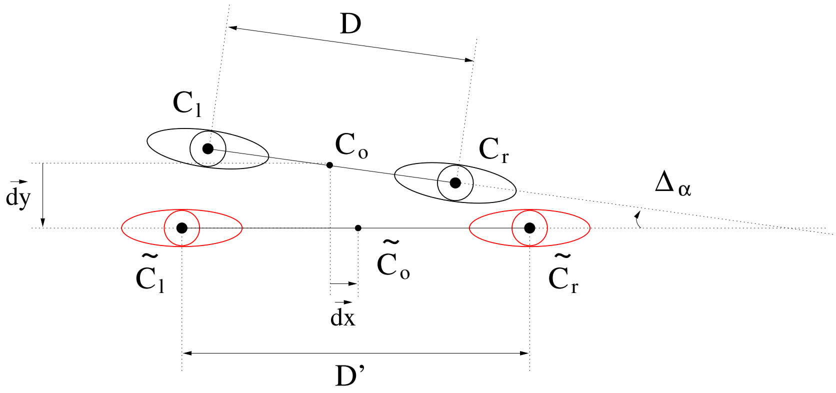


Figure 5.1: Summary of some basic measurements made in face localization. C_l and C_r (resp. $\tilde{C}_l$ and $\tilde{C}_r$) represent the true (resp. the detected) eye positions. C_0 (resp. $\tilde{C}_0$) is the middle of the segment $[C_l C_r]$ (resp. $[\tilde{C}_l \tilde{C}_r]$).

5.1 Performance Measures for Face Localization

5.1.1 Lack of Uniformity

Direct comparison of face localization systems is a very difficult task, mainly because there is no clear definition of what a good face localization means. While most concerned papers found in the literature provide localization and error rates, only few of them mention the way they count a correct/incorrect hit that leads to computation of these rates. Furthermore, when reported, the underlying criterion is usually not clearly described. For instance, in [86] and [38], a detected window is counted as a true or false detection based on the visual observation that the box includes both eyes, the nose and the mouth. According to Yang's survey [97], Rowley *et al.* [75] *adjust the criterion until the experimental results match their intuition of what a correct detection is (i.e. the square window should contain the eyes and also the mouth).* In some rare works, the face localization criterion is more precisely presented. In [48] for instance, Lienhart *et al.* count a correct hit if the Euclidean distance between the centers of the detected and the true face is less than 30% of the width of the true face, and the width of the detected face is within $\pm 50\%$ of the true face. In [27], the authors consider a true detection if the measured face position (through the position of the eyes) and size (through the distance between the eyes) do not differ more than 30% from the true values. Unfortunately, this lack of uniformity between reported results makes them particularly difficult to compare and reproduce.

5.1.2 A Relative Error Measure

Recently, Jesorsky *et al.* [39] introduced a relative error measure based on the distance between the detected and the expected (ground-truth) eye center positions. Let C_l (respectively C_r) be the true left (resp. right) eye coordinate position and let $\tilde{C}_l$ (resp. $\tilde{C}_r$) be the left (resp. right) eye position estimated by the localization algorithm. This measure can be written as

$$d_{eye} = \frac{\max(d(C_l, \tilde{C}_l), d(C_r, \tilde{C}_r))}{d(C_l, C_r)} \tag{5.1}$$

where $d(a, b)$ is the Euclidean distance between positions a and b. A successful localization is accounted if $d_{eye} < 0.25$ (which corresponds approximately to half the width of an eye).

This is, to the best of our knowledge, the first attempt to provide a unified face localization measure. We can only encourage the scientific community to use it and mention it when reporting detection/error rates when the task is localization only. Researchers seem to only start to be aware of this problem of uniformity in the reporting of localization errors and now sometimes report cumulative histograms of d_{eye} [4, 34] (detection rate vs. d_{eye}), but this still concerns only a minority of papers. Furthermore, a drawback of this measure is that it is not possible to differentiate errors in translation, rotation and scale.

5.1.3 A More Parametric Measure

More recently, Popovici *et al.* [69] proposed a new parametric scoring function whose parameters can be tuned to more precisely penalize each type of errors. Since face localization is often only a first step of a more complex face processing system (such as a face recognition module), analyzing individually each type of errors may provide useful hints to improve the performance of the upper level system.

In the same spirit as [69], let us now define four basic measures to represent the difference in horizontal translation (Δ_x), vertical translation (Δ_y), scale (Δ_s) and rotation (Δ_α):

$$\Delta_x = \frac{\overline{dx}}{d(C_l, C_r)} , \tag{5.2}$$

$$\Delta_y = \frac{\overline{dy}}{d(C_l, C_r)} , \tag{5.3}$$

$$\Delta_s = \frac{d(\tilde{C}_l, \tilde{C}_r)}{d(C_l, C_r)} , \tag{5.4}$$

$$\Delta_\alpha = \widehat{\overrightarrow{C_l C_r}, \overrightarrow{\tilde{C}_l \tilde{C}_r}} , \tag{5.5}$$

where $\overline{dx}$ is the algebraic measure of vector $\overrightarrow{dx}$. All these measures are summarized in Figure 5.1. The four delta measures are easily computed given the ground-truth eye positions (C_l and C_r) and the detected ones ($\tilde{C}_l$ and $\tilde{C}_r$). Furthermore, as it will appear useful later in this chapter, one can artificially create detected positions given these four delta measures. Note finally that both the choices of Jesorsky's threshold (0.25) and Popovici's weights on each of these delta measures (in order to obtain a single measure) still remain subjective.

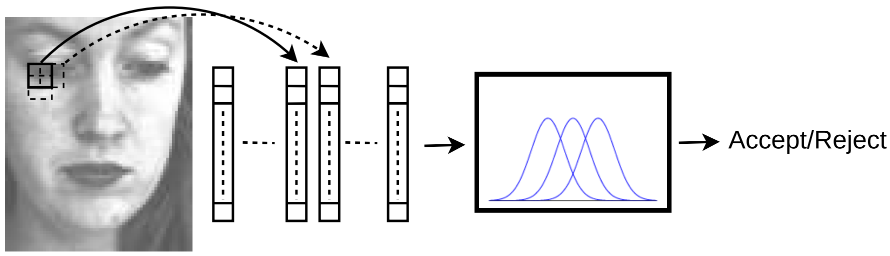
5.1.4 Application-Dependent Measure

Here, we argue that a universal objective measure for evaluating face localization algorithms *does not exist*. A given localized face may be correct for the task of initializing a face tracking system [37], but may not be accurate enough for a face authentication system [15]. We therefore think that there can be no absolute definition of what a *good face localization* is. We rather suggest to look for an application-dependent measure representing the final task. Moreover, we have previously demonstrated that the authentication score obtained with a perfect (manual) localization is significantly better than the authentication score obtained with a not-so-perfect (automatic) localization, which shows the importance of measuring accurately the quality of a face localization algorithm for authentication.

