Video person recognition strategies using head motion and facial appearance

Federico Matta

A doctoral dissertation submitted to:
University of Nice Sophia-Antipolis (UNSA)
in partial fulfilment of the requirements for the degree of:
DOCTOR OF PHILOSOPHY
Major subject: signal processing, image and vision

Approved by the following examining committee:

Supervisor:	Prof. Jean-Luc Dugelay
President of the jury:	Prof. Maurice Milgram
Examiner:	Prof. Alice Caplier
Examiner:	Prof. Carlo Regazzoni
Member:	Prof. Jean-François Bonastre
Member:	Eng. Lionel Martin

29th April 2008

Abstract

In this doctoral dissertation, we principally explore the use of the temporal information available in video sequences for person and gender recognition; in particular, we focus on the analysis of head and facial motion, and their potential application as biometric identifiers. We also investigate how to exploit as much video information as possible for the automatic recognition; more precisely, we examine the possibility of integrating the head and mouth motion information with facial appearance into a multimodal biometric system, and we study the extraction of novel spatio-temporal facial features for recognition.

We initially present a person recognition system that exploits the unconstrained head motion information, extracted by tracking a few facial landmarks in the image plane. In particular, we detail how each video sequence is firstly pre-processed by semi-automatically detecting the face, and then automatically tracking the facial landmarks over time using a *template matching* strategy. Then, we describe the geometrical normalisations of the extracted signals, the calculation of the feature vectors, and how these are successively used to estimate the client models through a *Gaussian mixture model (GMM)* approximation. In the end, we achieve person identification and verification by applying the probability theory and the *Bayesian decision rule* (also called *Bayesian inference*).

Afterwards, we propose a multimodal extension of our person recognition system; more precisely, we successfully integrate the head motion information with mouth motion and facial appearance, by taking advantage of a unified probabilistic framework. In fact, we develop a new temporal subsystem that has an extended feature space enriched by some additional mouth parameters; at the same time, we introduce a complementary spatial subsystem based on a probabilistic extension of the original *eigenface approach*. In the end, we implement an integration step to combine the similarity scores of the two parallel subsystems, using a suitable *opinion fusion* (or *score fusion*) strategy.

Finally, we investigate a practical method for extracting novel spatio-temporal facial features from video sequences, which are used to discriminate identity and gender. For this purpose we develop a recognition system called *tomofaces*, which applies the *temporal X-ray transformation* of a video sequence to summarise the facial motion and appearance information of a person into a single X-ray image. Then, we detail the linear projection from the X-ray image space to a low dimensional feature space, the estimation of the client models obtained by computing their cluster representatives, and the recognition of identity and gender through a *nearest neighbour classifier* using distances.

Résumé

Dans cette thèse, nous explorons principalement l'utilisation de l'information temporelle des séquences vidéo afin de l'appliquer à la reconnaissance de personne et de son genre; en particulier, nous nous concentrons sur l'analyse du mouvement de la tête et du visage ainsi que sur leurs applications potentielles comme éléments d'identification biométriques. De plus, nous cherchons à exploiter la majorité de l'information contenue dans la vidéo pour la reconnaissance automatique; plus précisément, nous regardons la possibilité d'intégrer dans un système biométrique multimodal l'information liée au mouvement de la tête et de la bouche avec celle de l'aspect du visage, et nous étudions l'extraction des nouveaux paramètres spatio-temporels pour la reconnaissance faciale.

Nous présentons d'abord un système de reconnaissance de la personne qui exploite l'information relative au mouvement spontané de la tête. Cette information est extraite par le suivi dans le plan image de certains éléments caractéristiques du visage. En particulier, nous détaillons la détection semi-automatique du visage dans chaque séquence vidéo, puis le suivi automatique dans le temps de certains éléments caractéristiques en utilisant une approche basée sur *l'appariement de bloques* (*template matching*). Ensuite, nous exposons les normalisations géométriques des signaux que nous avons obtenus, le calcul des vecteurs caractéristiques, et la façon dont ils sont utilisés pour estimer les modèles des clients, approximés avec des *modèles de mélange de gaussiennes*. Nous terminons par le module d'identification et vérification basé sur la théorie de probabilités et la *règle de décision bayésienne* (aussi appelée *inférence bayésienne*).

Nous proposons ensuite une extension multimodale de notre système de reconnaissance de la personne; plus précisément, nous intégrons à travers un cadre probabiliste unifié l'information sur le mouvement de la tête avec celle liée au mouvement de la bouche et à l'aspect du visage. A cet effet nous développons un nouveau sous-système temporel qui a un espace caractéristique étendu, lequel est enrichi par certains paramètres supplémentaires relatif au mouvement de la bouche; dans le même temps nous introduisons un sous-système spatial complémentaire au précédent, basé sur une extension probabiliste de l'*approche Eigenfaces* d'origine. Une étape finale d'intégration combine les scores de similarité des deux sous-systèmes parallèles, grâce à une stratégie appropriée de *fusion d'opinions*.

La dernière partie de la thèse nous avons voulu la consacrer à l'étude d'une methode pratique d'extraction de nouveaux paramètres spatio-temporels liés au visage à partir des séquences vidéo; le but est de distinguer l'identité et le genre de la personne. À cette fin nous introduisons un système de reconnaissance appelé *tomovisages* (*tomofaces*), qui utilise le *principe de la tomographie vidéo* pour résumer en une seule image l'information relative au mouvement et à l'aspect du visage d'une personne. Puis, nous détaillons la projection linéaire à partir de l'espace de l'image en rayons X à un espace caractéristique de dimension réduite, l'estimation des modèles des utilisateurs en calculant les représentants des clusters correspondants, et la reconnaissance de l'identité et du genre par le biais d'un *classificateur de plus proche voisin*, qui adopte des distances dans le sous-espace.

Riassunto

In questa tesi di dottorato esploriamo la possibilità di riconoscere l'identità e il sesso di una persona attraverso l'utilizzo dell'informazione temporale disponibile in alcune sequenze video, in particolare ci concentriamo sull'analisi del movimento della testa e del viso, nonché del loro potenziale utilizzo come identificatori biometrici. Esaminiamo inoltre la problematica relativa al fatto di sfruttare più informazione video possibile per effettuare il riconoscimento automatico della persona; più precisamente, analizziamo la possibilità di integrare in un sistema biometrico multimodale l'informazione relativa al movimento della testa e della bocca con quella dell'aspetto del viso, e studiamo il calcolo di nuovi parametri spazio-temporali che siano utilizzabili per il riconoscimento stesso.

In primo luogo presentiamo un sistema di riconoscimento biometrico della persona che sfrutti l'informazione legata al movimento naturale della testa, il quale è estratto seguendo la posizione nel piano immagine di alcuni elementi caratteristici del viso. In particolare descriviamo come in una sequenza video il volto venga dapprima individuato semiautomaticamente, e come poi alcuni suoi elementi caratteristici siano localizzati nel tempo tramite un algoritmo automatico di *messa in corrispondenza di modelli* (*template matching*) permettendo di seguirne la posizione. Spieghiamo quindi le normalizzazioni geometriche dei segnali che abbiamo ricavato, il calcolo dei vettori caratteristici, ed il modo in cui questi sono utilizzati per stimare i modelli degli utilizzatori, approssimandoli tramite delle *misture di distribuzioni gaussiane* (*Gaussian mixture models*). Alla fine otteniamo l'identificazione e la verifica dell'identità della persona applicando la teoria delle probabilità e la *regola di decisione o inferenza bayesiana*.

In seguito proponiamo un'estensione multimodale del nostro sistema di riconoscimento della persona; più precisamente, tramite un approccio probabilistico unificato, integriamo l'informazione sul movimento della testa con quelle relative al movimento della bocca e all'aspetto del viso. Infatti sviluppiamo un nuovo sottosistema temporale che possiede uno spazio caratteristico esteso, arricchito di alcuni parametri aggiuntivi legati al movimento della bocca; contemporaneamente, introduciamo un sottosistema spaziale complementare al precedente, basato su un'estensione probabilistica dell'*approccio Eigenfaces* originale. Alla fine implementiamo uno stadio di fusione, che metta insieme i valori di somiglianza dei due sottosistemi paralleli, attraverso un'appropriata strategia di *fusione delle opinioni*.