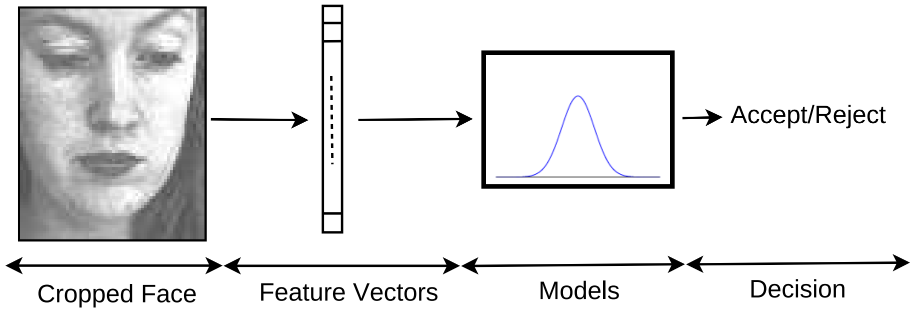
Hence, we will empirically show how face localization errors affect face authentication results, and how it can be more accurately measured than using currently proposed measures.

5.2 Robustness of Current Measures

The purpose of this Section is to analyze the relation between the tasks of face localization and face authentication, by observing how errors reported by the FL system affect the AFA system. We start by



(a) DCT/GMM



(b) PCA/Gaussian

Figure 5.2: Conceptual representations of the two face authentication systems

observing, in Section 5.2.1, the performance of an AFA system when we artificially introduce some localization errors in the tested face images. Then, in Section 5.2.2, we empirically demonstrate for a particular case that a generic face localization measure is not accurate. These preliminary experiments are performed on the XM2VTS database using the associated protocol. The experiments were carried out with the two different AFA approaches, namely DCT/GMM and PCA/Gaussian. These two systems have been introduced in Chapter 3 : while DCT/GMM refers to the GMM based system presented in Section 3.1.1, the PCA/Gaussian is similar to the reference PCA based system briefly described in Section 5.4. Conceptual examples of the DCT/GMM and PCA/Gaussian systems are represented in Figure 5.2.

The models are trained with manually located images and the decision threshold is chosen *a priori* at EER on the validation set (also using manually located images). The AFA systems are thus independent of the FL system used. The FAR, FRR and HTER performance measures are then computed with perturbed face images from the test set.

5.2.1 Effect of FL Errors

In Section 5.1.3, four types of localization errors were defined: horizontal and vertical translations (respectively Δ_x and Δ_y), scale (Δ_s) and rotation (Δ_α). As a preliminary analysis, we studied how each type of localization error affects the AFA performance. Specifically, the eye positions were artificially perturbed in order to generate a configurable amount of translation (horizontal and vertical), scale and rotation errors. Then experiments were performed for each type of errors independently; i.e. when we generated one type of perturbation, the others were kept null.

Figure 5.3 shows the AFA performance as a function of the generated perturbations for the two AFA systems. Several conclusions can be drawn from these curves:

1. Regarding HTER curves, as expected, the AFA performance is affected by localization errors. The minimum of the HTER curves are always obtained at the ground-truth positions.
2. In the tested range, only the FRR is sensitive to localization errors, the FAR is not significantly affected. In other words, localization errors in a reasonable range do not induce additional false acceptances.
3. HTER curves demonstrate that the two FL approaches are not affected in the same way. Generally, the DCT/GMM system is more robust to perturbed images than the PCA/Gaussian system; justification of this result is discussed further in Chapter 3. Moreover, we remark that the two systems are not sensitive to the same type of errors; while DCT/GMM is affected by scale and rotation errors and very robust to translation errors, the PCA/Gaussian system is very sensitive to all types of errors, including translation.

5.2.2 Indetermination of d_{eye}

In Section 5.1, we discussed the important problem of a universal measure to evaluate face localization performance, in order to get fair and clean system comparisons. We also introduced the currently unique existing measure, proposed by Jesorsky *et al.* [39], based on the true and the detected eye positions (5.1). We also underlined that this measure does not differentiate errors in translation, scale or rotation.

For the specific task of AFA, prior empirical evidence showed that the performance is closely related to the accuracy of the face localization system. In Section 5.2.1, we went further by explaining that this performance is closely related to the type of error introduced by the FL system and that this dependency varies from one AFA system to another (eg. DCT/GMM vs PCA/Gaussian). We then argued that a universal criterion like d_{eye} is not adapted to the final task of AFA and that we thus need to search for an application-dependent measure.

To illustrate this more clearly, let us look again at the d_{eye} measure and show why it is not adapted to the AFA task. In order to understand the limitations of this measure, we analyzed each type of localization error independently, as done in Section 5.2.1.

Table 5.1: For the specific case of $d_{eye} = 0.2$, the first column contains the corresponding Δ values and the third column contains the resulting HTER

delta error	d_{eye}	HTER
$\Delta_x = -0.2$	0.2	5.27
$\Delta_x = 0.2$	0.2	5.43
$\Delta_y = -0.2$	0.2	4.14
$\Delta_y = 0.2$	0.2	3.27
$\Delta_s = 0.6$	0.2	31.75
$\Delta_s = 1.4$	0.2	24.65
$\Delta_\alpha = 23°$	0.2	32.35
$\Delta_\alpha = -23°$	0.2	31.24

We first arbitrarily selected a value of $d_{eye} = 0.2$, which commonly means that the detected pattern is a face (since it is lower than 0.25). We then selected all kinds of delta errors which would yield $d_{eye} = 0.2$. Figure 5.4 shows examples of localizations obtained for each of these delta errors. The corresponding Δ values are reported in the first column of Table 5.1. The last column shows the resulting face authentication

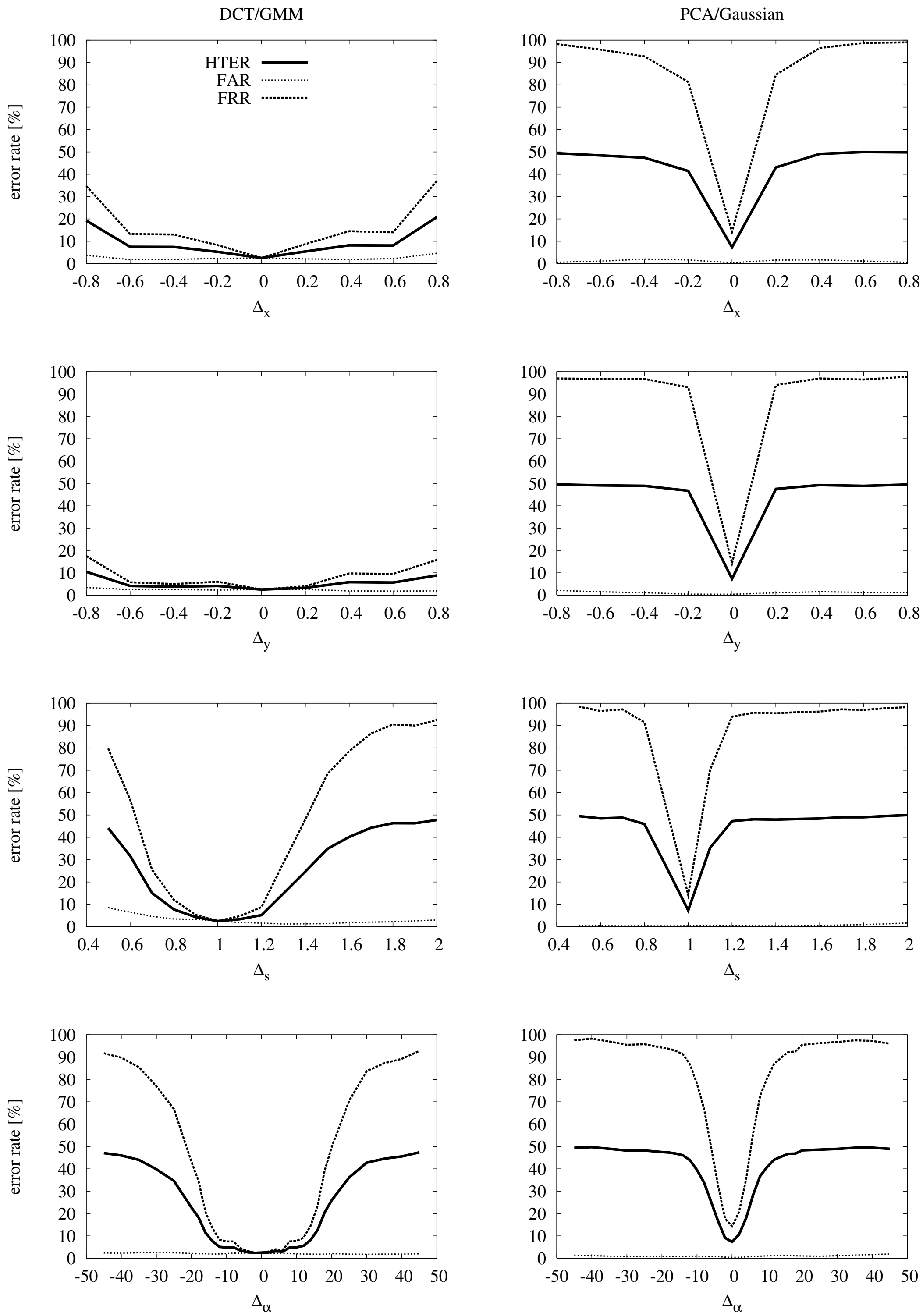


Figure 5.3: Face authentication performance (in terms of FAR, FRR and HTER error rates) as a function of face localization errors. The error rates are shown for the DCT/GMM (left column) and for the PCA/Gaussian (right column) face authentication systems.

performance, in terms of HTER, using the DCT/GMM face authentication system. This experiment basically shows the following:

1. There is a significant variation in HTER for the same value of d_{eye}.

2. The DCT/GMM system is more robust to errors in translation than to errors in scale or rotation (for the same $d_{eye} = 0.2$).

Note that in practice, a face detector does not fail only on one type of error. However, this experiment clearly shows that a face localization performance measure such as d_{eye} is not adapted if we want to take into account the performance of the whole system.

5.3 Approximate Face Authentication Performance

The preliminary experiments conducted in Section 5.2 should have convinced that current FL measures are not adapted to the AFA task, and we also argued that it is probably not adapted to any other particular task. Hence, as explained in Section 5.1, instead of searching for a universal measure assessing the quality of a face localization algorithm, we propose here to estimate a specific performance measure adapted to the target task. We here concentrate on the task of face authentication, hence a good face localization algorithm in that context is a module which produces a localization such that the expected error of the face authentication module is minimized. More formally, let $\mathbf{x}_i$ be the input vector describing the face of an access i, $\mathbf{y}_i = \mathrm{FL}(\mathbf{x}_i)$ be the output of a face localization algorithm applied to $\mathbf{x}_i$ (generally in terms of eye positions), $\mathbf{z}_i = \mathrm{AFA}(\mathbf{y}_i)$ be the decision taken by a face authentication algorithm (generally accept or reject the access) and $\epsilon = \mathrm{Error}(\mathbf{z}_i)$ be the error generated by this decision. The ultimate goal of a face localization algorithm in the context of a face authentication task is thus to minimize the following criterion:

$$\mathrm{Cost} = \sum_i \mathrm{Error}(\mathrm{AFA}(\mathrm{FL}(\mathbf{x}_i))) \; . \tag{5.6}$$