Infine investighiamo un metodo pratico per estrarre nuovi parametri spazio-temporali relativi al volto a partire da sequenze video, i quali sono utilizzati per distinguere l'identità ed il sesso della persona. A questo riguardo sviluppiamo un sistema di riconoscimento chiamato *tomovolti* (*tomofaces*), il quale utilizza la *tecnica della tomografia video* per riassumere in una sola immagine l'informazione relativa all'aspetto ed al movimento del volto di una persona. Poi descriviamo la proiezione lineare dallo spazio dell'immagine ai raggi X ad un spazio caratteristico di dimensione ridotta, la stima dei modelli degli utilizzatori attraverso il calcolo dei rappresentanti corrispondenti ad ogni cluster, ed il riconoscimento dell'identità e del genere attraverso un *classificatore al vicino più prossimo* (*nearest neighbour classifier*), che adopera le distanze nel sottospazio.

Acknowledgements

I am greatly indebted with my closest colleagues, Dr. Jihene Bannour and Ph.D. Usman M. Saeed, and my supervisor, Prof. Jean-Luc Dugelay. We have been profitably working together for a few years, sharing ideas and collaborating on research projects, and we mutually supported and encouraged during the hard times.

I am grateful to my past colleagues, Dr. Florent Perronnin, Dr. Mohammed Faouzi Benzeghiba, Dr. Luca Brayda and Prof. Christian Wellekens, for the productive and inspirational discussions that we shared together, which were of invaluable help for my research.

I would like to thank my fellow Ph. D. students and colleagues in Eurécom, for years of friendship, productive working environment, and much stimulating discussions: Eng. Caroline Mallauran, Dr. Emmanuel Garcia, Dr. Gwenaël Doërr, Prof. Benoit Huet, Prof. Nicholas Evans, Ph.D. Marco Paleari, Ph.D. Remi Trichet, Ph.D. Emilie Dumont, Dr. Joakim Jiten, Dr. Fabio Valente, Dr. Daniel Riccio, Ph.D. Slim Trabelsi, Ph.D Antony Schutz, Dr. Maxime Guillaud... and all those who offered their friendly and generous help during these four years.

I would like to show my appreciation to each one of the jury members, for having dedicated a part of their time reading and evaluating the research that I did for this thesis. My thanks also go to the Similar Network of Excellence and ST Microelectronics, for having funded my research activities and several pleasurable trips to conferences and workshops.

Most of all, I would like to express my deepest gratitude to my beloved Géraldine and my dear family. Without their sacrifice, support and encouragement in difficult moments, there would never have been any chance for this thesis to happen.

List of contents

List of figures

List of tables

List of abbreviations

AAM: active appearance model
CAR: correct acceptance rate
CIR: correct identification rate
CMS: cumulative (correct) match score
CRR: correct rejection rate
DCT: discrete cosine transform
DET: detection error trade-off
DVT: discrete video tomography
EER: equal error rate
EBGM: elastic bunch graph matching
EGM: elastic graph matching
EM: expectation-maximisation
FAR: false acceptance rate
FLD: Fisher's linear discriminant
FMR: false match rate
FNMR: false non-match rate
FRR: false rejection rate
FTCR: failure to capture rate
FTER: failure to enrol rate
GMM: Gaussian mixture model
GUI: graphical user interface
HDRT: hierarchical discriminative regression tree
HMM: hidden Markov model
KLT: Karhunen-Loeve transform
LDA: linear discriminant analysis
MAP: maximum a posteriori
PCA: principal component analysis

PDF: probability density function

RBFNN: radial basis function neural network

ROC: receiver operating characteristic

SIS: sequential importance sampling

TSSSM: time series state space model

Chapter I. Introduction

I.A. Motivation

There are numerous reasons that motivate our interest in studying novel person recognition approaches based on facial video sequences.

First of all, over the last few decades biometric person recognition has gained a vast interest in the scientific community and benefited by increasing investments in the most technologically advanced countries. In fact, the expansion of electronic commerce and finance, the need for accessing restricted areas and resources, and the development of worldwide travel have required simple and reliable person recognition tools. Furthermore, after the terrorist attacks of 9th September 2001, government agencies and corporations have been investing in biometric technology more than ever, in order to enforce public security and access to sensitive facilities.

Then, the human face is a fundamental element in our social lives because it provides a bewildering variety of important signals: for example, its bearer's identity, gender, age, emotion and interest. For this reason, human face recognition has been a central topic in the field of person recognition, and this biometric has demonstrated some valuable properties: it is non intrusive, easy to collect, and well-accepted by the public.

Afterwards, person recognition using facial video information has some advantages over image-based recognition. First of all, video frames can provide a huge amount of data compared to single pictures, and more robust and stable recognition can be achieved by integrating information and decisions from previous frames. Then, in addition to the physiological information already present in images, also the temporal one becomes available and can be exploited to improve the recognition task; consequently, nowadays researches have the possibility to analyse not only facial appearance but also head and facial motion, and human face starts to be considered as a hybrid biometric identifier, rather than only a physiological one. Finally, video data allows learning and updating the models over time.

In our research, we principally explore the use of the temporal information available in video sequences for person and gender recognition; in particular, we focus on the analysis of head and facial motion, and their potential application as biometric identifiers. We motivate our choice with the following considerations.

Currently, the research on person recognition using facial video information has been mostly focused on developing straightforward extensions of image-based approaches, which exploit only the spatial information in video sequences; furthermore, most of temporal strategies take only advantage of the evolution of facial appearance over time. Hence, the use of head and mouth motion for person recognition is still a largely unexplored topic. On the contrary, we believe that the way an individual moves his head or his face is somewhat characteristic, and that the dynamic patterns could be used to discriminate people. We are supported in this claim by the study of Knight and Johnston [41], which reveals that under non-optimal image conditions (like negative images) "moving faces are significantly better recognised than still faces".

Finally, in our research we also investigate how to exploit as much video information as possible for recognition; more precisely, we focus on the possibility of integrating the head and mouth motion information with facial appearance into a multimodal biometric system, and we study the extraction of novel spatio-temporal facial features for recognition. Again, there are some reasons that motivate our choice.

Until now, it is a common trend in literature to exploit only a part of the biometric information embedded in video sequences, mainly the physiological one related to facial appearance. Though, video data does not provide only abundant spatial information but also the temporal one, and as far as we know it has never been proposed a hybrid person recognition system, using the physiological and behavioural aspects of the face at the same time. In contrast, the integration of multiple sources of information typically has numerous advantages for biometric recognition systems: it can increase the accuracy of the systems, by exploiting complementary information, it can augment their reliability, by taking advantage of redundant and richer information that can compensate the individual weaknesses, and it can reduce their cost, by exploiting several cheap sensors. Hence, we considered taking advantage not only of the temporal facial information present in video sequences, but also of facial appearance, which is one of the traditional biometric identifiers for person recognition, and it has been largely studied during the last decades.

I.B. Original contributions

In this section we underline the original contributions of this thesis.

The major contribution in Chapter IV is the exploration of the unconstrained 2D head motion information for person recognition. In particular, we calculate the feature vectors by tracking a few facial landmarks in the image plane, we estimate the user model through a *Gaussian mixture model approximation (GMM)*, and we achieve identification and verification by applying the probability theory and the *Bayesian decision rule* (also called *Bayesian inference*). In addition, one minor contribution is the introduction of two novel similarity measures for person recognition using video data: the *video log-posterior probability* for the identification task, and the *video log-posterior probability ratio* for the verification one. Then, other minor contributions are: the experimental analysis of the effect of tracking noise on the discriminatory power of the head motion information, and the application of our biometric system to a gender recognition scenario.

Afterwards, the major contribution in Chapter V is the study of a multimodal person and gender recognition system, which integrates unconstrained head motion with mouth motion and facial appearance, in a unified probabilistic framework. In fact, we develop a new temporal subsystem that has an extended feature space enriched by some additional mouth parameters, and a complementary spatial subsystem based on a probabilistic extension of the original eigenface approach; then, we introduce an integration step to combine the similarity scores of the two parallel subsystems. In addition, one minor contribution is the extension of the eigenface technique to a probabilistic framework, by adopting a *Gaussian mixture model (GMMs)* approximation to represent the biometric features of each client, and *Bayesian inference* to calculate the similarity between tests and models. Then, other minor contributions are: the development of a *weighted summation fusion* strategy adapted to our probabilistic framework, and its probabilistic interpretation.