Our proposed solution for a meaningful FL measure adapted to a given task is thus to embed all subsequent functions (AFA and Error) into a single box and to estimate this box using some universal approximator:

$$\mathrm{Cost} = \sum_i f(\mathrm{FL}(\mathbf{x}_i); \theta) \tag{5.7}$$

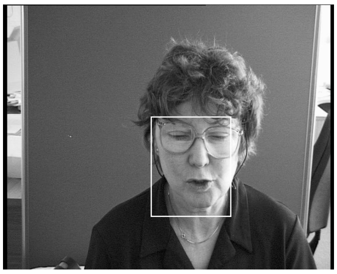
where $f(\cdot; \theta)$ is a parametric function that would replace the rest of the process following localization using parameters θ. In this chapter, we consider as function $f(\cdot)$ a simple K nearest neighbor (KNN) algorithm [9]. In order to be independent of the precise localization of the eyes, we modified in fact slightly this approach by changing the input of function $f(\cdot)$ in order to contain instead the error made by the localization algorithm in terms of very basic measures: Δ_x, Δ_y, Δ_s and Δ_α, as described in Section 5.1. Let $\mathrm{GT}(\mathbf{x}_i)$ be the ground-truth eyes position of $\mathbf{x}_i$ and $\mathrm{Err}(\mathbf{y}_i, \mathrm{GT}(\mathbf{x}_i))$ be the function that produces the face localization error vector; we thus have

$$\mathrm{Cost} = \sum_i f(\mathrm{Err}(\mathrm{FL}(\mathbf{x}_i), \mathrm{GT}(\mathbf{x}_i)); \theta) \; . \tag{5.8}$$

In order to train such a function $f(\cdot)$, we used the following methodology. First, in order to cover the space of localization errors, we create artificial examples based on all available training accesses. The training examples of $f(\cdot)$ are thus uniformly generated by adding small perturbations (localization errors) bounded by a reasonable range. For each generated example, an authentication is performed and a corresponding target value of 1 (respectively 0) is assigned when an authentication error appears (respectively does not appear).

(a) ground-truth ($d_{eye} = 0.0$)

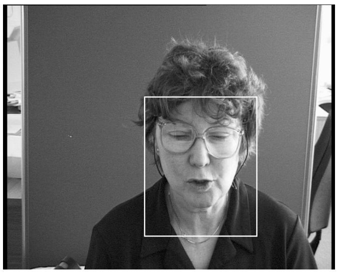
(b) $\Delta_x = 0.2$ ($d_{eye} = 0.2$)



(c) $\Delta_x = -0.2$ ($d_{eye} = 0.2$)

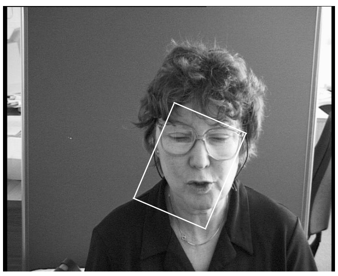
(d) $\Delta_y = 0.2$ ($d_{eye} = 0.2$)

(e) $\Delta_y = -0.2$ ($d_{eye} = 0.2$)

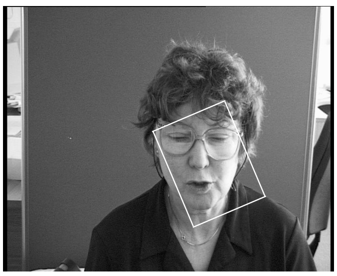


(f) $\Delta_s = 1.4$ ($d_{eye} = 0.2$)

(g) $\Delta_s = 0.6$ ($d_{eye} = 0.2$)



(h) $\Delta_\alpha = 23^\circ$ ($d_{eye} = 0.2$)



(i) $\Delta_\alpha = -23^\circ$ ($d_{eye} = 0.2$)

Figure 5.4: Figure (a) shows the face bounding box for the ground-truth annotation. For the given value of $d_{eye} = 0.2$, Figures (b) to (i) illustrate the bounding box resulting from perturbations in horizontal translation (b,c), vertical translation (d,e), scale (f,g) and rotation (h,i).

5.4 Experiments and Results

This section is devoted to verifying experimentally if our proposed method to measure the performance of localization algorithms in the context of a face authentication task improves with respect to other known measures.

5.4.1 Training Data

The XM2VTS database was used to generate examples to estimate our function $f(\cdot)$, which should yield the expected authentication error given a localization error. For each of the 1000 available client images[1], 50 localization errors were randomly generated following a uniform distribution in a predefined interval $[-1, 1]$ for Δ_x and Δ_y, $[0.5, 1.5]$ for Δ_s and $[-20°, 20°]$ for Δ_α. The training set thus contains 50000 examples. An authentication is performed for each example, which will be assigned a target value of 1 (respectively 0) when the authentication algorithm accepts the client (respectively rejects him). Furthermore, a separate validation set of 50000 examples was created using the same procedure (with the same set of clients, but a different random seed). The hyper-parameter K of the KNN model, which controls the capacity [90] of $f(\cdot)$, was then chosen as the one which minimized the out-of-sample error on the validation set.

5.4.2 Face Localization Performance Measure

Given the set of errors $\Delta = \{\Delta_x, \Delta_y, \Delta_s, \Delta_\alpha\}$ generated by the FL algorithm on an image n we define the error of the KNN localization algorithm as:

$$E_{\text{KNN}}(\Delta^n) = \frac{1}{K} \sum_{k \in \text{KNN}(\Delta^n)} C_k \tag{5.9}$$

where KNN(Δ^n) is the set of the K nearest training examples of Δ^n and C_k is the error made on example k defined as:

$$C_k = \begin{cases} 0 & \text{if} \quad \text{Accepted Client} \\ 1 & \text{if} \quad \text{Rejected Client} \ . \end{cases} \tag{5.10}$$

We then estimate the performance of the FL system on a set of N images using:

$$E_{\text{KNN}} = \frac{1}{N} \sum_{n=1}^{N} E_{\text{KNN}}(\Delta^n) \ . \tag{5.11}$$

Similarly, we measure the error made by the d_{eye} measure as follows:

$$E_{eye}(n) = \begin{cases} 0 & \text{if} \quad \text{Accepted Client and } d_{eye}(n) < 0.25 \\ 1 & \text{if} \quad \text{otherwise} \end{cases} \tag{5.12}$$

and

$$E_{eye} = \frac{1}{N} \sum_{n=1}^{N} E_{eye}(n) \ . \tag{5.13}$$

[1] The preliminary analysis of Section 5.2.1 showed that FAR is not significantly affected by localization errors, so we did not use any impostor access for this step.

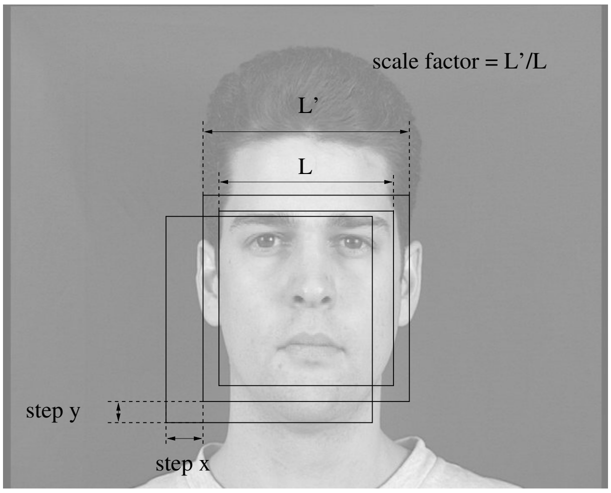


Figure 5.5: Face localization scanning parameters: step x, step y and scale factor. The choice of these parameters both affects the speed of the system as well as accuracy.

5.4.3 KNN Function Evaluation

In order to verify that the obtained KNN function is robust to the choice of the training dataset, we chose to evaluate it on another dataset, namely the English BANCA corpus (see Chapter 2). In order to extract the faces from the access images, we used, similarly to previous chapters, the face detector proposed by Fröba and Ernst in [26] described briefly in Chapter 3, Section 3.3.1. This system involves some scanning parameters typically chosen empirically, such as horizontal and vertical steps and scale factor (see Figure 5.5).

When minimizing these parameters, the localization is expected to be more accurate, however the computational cost then becomes intractable. These two parameters should thus be selected in order to have a good *performance/computational cost* trade-off. In order to obtain a good trade-off we can either favor translation accuracy by reducing horizontal and vertical steps or scale accuracy by reducing the scale factor.

We thus decided to test two different versions of the localization system, as follows:

1. The first system, FL_{shift}, uses larger values for horizontal and vertical step factors. This system is expected to introduce more errors in translation.

2. The second system, FL_{scale}, uses finer step factors, but a larger scale factor, expected to introduce errors in scale.

We aim to verify that our KNN function is able to measure correctly which one of this two FL system is the best (ie. the one that minimize the AFA error). Table 5.2 compares the localization errors (the smaller the better for all compared measures) obtained with the d_{eye} criterion (second column) computed using Equation (5.13), our proposed function (third column) computed using equation (5.11), and the actual authentication score decomposed into its FAR, FRR and HTER components obtained with the DCT/GMM AFA, on all the accesses of the BANCA database using protocol P. The findings of this experiment can be summarized as follow:

1. As expected, the best authentication score (HTER $= 19.5$) is obtained with perfect localization similarly to the first conclusion of Section 5.2.1 Then, follows the FL_{shift} system (HTER $= 21.0$) and finally the FL_{scale} system. This ordering is the one we wish to find using a face localization quality measure.

2. Our proposed function correctly order the localization system (FL_{shift}), while the d_{eye}-based measure fails to order the two FL systems. This can be mainly explained because the d_{eye} measure does not

differentiate errors in translation, shift or rotation, while the DCT/GMM system is more affected by some specific types of error (third conclusion of Section 5.2.1).

3. We remark that the FAR corresponding to the FL_{shift} system (11.7) and the FL_{shift} system (14.7) are lower than the FAR with perfect localization (15.1). We think this is because, for impostor accesses, a bad face localization only helps the system to reject the access, yielding a lower FAR.

While the d_{eye}-based measure fails to identify the best localization system (hence the system which minimizes the AFA error), our proposed function correctly orders the two modules.

Furthermore, the proposed KNN measure is very fast to compute. It only takes 20 ms on a PIV 2.8 Ghz to evaluate an image access, while it would take 350 ms for the DCT/GMM system (preprocessing, feature extraction and classification).

Table 5.2: Comparison of two FL performance measures for two face localization systems as well as for a perfect localization (ground-truth). The last 3 columns contains the face authentication score in terms of FAR, FRR and HTER for the DCT/GMM system.

FL Systems	Measures		Authentication		
	E_{eye}	E_{KNN}	FAR [%]	FRR [%]	HTER [%]
ground-truth	0.00	0.05	15.1	23.9	19.5
FL_{shift}	0.10	0.12	11.7	30.3	21.0
FL_{scale}	0.04	0.15	14.7	33.8	24.3

5.5 Conclusion

In this chapter, we have proposed a novel methodology to compare face localization algorithms in the context of the particular application of face authentication. Note that the same methodology could have been applied to any other task that builds on localization, such as face tracking. We have first shown that current measures used in face localization are not accurate. We have thus proposed a method to estimate the authentication errors induced specifically by the use of a particular face localization algorithm. This measure can then be used to compare more precisely several localization algorithms. We tested our proposed measure using the BANCA database on a face authentication task, comparing two different face localization algorithms. Results show that our measure does indeed capture more precisely the differences between localization algorithms (when applied to authentication tasks), which can be useful to select an appropriate localization algorithm. Furthermore, our function is robust to the training dataset (training on XM2VTS and test on BANCA) and compared to the DCT/GMM face authentication system, the KNN performs more than 15 times faster (no preprocessing and feature extraction steps). Finally, in order to compare FL modules, we do not need to run face authentication on the entire database, but we only use our function on a subset of face images.

In fact, one can view the process of training a localization system as a selection procedure where one simply selects the best localization algorithm according to a given criterion. In that respect, an interesting future work could concentrate on the use of such a measure to effectively *train* a face localization system for the specific task of face authentication.

Chapter 6

Conclusion and Perspectives

6.1 General Summary

In this thesis, the problem of fully automatic face authentication in weakly constrained environment was investigated. Particular attention was devoted to generative model based approaches. Main drawbacks of existing algorithms using generative models have been enlightened. Then an alternative training strategy has been proposed as well as new features and models which allow good performance / speed trade-offs. Furthermore, we have pointed out the importance of using an automatic localization system to evaluate face authentication performances. And finally, we proposed an evaluation measure for the face localization module in order to choose the most accurate for a specific face authentication system. The most important results achieved in this thesis are summarized in the following:

- Using Maximum *a Posteriori* based training results in considerably more precise models than with ML training approaches, leading to higher discrimination performance.
- The P2D-HMM approach is overall the most robust and obtains the best discrimination performance, when compared to the 1D-HMM and GMM based approaches. However, it is also the most computationally intensive approach, making it impractical for application use on current hardware.
- The performance of the GMM approach can be improved by embedding positional information in the feature vectors.
- In Chapter 4, we proposed an alternative 1D-HMM topology which allows the use of block based features (unlike the traditional 1D-HMM approach which use strip based feature vectors). Due to its low complexity, it is also eight times faster than P2D-HMM with the cost of a lower accuracy.
- Systems that utilize rigid spatial constraints between face parts (such as PCA and 1D-HMM based systems), are easily affected by face localization errors, which are caused by an automatic face locator. In contrast, systems which have relaxed constraints (such as GMM and P2D-HMM based systems), are quite robust.
- Good performance on manually located faces does not necessarily reflect good performance in real life conditions, where an automatic localization system must be used. As automatic localization cannot guarantee perfect face localization, this indicates that any new technique must be designed from the ground up to handle imperfectly located faces.
- We have demonstrated empirically that the natural decomposition of the face from top to bottom (forehead, eyes, nose, mouth) is also the most efficient for automatic face recognition.