Finally, the main original contribution in Chapter VI is the exploration of novel spatio-temporal facial features for person and gender recognition. In particular, we propose a biometric system called *tomofaces*, which applies the temporal X-ray transformation of a video sequence to summarise the facial motion and appearance information of a person into a single X-ray image.

We conclude this section by underlying the importance of the video database of Italian TV speakers that we have been collecting for some months (Chapter VIII): without the manual work and this precious data, there would never have been any chance for this research to happen.

I.C. Outline

This doctoral dissertation is organised as follows:

- In Chapter II we provide an introduction to the discipline of biometrics and its evolution towards multi-biometrics.
- In Chapter III we review the literature on person recognition using facial video information.
- In Chapter IV we present a novel person recognition system that exploits the unconstrained head motion information, extracted by tracking a few facial landmarks in the image plane.
- In Chapter V we propose a multimodal extension of our person recognition system; in particular, we successfully integrate the head motion information with mouth motion and facial appearance, by taking advantage of a unified probabilistic framework.
- In Chapter VI we investigate a practical method for extracting novel spatio-temporal facial features from video sequences, which are used to discriminate identity and gender in a recognition system called *tomofaces*.
- In Chapter VII we conclude this dissertation with a summary and some comments on future perspectives.

- In Chapter VIII we detail our database of Italian TV speakers, which is used in the experiments to evaluate the performance of the recognition systems.

In the following chapters there are some minor intentional repetitions and redundant references, because we intend to provide as much consistent and complete information possible, and to generate almost self sufficient chapters.

Chapter II. Generalities on biometrics

II.A. Introduction

Over the last few decades biometric person recognition [20][37][38][52][67][68] has gained a vast interest in the scientific community and benefited by increasing investments in the most technologically advanced countries. In fact, the expansion of electronic commerce and finance, the need for accessing restricted areas and resources, and the development of worldwide travel have required simple and reliable person recognition tools. Furthermore, after the terrorist attacks of 9th September 2001, government agencies and corporations have been investing in biometric technology more than ever, in order to enforce public security and access to sensitive facilities. This growing interest in biometrics can be evaluated by considering the total revenue market of the last decade [52], which is exponentially increasing as shown in Figure 1.

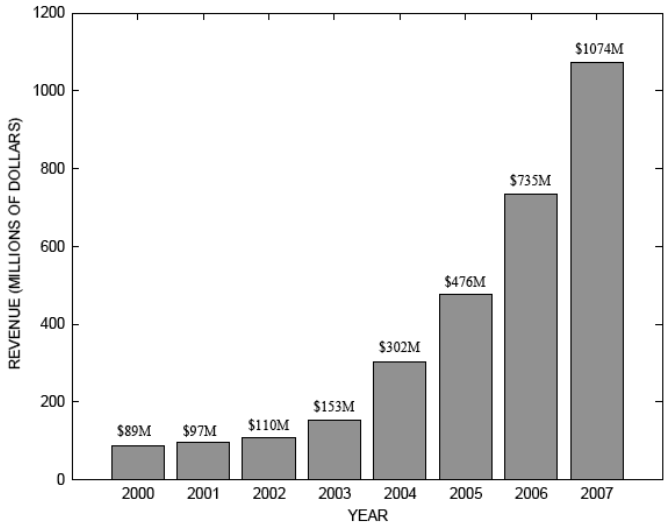

Figure 1: total biometric revenue market: 2000-2007 [52].

II.B. Definitions and properties

Traditionally, verification systems were of only two types: knowledge-based and token-based. *Knowledge-based systems* make use of something someone knows, like a password or a PIN, while *token-based systems* take advantage of something someone owns, like a badge or a key. These verification approaches are inherently insecure, because knowledge can be forgotten or guessed and a token can be stolen or lost. On the other hand, identification of criminals has exploited biometric characteristics since the beginning of XX century, but the need of manual inspection and comparison of numerous fingerprints moderated its use and diffusion. Nowadays, the advent of a new generation of systems, which are based on the automatic recognition of biometric traits, has been able to provide higher security and to extend the potential domains of application.

II.B.1. Biometric identifier

A *biometric* or *biometric identifier* is originally defined as an objective measurement of a physical characteristic of an individual which, when captured in a database, can be used to verify the identity or check against other entries in the database. The term *biometric* has a Greek origin: it is composed by *bios* (life) and *metron* (measure) and means "measure of life".

A biometric identifier should ideally possess the following properties to be exploited in a recognition system:

- *Universal*: each user should have it.
- *Permanent*: it should not vary over time.
- *Distinctive*: inter-class variability should be as large as possible, which means that captured patterns from distinct users should be as different as possible.
- *Robust*: intra-class variability should be as small as possible, which means that different captured patterns from the same user should be as close as possible.
- *Collectable*: it should be easy to collect.
- *Accessible*: it should be easy to present to the sensor.
- *Acceptable*: to be well accepted by the public a biometric trait should be perceived as non obtrusive and non intrusive.
- *Hard to circumvent*: it should be difficult to alter or reproduce by an impostor who wants to fool the system.

II.B.2. Biometric recognition

Biometric recognition can be considered as: the automatic person identification or identity verification of an individual, based on *physiological and/or behavioural biometric identifiers*. It is not possible to classify all biometrics with a clear distinction between physiological and behavioural traits; in some cases a biometric is a combination of both elements, so we will introduce the third class of hybrid biometrics.

The most important *physiological biometrics* are the following:

- *Fingerprint* [36][55]: a fingerprint is the pattern of ridges and valleys on the surface of a fingertip. Fingerprints of identical twins are different and so are the prints on each finger of the same person; for these reasons fingerprints possess good discriminatory power and were the first biometric identifiers to be used in real recognition systems. In 2002, fingerprint recognition was the most important technology, with the biggest market share [52].
- *Iris* [19]: the iris is the annular region of the eye bounded by the pupil and the sclera on either side. The complex iris texture carries very distinctive information: each iris is different and irises of identical twins are different. Iris recognition is a very promising biometric, in terms of accuracy and speed, but it requires considerable user cooperation.
- *DNA*: Deoxyribonucleic acid (DNA) is a nucleic acid that contains the genetic instructions for the development and functioning of all living organisms. It represents the one dimensional ultimate unique code for one's individuality, except for the fact that identical twins have identical DNA patterns. However, its practical application has been limited due to problematical chemical analyses and privacy issues.
- *Retina*: the retina is a thin layer of neural cells that lines the back of the eyeball; the retinal vasculature is rich in structure and it is supposed to be characteristic of each individual and each eye. Although this biometric is considered as one of the most secure, the intrusiveness and the need for user cooperation are its major drawbacks.
- *Hand and finger geometry* [79]: hand geometry recognition systems are based on a number of measurements taken from the human hand and fingers; the geometry of the hand is an inexpensive technique, well accepted and easy to collect, but not one of the most discriminating.
- *Ear* [72][94]: it has been suggested that the shape of the ear and the structure of the cartilaginous tissue of the pinna are distinctive. It is a relatively new biometric and its accuracy and scalability are not well known yet.
- *Palm print* [95]: the palms of the human hands contain pattern of ridges and valleys much like the fingerprints; human palms also contain additional distinctive features such as principal lines and wrinkles. Palm prints can be considered as an evolution of fingerprints: they are more accurate but acquisition sensors are more expensive.
- *Infrared thermo grams of body parts (face, hand, hand veins)* [42]: the pattern of heat radiated by human body is a characteristic of an individual and can be captured by an infrared camera. It is a well accepted biometric and can be exploited for covert applications, but the acquisition process is very sensible to heat emanating surfaces (room heaters, vehicles...).

- *Odour*: it is known that each object exudes an odour that is characteristic of its chemical composition and that a component of the odour emitted by a human body is distinctive to a particular individual. On the other hand, it is not known how the use of deodorants and the chemical composition of the surrounding environment affect its performance.

Then, a few examples of *behavioural biometrics* are:

- *Gait* [62][63]: gait is the particular way one walks and is a complex spatio-temporal biometric. Gait it is generally not as distinctive and may not remain invariant over time, but it is well accepted by the population.

- *Keystroke dynamics* [59]: it is hypothesized that each person types on a keyboard in a characteristic way. This behavioural biometric is not expected to be unique and one may presume to observe large variations in typical typing patterns.