- We have shown that current measures used in face localization are not accurate when localization is used as a preliminary step of a face authentication system.
- We have proposed a novel methodology to compare face localization algorithms for a face authentication task. We have thus investigated a method to estimate the authentication errors indused specifically by the use of a particular face localization algorithm. Results show that our measure does indeed capture more precisely the differences between localization algorithms when applyed to authentication task which can be useful to select an appropriate localization algorithm.

6.2 Possible Future Directions

Although the algorithms presented in this thesis have shown to perform well, interesting issues remain open.

Illumination Invariant Face Recognition. In this thesis, we considered closely the problems of feature extraction and classification as well as the performance measure of the face localization, whereas the lighting normalization step was neglected. In our experiments, we used a simple histogram normalization. In [83] it has been experimentally demonstrated that more advanced methods in addition to histogram normalization improved significantly the performance of face authentication. This result can probably be extended to generative model based face authentication. Therefore image normalization approaches based on Land's "retinex" theory [46] such as Gross and Brajovic's algorithm [32] is worth investigating and should make the systems proposed in this thesis more robust to illumination variation. This issue is in fact currently under investigation [36, 35].

Discriminant Model Parameters. In the literature, several papers propose the use of discriminant models for face recognition [52, 12, 42]. However all these algorithms use holistic features which do not allow spatial alignment (problems of holistic features are discussed in Chapter 3). Moreover the model parameters are difficult to estimate due to the large size of the feature vectors and the small amount of training data available which could lead to *overtrained* models; this problem is pointed out in [15]. In this thesis we propose the use of generative models trained using MAP adaptation which overcomes the problems of alignment and lack of training data. Although few parameters are found in a discriminant fashion (hyper-parameters and the decision threshold are chosen to minimize EER on a validation set), the majority of the parameters of the proposed models (variances, means and weights of the gaussians as well as transition probabilities) are not trained specifically to discriminate between the true client and impostors. It would be interesting to investigate a way to train these parameters in a discriminant fashion.

P2D-HMM for Face alignment. We have seen in this thesis that P2D-HMM based approach obtains the overall best performance for the face authentication task. That result suggests that P2D-HMM is a good model for face segmentation and thus could be used for face and facial feature localization. In some ways some state of the model can be attributed to the face and some can be attributed to the background. A similar approach has been proposed for body tracking in [72]. This method would be particularly attractive since it would not involve explicit shape models and thus can be expected to be robust for any kind of face shape.

Appendix A

Acronyms

1D-HMM	One-dimensional hidden Markov model
2D	Two-dimensional
AFA	Automatic face authentication
AFR	Automatic face recognition
DCT	Discrete cosine transform
DET	Detection Error trade-off
EER	Equal error rate
EM	Expectation-maximization
FA	False acceptance
FAR	False acceptance rate
FL	Face localization
FR	False rejection
FRR	False rejection rate
GM	Generative models
GMM	Gaussian mixture models
HMM	Hidden Markov model
LDA	Linear discriminant analysis
LP	Lausanne protocol
MAP	Maximum *a posteriori*
ML	Maximum likelihood
NC	Normalized correlation
PCA	Principal component analysis
P2D-HMM	Pseudo two-dimensional hidden Markov model
ROC	Receiver operating characteristic
SVM	Support vector machine
WCE	Weakly constrained environment

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, July 1997.

[2] W. Atkins. A testing time for face recognition technology. *Biometric Technology Today*, 9(3):8–11, 2001.

[3] 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*, Guilford, UK, 2003.

[4] S. Behnke. Face localization in the neural abstraction pyramid. In *International Conference on Knowledge-Based Intelligent Information and Engineering Systems, KES*, pages 139–145, 2003.

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

[6] S. Bengio and J. Mariéthoz. The expected performance curve: a new assessment measure for person authentication. In *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop*, pages 279–284, Toledo, 2004.

[7] S. Bengio and J. Mariéthoz. A statistical significance test for person authentication. In *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop*, 2004.

[8] J.R. Beveridge, K. She, B.A. Draper, and G.H Givens. A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 535–542, USA, 2001.

[9] C. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[10] V. Blanz and S. Romdhani. Face identification across different poses and illuminations with a 3d morphable model. In *AFGR '02: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, page 202, Washington, DC, USA, 2002. IEEE Computer Society.

[11] F. Cardinaux. Local Features and 1D-HMMs for Fast and Robust Face Authentication. IDIAP-RR 17, IDIAP, 2005.

[12] F. Cardinaux and S. Marcel. Face verification using MLP and SVM. In *XI Journees NeuroSciences et sciences pour l'Ingenieur (NSI 2002)*, number 21, La Londe Les Maures, France, 2002.

[13] F. Cardinaux, C. Sanderson, and S. Bengio. Face Verification Using Adapted Generative Models. In *The 6th International Conference on Automatic Face and Gesture Recognition, FG2004*, Seoul, Korea, 2004. IEEE.

[14] F. Cardinaux, C. Sanderson, and S. Bengio. User Authentication via Adapted Statistical Models of Face Images. *To appear in IEEE Transaction on Signal Processing*, 2005.

[15] F. Cardinaux, C. Sanderson, and S. Marcel. Comparison of MLP and GMM Classifiers for Face Verification on XM2VTS. In *4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA*, pages 911–920, Guilford, UK, 2003.

[16] H. Chen, P. Belhumeur, and D. Jacobs. In search of illumination invariants. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 254 – 261, 2000.

[17] D. Comaniciu and V. Ramesh. Robust detection and tracking of human faces with an active camera. In *Proceedings of the Third IEEE International Workshop on Visual Surveillance, IEEE Computer Society, Dublin, Ireland*, pages 11–18, 2000.

[18] J. Czyz. *Decision Fusion in Identity Verification using Facial Images*. PhD thesis, Université catholique de Louvain, December 2003.

[19] J. Daugman. Uncertainty relation for resolution in space, spacial frequency and orientation optimized by two-dimentional visual cortical filters. *Journal of Optical Society of America*, 2(7), 1985.

[20] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B*, 39:1–38, 1977.

[21] A. Douglas, Reynolds, F. Q. Thomas, and B. D. Robert. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1–3), 2000.

[22] R. O. Duda, P. E. Hart, and G. S David. *Pattern Classification*. Wiley, 2001.

[23] J.-L. Dugelay, J.-C. Junqua, C. Kotropoulos, R. Kuhn, F. Perronnin, and I. Pitas. Recent advances in biometric person authentication. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, volume IV, pages 4060–4062, 2002.

[24] S. Eickeler, S. Müller, and G. Rigoll. Recognition of jpeg compressed face images based on statistical methods. *Image and Vision Computing*, 18(4):279–287, 2000.

[25] Y. Freund and R.E. Shapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, (14):771–780, 1999.

[26] B. Fröba and A. Ernst. Face detection with the modified census transform. In *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition (AFGR)*, pages 91–96, 2004.

[27] B. Fröba and C. Küblbec. Robust face detection at video frame rate based on edge orientation features. In *IEEE Conference on Automatic Face and Gesture Recognition, AFGR*, 2002.

[28] J. L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observation of markov chains, 1994.

[29] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):643–660, 2001.

[30] R. C. Gonzales and R. E. Woods. *Digital Image Processing*. Addison-Wesley, 1992.

[31] M. Gray, J. Movellan, and T. Sejnowski. A comparison of local versus global image decomposition for visual speechreading. In *the 4th Joint Symposium on Neural Computation*, 1997.

[32] R. Gross and V. Brajovic. An image preprocessing algorithm for illumination invariant face recognition. In *4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*. Springer, June 2003.

[33] R. Gross, J. Yang, and A. Waibel. Growing gaussian mixture models for pose invariant face recognition. In *Proceedings of the15th International Conference on Pattern Recognition (ICPR 2000)*, volume 1, pages 1088 – 1091, September 2000.

[34] M. Hamouz, J. Kittler, J.-K. Kamarainen, P. Paalanen, and H. K älviäinen. Affine-invariant face detection and localization using gmm-based feature detector and enhanced appearance model. In *IEEE Conference on Automatic Face and Gesture Recognition, AFGR*, pages 67–72, 2004.

[35] G. Heusch, F. Cardinaux, and S. Marcel. Efficient diffusion-based illumination normalization for face verification. IDIAP-RR 46, IDIAP, 2005.

[36] G. Heusch, F. Cardinaux, and S. Marcel. Lighting normalization algorithms for face verification. IDIAP-COM 03, IDIAP, 2005.

[37] K.S. Huang and M.M. Trivedi. Robust real-time detection, tracking, and pose estimation of face in video streams. In *International Conference on Pattern Recognition, ICPR*, 2004.

[38] R.-J. Huang. *Detection Strategies for Face Recognition Using Learning and Evolution*. PhD thesis, George Mason University, Fairfax, Virginia, 1998.

[39] O. Jesorsky, K. Kirchberg, and R. Frischholz. Robust face detection using the hausdorff distance. In *3rd International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA*, pages 90–95, 2001.

[40] D. J. Jobson, Z.-R., and G. A. Woodell. A multi-scale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image Processing*, 6(7), july 1997.

[41] D. J. Jobson, Z.-R., and G. A. Woodell. Properties and performance of a center/surround retinex. *IEEE Transactions on Image Processing*, 6(3):451–462, march 1997.

[42] K. Jonsson, J. Kittler, Y. P. Li, and J. Matas. Support vector machines for face authentication. *Image and Vision Computing*, 2002.

[43] B. Kamgar-Parsi, B. Kamgar-Parsi, A. Jain, and J. Dayhoff. Aircraft detection: A case study in using human similarity measure. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(12):1404–1414, 2001.

[44] J. Kittler, Y. Li, and J. Matas. On matching scores for LDA-based face verification. In *The 11th British Machine Vision Conference (BMVC2000)*, pages 42–51, Bristol, UK, September 2000.

[45] M. Lades, J. Vorbruggen, J. Buhmann, J. Lange, C. V. D. Malburg, and R. Wurtz. Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Comput.*, 42:300–311, 1993.

[46] E. H. Land and J. J. McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61:1–11, 1971.

[47] Y. Li, J. Kittler, and J. Matas. On matching scores of LDA-based face verification. In *Proceedings of the British Machine Vision Conference BMVC2000*, 2000.

[48] R. Lienhart, A. Kuranov, and V. Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. Technical report, Intel Corporation, USA, 2002.

[49] M. Lockie. Facial verification bureau launched by police it group. *Biometric Technology Today*, 10(3):3–4, 2002.

[50] J Lüttin. Evaluation protocol for the the XM2FDB database (lausanne protocol). Technical Report COM-05, IDIAP, 1998.

[51] J. Lüttin and G. Maître. Evaluation protocol for the extended m2vts database (xm2vtsdb). IDIAP Communication 98-05, IDIAP Research Institute, Martigny, Switzerland, 1998.

[52] S. Marcel, C. Marcel, and S. Bengio. A state-of-the-art Neural Network for robust face verification. In *Proceedings of the COST275 Workshop on The Advent of Biometrics on the Internet*, Rome, Italy, 2002.

[53] J. Mariéthoz and S. Bengio. A comparative study of adaptation methods for speaker verification. In *International Conference on Spoken Language Processing ICSLP*, pages 581–584, Denver, CO, USA, 2002.

[54] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Proceedings of Eurospeech'97, Rhodes, Greece*, pages 1895–1898, 1997.

[55] A. M. Martinez and A. C. Kak. PCA versus LDA. In *IEEE Transactions on pattern analysis and machine intelligence, Vol 23*, 2001.

[56] K. Messer, J. Kittler, M. Sadeghi, M. Hamouz, A. Kostyn, S. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, N. Poh, Y. Rodriguez, K. Kryszczuk, J. Czyz, L. Vandendorpe, J. Ng, H. Cheung, and B. Tang. Face authentication competition on the BANCA database. In *International Conference on Biometric Authentication, ICBA*, 2004.

[57] B. Moghaddam and A. Pentland. *Probabilistic Visual Learning for Object Representation*. Oxford University Press, 1996.