Finally, some examples of *hybrid biometrics*:

- *Voice* [27]: the acoustic patterns used in speaker recognition reflect both anatomy (size and shape of the throat and mouth) and behavioural patterns (voice pitch and prosody). Voice it is generally not as discriminating; it suffers the presence of background noise and may not remain invariant over time.

- *Signature* [70]: the way a person signs his name is known to be a characteristic of that individual; however, signatures of some people vary substantially. The shape of the signature is typically a physiological pattern, while the speed and the inclination during the signature are behavioural ones. Even if this biometric is well accepted and widespread, it is not robust and can be reproduced by professional forgers.

- *Face* [12][67][96]: face recognition is a non intrusive biometric and probably the most common biometric characteristic used by humans to recognise people. Facial appearance is a physiological trait, whereas head and facial motion are behavioural ones. Face is a very promising biometric, easy to collect and well accepted, but at the moment its accuracy is quite low, due to the capricious variations of the facial data acquired with cameras.

We have seen that a number of biometrics exist and are in use in various applications (see Figure 2). It is important to notice that no biometric is "optimal", presenting "ideal" properties and outperforming recognition results, but everyone has various strengths and weaknesses. That is the reason why person recognition is such a big research domain with numerous alternative approaches that must be carefully evaluated, depending on the application in question.

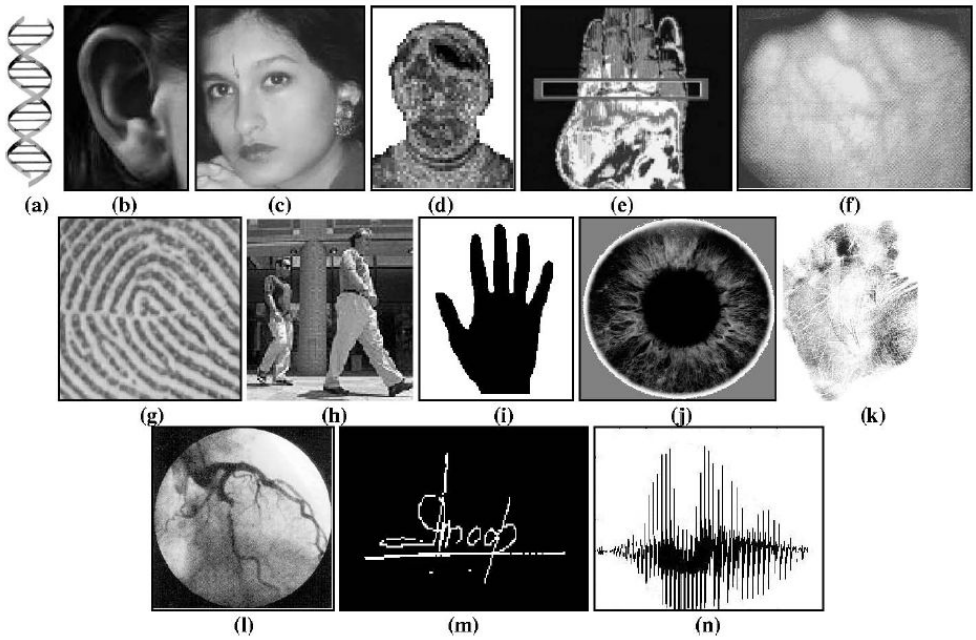

Figure 2: examples ofbiometric identifiers [37]: (a) DNA, (b) ear, (c) face, (d) facial thermo gram, (e) hand thermo gram, (f) hand vein, (g) fingerprint, (h) gait, (i) hand geometry, (j) iris, (k) palm print, (l) retina, (m) signature and (n) voice.

II.B.3. Applications: properties and examples

Biometric recognition systems can be classified by considering the following properties ofpotential applications:

- *Cooperative/non-cooperative*: reflects the necessity for a user to actively cooperate during the recognition process, or not.
- *Overt/covert*: if the user is aware or not of the measurement, during the capture ofhis biometric identifier.
- *Habituated/non-habituated*: indicates the frequency of interaction of a user with the recognition system.
- *Attended/non-attended*: expresses whether the capture or the recognition process in general has to be supervised, observed or guided by an operator.
- *Standard/non-standard environment*: considers if the conditions ofoperation of the device are familiar and usual, or not.
- *Public/private*: states whether the clients of the system are external customers or the device operators.
- *Open/closed*: reflects the requirement of exchanging data with peripheral applications, or not.

The main domains of application for biometric recognition systems are the following:

- *Access control*: we can divide access control applications into physical and virtual categories. Physical access control regulates the entrance to physical locations, like buildings and offices; virtual access control regulates the usage of resources or services, like networks, computers, cellular phones, PDAs, etc.
- *Commercial transaction authentication*: these applications are related to banking and business activities and need to accurately verify the identity of clients. Examples are automatic teller machines (ATM), electronic fund transfers, credit card transactions, internet commerce (e-commerce), etc.
- *Citizen identification*: mainly used by government agencies to enforce security and law. Examples are identification of criminals and corpses, terrorist detection, parent determination for abandoned children, etc. Automatic biometric recognition plays a major role in the increase of robustness and efficiency of traditional methods, based on national ID cards, drivers' licences, passports or manual inspection of fingerprints and DNA patterns.
- *Personalisation*: biometrics can also be used by marketing companies to easily identify customers and provide personalised services. This is a promising business for recognition technology, yet to be fully developed.

II.C. Operational modes

A biometric recognition system has two main operational modes: verification (or authentication) and identification. Besides, in this dissertation we use the generic term *recognition* when we do not want to refer to any particular operational mode, and the expression *full recognition* when we want to consider both of them.

II.C.1. Verification (or authentication)

In a *verification (or authentication)* scenario, a user presents his biometric identifier to a sensor and claims an identity; after that, the recognition system verifies his claim and decides to accept it or reject it. The authentication process is done through a one-to-one comparison between the unidentified biometric pattern and the claimed model pattern stored in the system. An *open-set* is generally assumed, which means that the input sample may correspond to an individual who is not enrolled in the system. Biometric verification is an alternative solution to knowledge-based and token-based traditional systems in *positive recognition* applications, those that prevent multiple individuals by using the same identity. In fact, if multiple users claim the same identity, the system will authenticate only one of them (the *true client* or simply the *client*) and will reject the others (the *impostors*).

We can formally describe the verification problem as follows. If we consider an input feature vector, $\mathbf{x}$, and a claimed identity, ϖ, then a verification system must determine if the pair $(\varpi, \mathbf{x})$ belongs to class k_ϖ or $\bar{k}_\varpi$, where k_ϖ is the client class (claim is true) and $\bar{k}_\varpi$ is the impostor one (also called the *alternative hypothesis*, for which the claim is false). If we represent the stored model pattern for user k_ϖ as $\boldsymbol{\Theta}_{k_\varpi}$, the *decision rule* is the following:

$$(\varpi, \mathbf{x}) \in \begin{cases} k_\varpi & \text{if} \quad S^{(VER)}\left(\mathbf{x}, \boldsymbol{\Theta}_{k_\varpi}\right) \geq \theta \\ \bar{k}_\varpi & \text{otherwise} \end{cases}$$

where θ is a predefined *threshold*, and $S^{(VER)}\left(\mathbf{x}, \boldsymbol{\Theta}_{k_\varpi}\right)$ is the *similarity (or matching) score* between the test $\mathbf{x}$ and the model $\boldsymbol{\Theta}_{k_\varpi}$. It is important to notice that, in practical applications, independent biometric measurements of the same individual are somewhat different (even slightly); for this reason, it is not possible to obtain a *perfect match* and a threshold value must be introduced in the decision rule.

II.C.2. Identification

In an *identification* scenario, a user presents his biometric identifier to the sensor and makes no claim on his identity; then the system performs a search through the database to find the most likely identity. In this case, the unidentified pattern is matched up to all the model patterns present in the system, in a one-to-many comparison. A *closed-set* is generally assumed, which means that the input sample belongs to an individual who is enrolled in the system. Biometric identification is the only solution for *negative recognition* applications, those that prevent a single individual from using multiple identities. In fact, if a user aims to exploit diverse identities or to hide his real one, then a biometric identification system can determine its unique authentic identity in each occasion. Moreover, this operational mode can be applied in *positive recognition* applications as well (refer to section II.C.1 for more details): in this case it is more convenient than verification, because the user need not make a claim, but it can be more complex and less robust, due to a more difficult one-to-many scenario and there being less information available (the claim).

Similarly to the verification case, we can formally describe the identification problem as follows. If we consider an input feature vector, $\mathbf{x}$, then an identification system must determine the identity of the user, $k \in \mathrm{N}$, where $\{k \mid k = 1, \ldots, K\}$ are the clients enrolled in the system and $k = K + 1$ represents the reject case. If we denote the stored model pattern for identity k as $\boldsymbol{\Theta}_k$, and with θ the predefined *threshold*, the *decision rule* is the following:

$$\mathbf{x} \in \begin{cases} k & \text{if} \quad \max_k \left\{ S^{(ID)}\left(\mathbf{x}, \boldsymbol{\Theta}_k\right) \right\} \geq \theta \quad k = 1, \ldots, K \\ K+1 & \text{otherwise} \end{cases}$$

where θ is a predefined *threshold* and $S^{(ID)}\left(\mathbf{x}, \boldsymbol{\Theta}_k\right)$ is the *similarity (or matching) score* between the test $\mathbf{x}$ and the model $\boldsymbol{\Theta}_k$.

II.D. Architecture

A typical automatic recognition system is composed by two mandatory modules, for the enrolment and recognition tasks, and can optionally have a third one, for the adaptation of user models. An overview of the common architecture of a biometric system is illustrated in Figure 3: the main elements in the picture are described in the following sections.

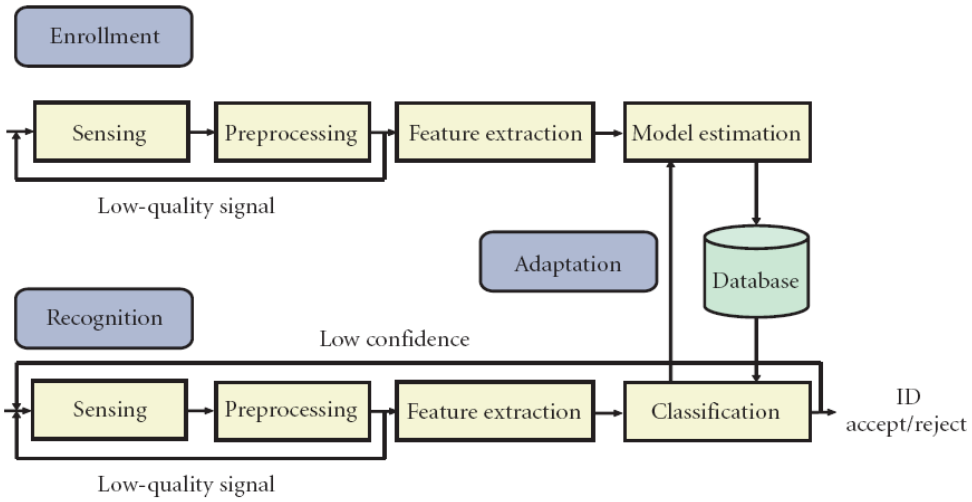

Figure 3: common architecture of a biometric system.

II.D.1. Enrolment

The *enrolment* module is required for the registration of new users in the recognition system. First of all, a client presents his biometric identifier to a *sensing* device, which captures the signal and represents it in a digital form. In general, a *pre-processing* step enhances and normalises the acquired information, in order to achieve a better representation for the extraction of discriminative features. For example, pre-processing of image data usually consists of: object detection and segmentation, photometric compensation, noise reduction and geometric normalisations. Next, the quality of the pre-processed signal is checked to estimate if reliable features could be extracted from it or not; in the latter case, the sample is discarded and a new acquisition is needed. This situation is often referred to as *failure to enrol*, which is detailed in Section II.E.3. After that, the *feature extraction* step transforms the signal, trying to isolate the significant features that characterise the individual and to discard the irrelevant and redundant information. In most cases, the feature extractor computes a reduced representation of the acquired pre-processed signal, which can be seen as a non reversible compression technique. In the end, the enrolment module estimates a model of the client, representing the potential range of biometric features for that user, and stores it in its internal database (*model estimation* step).

II.D.2. Recognition

The *recognition* module verifies and/or identifies users, by comparing new acquisitions of the biometric identifier with the models of the clients stored in the database. When a user needs to be recognised, he presents his biometric identifier to a captor and then the sensing, pre-processing and feature extraction phases are exactly the same as in the enrolment module. Afterwards, the *classification* step compares the discriminative features of the unknown user with the model patterns retrieved from the database, and computes the similarity score for each possible match (one or more, depending on the operational mode). The final decision of the system is determined by the operational mode in question: in a verification task the user claim is confirmed or rejected, while in an identification one the user identity is established.

II.D.3. Adaptation

The *adaptation* module is optional and it is useful for updating the user models stored in the database. Most of the biometrics are not permanent and vary over time, especially the behavioural and hybrid ones like: gait, voice, signature and face. Consequently, the actual biometric identifier of a client gradually differs from the original acquisitions stored at his first enrolment, and may eventually lead to a degradation in performance. Therefore, the adaptation module is meant to cope with those variations and to provide an updated representation of each user; in general, it progressively adds new acquisitions to those already stored, and adapts the model estimate using this new data.

II.E. Performance evaluation

Automatic biometric person recognition is a challenging issue. In fact, two samples of the same biometric characteristic from the same individual are not exactly identical; this effect can be caused by multiple reasons: for example, variable acquisition conditions, changes in the user's physiological or behavioural biometric identifier, diverse ambient conditions, dissimilar user interaction with the sensing device, etc. It is then fundamental to evaluate the performance of a biometric system and to understand its strengths and limitations: this section is dedicated to performance evaluation measures and their uncertainties.

It is possible to consider three different scenarios for assessing the recognition capability of a biometric system. These are the following [56]:

1. *Technology evaluation*: it is largely the most common in the scientific community. The goal of a technology evaluation is to compare competing algorithms for a single technology. Testing of all algorithms is carried out using offline processing of a standard database, collected by a "universal" sensor; because the database is fixed, the results of technology tests are repeatable.

2. *Scenario evaluation*: the goal of scenario testing is to determine the overall system performance in a prototype or simulated application: testing is carried out on a complete system, in an environment that models a real-world target application. Each tested system will have its own acquisition sensor and so will receive slightly different data; consequently, care will be required that data collection across different systems is in the same environment with the same population. Depending on the storage capabilities of each device, testing might be a combination of offline and online comparisons. Scenario evaluation has the advantage of taking into account not only the technology, but also the human-machine interaction, with relative problems and errors. Test results will be repeatable only to the extent that the modelled scenario can be carefully controlled.
3. *Operational evaluation*: the goal of operational testing is to determine the performance of a complete biometric system in a specific application environment with a specific target population. Depending upon storage capabilities of the tested device, offline testing might not be possible; moreover, due to unknown and undocumented differences between working environments, in general operational test results are not repeatable.

II.E.1. Measures for verification (or authentication)

A system operating in a verification (or authentication) mode can make two major types of decision errors:

- *False rejection* (or *Type-I error*): occurs when a *client* – a person who makes a true identity claim – is erroneously rejected.
- *False acceptance* (or *Type-II error*): occurs when an *impostor* – a person who makes a false identity claim – is erroneously accepted.

It is then possible to define the following four decision error measures:

- *False rejection rate (FRR)* (or *miss*): the expected proportion of transactions with true claims incorrectly denied.
- *False acceptance rate (FAR)* (or *false alarm*): the expected proportion of transactions with false claims incorrectly confirmed.
- *Correct acceptance rate (CAR)*: the complementary measure to the FRR, and represents the expected proportion of transactions with true claims correctly confirmed. Mathematically: $\eta_\theta^{(CAR)} \equiv 1 - \xi_\theta^{(FRR)}$ for $\forall \theta$.
- *Correct rejection rate (CRR)*: it is the complementary measure of the FAR, and represents the expected proportion of transactions with false claims correctly denied. Mathematically: $\eta_\theta^{(CRR)} \equiv 1 - \xi_\theta^{(FAR)}$ for $\forall \theta$.

Clearly, only two of these four measures are needed to characterise the performance capability of a verification system: from now on, we will mostly use FRR and FAR.