[58] B. Moghaddam, W. Wahid, and A. Pentland. Beyond eigenfaces: Probabilistic matching for face recognition. In *The 3rd IEEE Int'l Conference on Automatic Face and Gesture Recognition*, pages 30–35, Nara, Japan, April 1998.

[59] A. Nefian and M. Hayes. Face recognition using an embedded HMM. In *Proceedings of the IEEE Conference on Audio and Video-based Biometric Person Authentication, AVBPA*, pages 19–24, 1999.

[60] A. Nefian and H. Monson. Maximum likelihood training of the embedded HMM for face detection and recognition. In *IEEE International Conference on Image Processing*, volume 1, pages 33–36, Vancouver, BC, Canada, 2000.

[61] J. Ortega-Garcia, J. Bigün, D. Reynolds, and J. Gonzales-Rodriguez. Authentication gets personal with biometrics. *IEEE Signal Processing Magazine*, 21(2):50–62, 2004.

[62] H. Othman and T. Aboulnasr. A separable low complexity 2D HMM with application to face recognition. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, 2003.

[63] C. Padgett and G. Cottrell. Representing face images for emotion classification. *Advances in Neural Information Processing Systems*, 9, 1997.

[64] P. Penev and J. Atick. Local feature analysis: a general statistical theory for object representation. *Network:Computation in Neural Systems*, 7(3), 1996.

[65] F. Perronnin and J.-L. Dugelay. A model of illumination variation for robust face recognition. In *MMUA'03, Workshop on Multimodal User Authentication*, pages 11–12, Dec 2003.

[66] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[67] J. Picone. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9):1215–1247, 1993.

[68] N. Poh and S. Bengio. Improving single modal and multimodal biometric authentication using f-ratio client-dependent normalisation. IDIAP-RR 52, IDIAP, 2004.

[69] V. Popovici, Y. Rodriguez, J.-P. Thiran, and S. Marcel. On performance evaluation of face detection and localization algorithms. In *International Conference on Pattern Recognition, ICPR*, pages 313–317, 2004.

[70] L. R . Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In A. Waibel and K.-F. Lee, editors, *Readings in Speech Recognition*, pages 267–296. Kaufmann, San Mateo, 1990.

[71] Z. Rahmann, G. Woodell, and D. Jobson. A comparison of the multiscale retinex with other image enhancement techniques. In *Proceedings of IS&T 50th Anniversary Conference*, pages 19–23, May 1997.

[72] G. Rigoll, H. Breit, and F. Wallhoff. Robust tracking of persons in real-world scenarios using a statistical computer vision approach. *Image Vision Computing*, 22(7):571–582, 2004.

[73] Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariéthoz. Estimating the quality of face localization for face verification. In *IEEE International Conference on Image Processing, ICIP '04*, volume 1, pages 581– 584, Singapore, 2004.

[74] Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariéthoz. Measuring the performance of face localization systems. IDIAP-RR 53, IDIAP, 2005.

[75] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *Transactions on Pattern Analysis and Machine Intelligence*, 20(1), 1998.

[76] M. Sadeghi, J. Kittler, A. Kostin, and K. Messer. A comparative study of automatic face verification algorithms on the banca database. In *4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA*, pages 35–43, Guilford, UK, 2003.

[77] F. Samaria. *Face Recognition Using Hidden Markov Models*. PhD thesis, University of Cambridge, 1994.

[78] C. Sanderson. *Automatic Person Verification Using Speech and Face Information*. PhD thesis, Griffith University, Brisbane, Australia, 2002.

[79] C. Sanderson and S. Bengio. Extrapolating single view face models for multi-view recognition. In *Proc. Int. Conf. Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, pages 581–586, Melbourne, 2004.

[80] C. Sanderson and K. K. Paliwal. Fast features for face authentication under illumination direction change. *Pattern Recognition Letters*, 24(14):2409–2419, 2003.

[81] C. Sanderson and K.K. Paliwal. Polynomial features for robust face authentication. In *International Conference on Image Processing, Rochester, New York*, 2002.

[82] C. Sanderson and K.K. Paliwal. Identity verification using speech and face information. *Digital Signal Processing*, 14(5):449–480, 2004.

[83] J. Short, J. Kittler, and K. Messer. A comparison of photometric normalisation algorithms for face verification. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 254–259, 2004.

[84] F.K. Soong and A.E Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Trans. Acoustics, Speech and Signal Processing*, 36(6):871–879, 1988.

[85] S. Spors and R. Rabenstein. A real-time facetracker for color video. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[86] F.B. Tek. Face detection using learning networks. Master's thesis, The Middle East Technical University, Ankara, Turkey, 2002.

[87] H. L. Van Trees. *Detection, Estimation, and Modulation Theory: Radar-Sonar Signal Processing and Gaussian Signals in Noise*. Krieger Publishing Co., Inc., Melbourne, FL, USA, 1992.

[88] F. Tsalakanidou, S. Malasiotis, and M.G. Strintzis. Face localization and authentication using color and depth images. *IEEE Transactions on Image Processing*, 14(2):152–168, 2005.

[89] M. Turk and A. Pentland. Eigenface for recognition. *Journal of Cognitive Neuro-science*, 3(1):70–86, 1991.

[90] V. Vapnik. *Statistical Learning Theory*. Wiley, Lecture Notes in Economics and Mathematical Systems, volume 454, 1998.

[91] V. Vezhnevets. Method for localization of human faces in color-based face detectors and trackers, 2002.

[92] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features'. In *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 511–518, Seoul, 2001.

[93] J.L. Wayman. Digital signal processing in biometric identification: a review. In *Proc. IEEE Int. Conf. Image Processing (ICIP)*, volume 1, pages 37–40, 2002.

[94] K. Weber. *HMM Mixtures (HMM2) for Robust Speech Recognition*. PhD thesis, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland, 2003.

[95] 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, 1997.

[96] J.D. Woodward. Biometrics: Privacy's foe or privacy's friend? *Proceedings of the IEEE*, 85(9):1480–1492, 1997.

[97] M.-H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 24:34–58, 2002.

[98] K. Yow. *Automatic human face detection and localization*. PhD thesis, University of Cambridge, Department of Engineering, 1998.

[99] W. Zhao, R. Chellappa, and P. Phillips. Subspace linear discriminant analysis for face recognition. Technical Report CAR-TR-914, Center for Automation Research, University of Maryland, College Park, 1999.