FRR, FAR, CAR and CRR are performance measures from detection/recognition theory and are directly related to recall and precision, which are information retrieval quantities [84]. The *recall* value, R, is defined as the proportion of all the material in the database which is of relevance, while the *precision* value, P, represents the proportion of retrieved material which is relevant. If we consider a uniform *richness*, P_t, which is the probability of appearance of each model in the database, then we can obtain the following relations:

$$R_\theta = 1 - \xi_\theta^{(FRR)} = \eta_\theta^{(CAR)}$$

$$P_\theta = \frac{P_t * \left(1 - \xi_\theta^{(FRR)}\right)}{P_t * \left(1 - \xi_\theta^{(FRR)}\right) + \left(1 - P_t\right) * \xi_\theta^{(FAR)}} = \frac{P_t * \eta_\theta^{(CAR)}}{P_t * \eta_\theta^{(CAR)} + \left(1 - P_t\right) * \xi_\theta^{(FAR)}}$$

It is important to notice that FRR and FAR are decision error measures and are defined over *transactions*, an attempt by a user to be authenticated by submitting one or more biometric samples, as allowed by the system decision policy. To avoid ambiguity with systems allowing multiple attempts or having multiple models per client, we define the following matching error measures, which consider a *single comparison* of a submitted sample against a single enrolled model:

- *False non-match rate (FNMR)*: it is the equivalent of the FRR for comparisons: it is the expected proportion of comparisons erroneously not matched. In a system where an individual transaction implies a single comparison, FRR and FNMR are identical.
- *False match rate (FMR)*: it is the equivalent of the FAR for comparisons: it is the expected proportion of comparisons erroneously matched. In a system where an individual transaction implies a single comparison, FAR and FMR are identical.

Figure 4 shows an example of client (blue curve) and impostor (red curve) distributions of normalised similarity scores; a client similarity score is calculated by matching a test pattern of a client with its model pattern, while an impostor similarity score is computed by matching a test pattern of a user (the impostor) with a different model pattern (the claimed client).

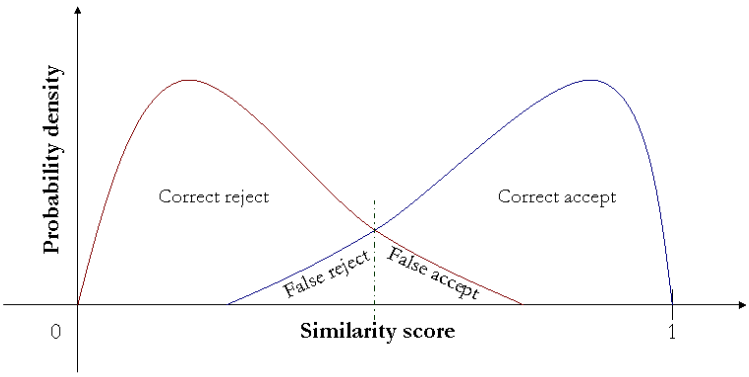

Figure 4: example ofclient (blue curve) and impostor (red curve) distributions of normalised similarity scores.

Ideally, these two distributions should be disjointed and the decision rule between clients and impostors would be immediate and flawless. Unfortunately in real cases these two distributions overlap, and a threshold (green dotted line) is needed to delimit the acceptance and rejection decision regions; if the similarity score is lower than the threshold value then the claim is considered false (rejection), while if it is higher then the claim is considered true (acceptance). Therefore, the overlapping of client and impostor score distributions and the choice of a decision threshold create two regions of errors for false rejects and false accepts. When the threshold value is increased, the system becomes more secure with higher FRR and lower FAR; on the other hand if the threshold value is decreased, the system becomes more convenient for the user with lower FRR and higher FAR. The choice of a threshold value should be made carefully, by evaluating the real context and requirements of the application in question, in order to find the best trade-off between security and user convenience.

A practical way for presenting the performance of a verification system is the *receiver operating characteristic (ROC)* curve as shown in Figure 5, in which FARs are plotted as a function of FRRs (or CARs). For drawing this graph, it is necessary to compute several pairs of FRRs and FARs at various threshold values; by eventually spanning the whole space of thresholds, $\theta \in [0,1]$, it is possible to obtain a full overview on the performance of a verification system, from low to high security configurations. Moreover, ROC curves are threshold independent, allowing performance comparison of different systems under similar conditions, or of a single system under differing conditions.

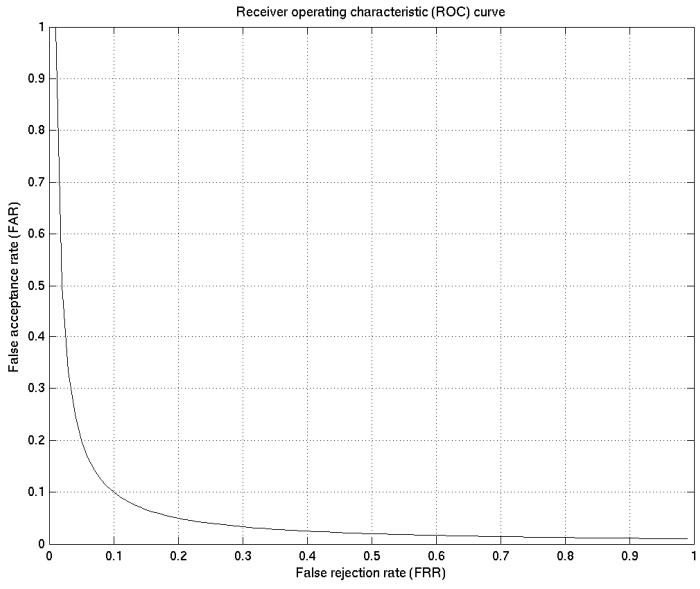

Figure 5: example of a receiver operating characteristic (ROC) curve.

Even if it is preferred to present verification results by plotting the ROC curve, some authors just report a single error measure; it is the *equal error rate (EER)*, a precise point on the ROC curve at which FRR and FAR are equal: $\xi_{\theta_{EER}}^{(FRR)} \equiv \xi_{\theta_{EER}}^{(FAR)}$.

Finally, there exists a modified ROC curve known as *detection error trade-off (DET)* curve. A DET curve plots error rates on both axes, like the ROC one, but the graph is drawn on a logarithmic scale (on both axes); this spreads out the plot and facilitates to distinguish and compare different well-performing technologies.

II.E.2. Measures for identification

A system operating in an identification mode makes an error when the identity matched with the input pattern is not the right one: the *correct identification rate (CIR)* is defined as the proportion (or percentage) of identification requests correctly answered. In this dissertation, as in the large majority of the scientific literature, we will compute the CIR by considering only the highest similarity score in each test (also called the *best match*).

For a better insight on the recognition capabilities of a system, it is possible to sort the similarity scores of each test and consider the ranking of the *correct match*, which is the score computed with the model of the input client. Ideally, in a flawless system the correct match is always the highest similarity value and $\eta^{(CIR)} = 1$; however in real cases the correct match can be lower than the top score and it may be interesting to consider its location.

For this purpose, we introduce the *cumulative (correct) match score (CMS)*, that is defined as the proportion (or percentage) of identification requests for which the correct match is among the highest M values. Clearly, when $M = 1$ we are just considering the top matches and $\eta_1^{(CMS)} \equiv \eta^{(CIR)}$. Figure 6 shows an example of CMSs plotted as a function of the M best similarity values retained: it is a monotone increasing curve which, at some point, clips at 1.

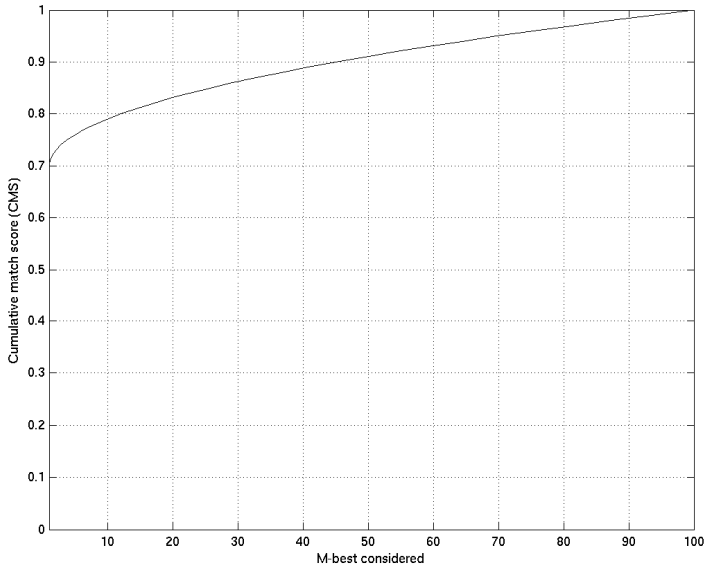

Figure 6: example ofcumulative match scores (CMSs) plotted as a function ofthe M-best similarity values retained.

In this operational mode, an unidentified pattern is matched up to all the model patterns present in the system, in a one-to-many comparison, and for very large databases this could be computationally demanding. Depending on the properties of the biometric in question, it may be possible to partition the whole database into smaller sub-datasets, and cleverly reduce the number of comparisons needed for each test. For example, if we consider a recognition system using fingerprints, it may be possible to partition the dataset into a few classes, based on the global pattern at the centre of the fingerprint: the arch, the loop, the whorl, etc. During the recognition phase, the test pattern is first used to find out the class to which it belongs, and then it is matched up with the models of that class. Ideally, the system should always associate the input pattern with the correct class, but in practical cases it may select an erroneous partition, causing a *binning error*.

In an identification system exploiting a partitioning strategy, we can compute two more measures of performance [56]: the *binning error rate* (or *retrieval error rate*) is defined as the expected number of model patterns wrongly discarded due to a binning error; the *penetration rate* is defined as the expected proportion of model patterns to be searched under the rule that the search proceeds through the entire partition, regardless of whether a match is found. In general, it is desirable to have the highest number of sub-datasets in order to reduce as much as possible the amount of comparisons; consequently, the highest is the number of partitions, the lower the penetration rate and the higher the binning error rate.

II.E.3. Other measures

For scenario and operational evaluations, two more measures of performance can be considered: failure to capture rate and failure to enrol rate. A recognition system with an automatic capture device may accidentally be unable to sense the biometric identifier and properly represent it in digital form. For this reason it can be interesting to measure the *failure to capture rate (FTCR)*, which is defined as the expected proportion of failures to capture a sample. Moreover, regardless of the universality property of the biometric identifier in question, it may happen that in practical applications some users can not be successfully enrolled and recognised with that identifier. For example, fingerprints of a small part of the population may be unsuitable for automatic recognition, because of genetic factors, aging, environmental or occupational reasons (manual workers may have a large number of cuts and bruises). The *failure to enrol rate (FTER)* is then defined as the expected proportion of the population not able to enrol in the recognition system. It is important to notice that the FTER is closely related to the quality checker module and can affect the global performance of a system; in fact, the higher is the FTER, the better is the quality of the database and the lower are the recognition error rates (FRR and FAR for example).

II.E.4. A glance on testing errors and uncertainty of estimates

It is out of the scope of this dissertation to provide a detailed analysis on testing errors and uncertainty of estimates; in this section we discuss only a few main issues on these topics.

Typically, the experiments run for assessing the performance of a recognition system can be affected by two kinds of errors: systematic errors and random errors.

Systematic errors are those due to bias in the test procedure. It may happen that a few categories of the population are over or under-represented in the database of use. A solution to reduce this bias is to carry out experiments on as much varied databases as possible. Another potential bias may arise when parameter tuning, client enrolment and performance assessment are done using the same data set. This choice produces an experimental situation over fitted to the data in question, not robust to changes in the operational environment and surely overoptimistic on the actual recognition capabilities of the technology. To avoid this problem it is strongly recommended to use disjoint datasets for tuning parameters, enrolling users and testing the system.

Random errors are related to the natural variation in clients and biometric samples, and are unavoidably caused by the limited amount of tests that can be done. In fact, the size of an evaluation, in terms of the number of users and the amount of attempts made, affects how accurately we can measure error rates: the larger the test, the more accurate results are likely to be. It can be useful to collect multiple biometric identifiers per person, but the number of people tested is more significant in determining test accuracy, rather than the total amount of attempts.

II.F. Limitations and issues

Automatic person recognition is a very challenging and elusive pattern recognition problem; in general it is perceived as an easy task, but the efforts required to achieve satisfactory performances have been largely underestimated. In fact, given a few samples of a biometric identifier, the key challenge is to be able to conceive a realistic representational model of the individual, and then formally extract the discriminative information present in the signal from its samples. By looking at the results of recent public evaluation campaigns [10][54][69][71], we have the proof that even after several decades of research, biometric recognition techniques are still not mature enough for real applications, and that their present market is quite restricted. In this section we analyse the main technological issues and operational limitations which negatively affect the performances of biometric devices.

II.F.1. Accuracy

The main reasons that reduce the accuracy of a biometric system can be grouped into three categories.

The first class is related to *information limitation*: the discriminative information content in the pattern samples may be inherently partial, due to the intrinsic signal capacity limitations in the biometric identifier. In other worlds, the signal captured from the biometric identifier and represented in a digital form may not be discriminating enough to distinguish between multiple different identities. Another aspect of information limitation is represented by non universal biometric identifiers, for which it is not possible to obtain functional biometric samples from a part of the population. For example it has been estimated that around 4% of the population cannot be identified through fingerprints, because of the poor quality of their ridges.

The second category denotes the *representation limitation*. The ideal recognition system should retain all invariant and discriminative information from the sensed measurements; nevertheless, in practice the feature extraction step includes some redundant and erroneous elements as well, and may fail to preserve the entire significant and distinctive information from the signal. The consequences of this imperfect representation are that: the potential power to discriminate user identities by the system is reduced, the inter-class similarities become more influent and in general the error rates gets higher.

The last category is related to *invariance limitation*. An ideal matcher should be able to precisely model the variations between different biometric samples of the same user, and to provide a robust and invariant relationship of similarity. In practice, due to data scarcity and approximated modelling of client identifiers, actual systems are not able to efficiently deal with various signals captured from the same user, and recognition performances are poor. These intra-class variations are clearly unwanted but impossible to remove, and are caused by inconsistent methods of signal acquisition. In fact, if we consider a face recognition device as an example, multiple face variations can be generated by: defective or improperly maintained sensors, unfavourable ambient conditions (illumination, light beams and shadows), differences in pose and facial expressions, occlusions, presence or absence of eyeglasses and hair, aging, etc. Data scarcity can become a major problem in modelling intra-personal variations for those techniques that require large training databases, like some behavioural and hybrid approaches.

II.F.2. Scale

How does the number of clients enrolled in database affect the performance of a recognition system? In a verification scenario the size does not really matter, because each transaction requires a one-to-one comparison between the unidentified biometric pattern and the claimed model pattern stored in the system. On the contrary, in an identification scenario there might be a need for an efficient *scaling* when the number of clients is very large, because in each transaction the unidentified pattern is matched up to all the model patterns present in the system, in a one-to-many comparison. Typical approaches to scaling include using multiple hardware units in parallel and coarse to fine pattern partitioning. Unfortunately, when using a parallelisation strategy the amount of required hardware units increases linearly with the number of clients, and this is not a feasible option in practical situations with thousands or millions of users; however, coarse to fine pattern partitioning may be a solution, but it is not easy to define a criterion to efficiently cluster biometric identifiers, providing relevant scaling advantages while maintaining good recognition results. Another possible strategy may be to index the biometric patterns like the conventional database records; however, due to large intra-class variations, it is not obvious how to ensure the samples from the same client fall into the same index bin, and consequently to obtain a low binning error rate.

II.F.3. Privacy and security

Privacy and security issues are closely interconnected. Considering that a reliable biometric device provides an irrefutable proof of identity of a person, there are serious *privacy concerns* in the application of biometric recognition systems. In fact, it may be possible to track clients from their overt and covert biometric captures, infringing the individual right to privacy; also personal biometric data may be abused for unintended or criminal purposes. Nevertheless the safeguard of individual privacy can be achieved through a clever legislation and through the design of reliable and secure recognition applications. Concerning the security aspect, it is critical to assure that the input identifier is presented by the legitimate owner, and that the captured sample is matched with a genuinely enrolled model pattern.

If we analyse the problem of having the *legitimate owner*, it should not be possible for an impostor to spoof the biometric trait of a client, and then use it to be recognised at his place. If some identifiers can be easily kept secret, like iris, DNA and retina, some others may be hiddenly spoofed and possibly replicated, like fingerprint, voice, face or signature. A potential solution to this kind of attacks may be given by *aliveness detection techniques*, which can assure that the input measurement is not captured from an unanimated object, like audio recorders or digital displays. Additionally, multi-biometric (or multimodal) recognition approaches can enforce the security of a system by requiring the presentation of various biometric identifiers, instead of only one; this clearly reduces the feasibility to spoof and correctly replicate multiple traits. Finally, it is also possible to combine aliveness and multimodality, or to exploit the synchronisation between various biometrics, like voice and face.

The second security issue is related with the *integrity of the models* enrolled in the database. In most cases, the enrolment of a new user is supervised by an operator, who can check the identity of the client and assures that the captured patterns are authentic. Though, when the enrolment can be unsupervised or when there is an adaptive procedure to update the client's model, it might be possible for an attacker to inject counterfeit biometric samples into the system, in order to corrupt the final decision results. In this case, an integrity check technique is needed and it should be possible to revoke those biometric identifiers that have been compromised. In addition, the access and management of enrolled model and stored patterns must be regulated with care. In fact, operators and administrators should not be able to willingly access the biometric identifiers present in the database, nor to recreate the original signals from their digital representations or to retrieve personal and medical information. A possible solution may be represented by administrator logging, which records all accesses and modification to stored data; an alternative and promising research direction may be *biometric cryptography* [88], which studies the generation of cryptographic keys based on biometric samples.

II.G. Multi-biometrics and multimodal biometric systems

A *multimodal biometric recognition system* [35][37][52][80] is strictly defined as a biometric recognition system that is using more than one biometric identifier to recognise a person. As discussed in detail in the following section (II.G.1), there are several "multimodal" approaches based on the same biometric identifier and a few ones based on a single sample. Considering the stringent similarities in the integration of these multiple pieces of information, we decide to extend the previous definition in order to include these particular approaches as well. Therefore, in this dissertation we generalise a *multimodal biometric recognition system* as a biometric recognition system that exploits diverse and multiple sources of biometric information to recognise a person.

The integration of multiple sources of information can have numerous advantages for biometric recognition systems: it can increase the accuracy of the systems, by exploiting complementary information, it can augment their reliability, by taking advantage of redundant and richer information that can compensate the individual weaknesses, and it can reduce their cost, by exploiting several cheap sensors. Moreover, multimodal biometric systems naturally have enforced anti-spoofing protection, because it becomes more difficult for an attacker to simultaneously spoof multiple traits, and enable more sophisticated aliveness detection techniques, by asking to present one or more unpredictable identifiers of a user. Finally, humans constantly make use of information fusion in everyday activities; in fact they naturally integrate sight, hearing, smell and touch, in order to have an augmented sensorial perception of the environment.

II.G.1. Sources of biometric information

There are different sources of biometric information that can be integrated in a multimodal system; we can classify them into one among the following five scenarios [37]:

1. *Multiple simultaneous captures of the same identifier* : in this case, various signals of the same biometric identifier are captured simultaneously by using different sensors. For example, multiple cameras can acquire different views of a face, or optical and capacitance sensing devices can capture diverse signals of a fingerprint.
2. *Multiple biometric identifiers*: this is the most common scenario, in which multiple dissimilar biometric identifiers are combined. In general, there is one sensing device for each biometric; for example, a video camera and a microphone can acquire facial video and voice data.
3. *Multiple units of the same biometric* : in this case, distinct units of the same biometric are captured and integrated for recognition. As an example, a system may acquire the scans of both irises or the fingerprints of two or more fingers.
4. *Multiple repeated captures of the same identifier* : in this scenario, repeated captures of the same biometric identifier are used to increase the accuracy of the system. For example, a video camera can provide many images of a face or a fingerprint sensing device can acquire several snapshots of a finger.
5. *Multiple representations/matching algorithms of the same capture* : in this case a biometric identifier is captured once, and then diverse feature extraction techniques and/or matching algorithms are applied for recognition. For example, one "minutiae" and one "non-minutiae" based fingerprint recognition technique can be applied on the same acquired data. The general idea behind the integration of multiple non-homogeneous features or classifiers is that their fusion may overcome the "bad properties" of each one.

Figure 7 shows a visual representation of the various scenarios depicted above. It is important to notice that diverse combinations of the previous scenarios are also possible; for example, multiple facial images can be acquired through a video camera (scenario 4), then integrated with voice data (scenario 2).

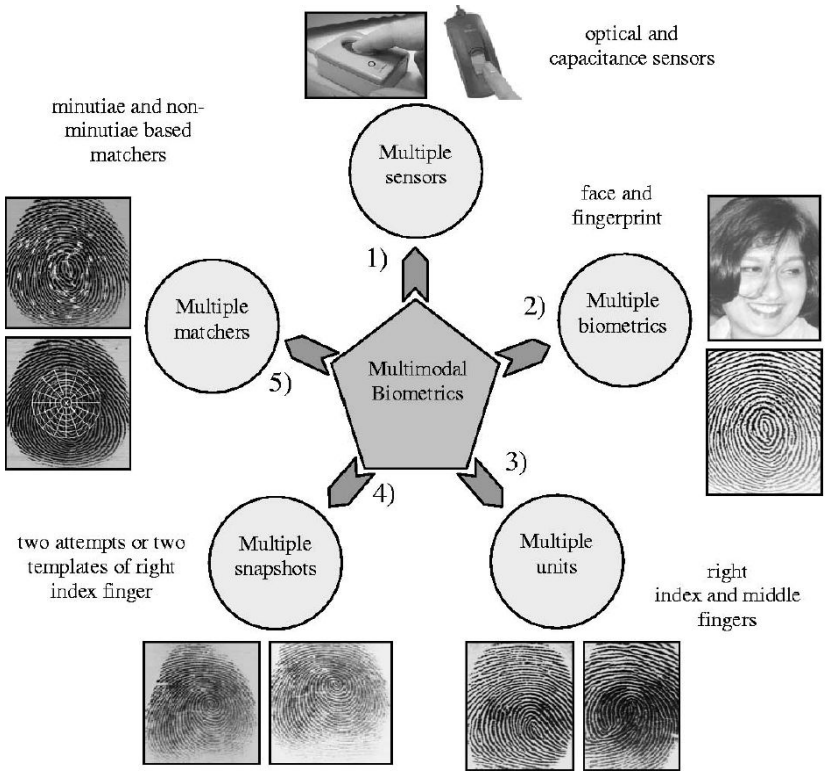

Figure 7: various scenarios ofmultimodal biometric systems [37].

During the design process of a multimodal biometric recognition system, it is very important to consider the relations among different sources of information to be integrated. In fact, it is a general trend in literature that the combination of uncorrelated modalities (like fingerprint and face) or loosely correlated ones (like face and iris) are expected to provide a better improvement of performances than the combination of correlated identifiers. This phenomena is probably caused by the fact that independent data convey the most complementary and rich information. Referring to the previous classification, scenario 2 has the highest potential since it is supposed to integrate the most independent sources of information, while scenarios 1, 4, and 5 combine the most correlated ones.

II.G.2. Integration schemes

Diverse and multiple sources of biometric information can be integrated in a multimodal recognition system by applying one among the following schemes:

- *Serial scheme*: in this situation, the output of each unimodal recognition system is linked to the input of another one, in a serial way. This mode can serve as an indexing scheme, narrowing down the number of candidates by considering only the best matches at each step. Moreover, the final decision on a transaction can be reached without acquiring or processing all biometric sources of information. This is the integration design that mostly reduces the overall recognition time of a system; in contrast, its improvements on the accuracy of the final decisions are usually the lowest.
- *Parallel scheme*: in this scenario all sources of biometric information are used simultaneously. This is the most common and powerful scheme, because the integration of multimodal information can be realized at numerous different levels: sensor data, feature, decision and opinion; next section (II.H) presents an insight on the main techniques of information fusion related to the parallel scheme.
- *Hybrid scheme*: this mode includes any hierarchical combination of serial and parallel schemes, in which the unimodal classifiers are integrated in a treelike structure; it also contains any fuzzy fusion scheme that cannot be merely considered as serial or parallel.

II.H. Information fusion for multi-biometrics

Information fusion is an independent scientific domain that studies the combination of different sources of information. Multi-biometrics and decision making problems are focused on a specific goal among those of information fusion: they are interested in the integration of multiple biometric indicators, in order to generate a richer representational format and to reach a more precise decision.

Following the proposal in [80], we divide the most important fusion strategies in three categories: pre-mapping, midst-mapping and post-mapping fusion. Figure 8 sketches a non-exhaustive tree representing the suggested classes and their more representative techniques.

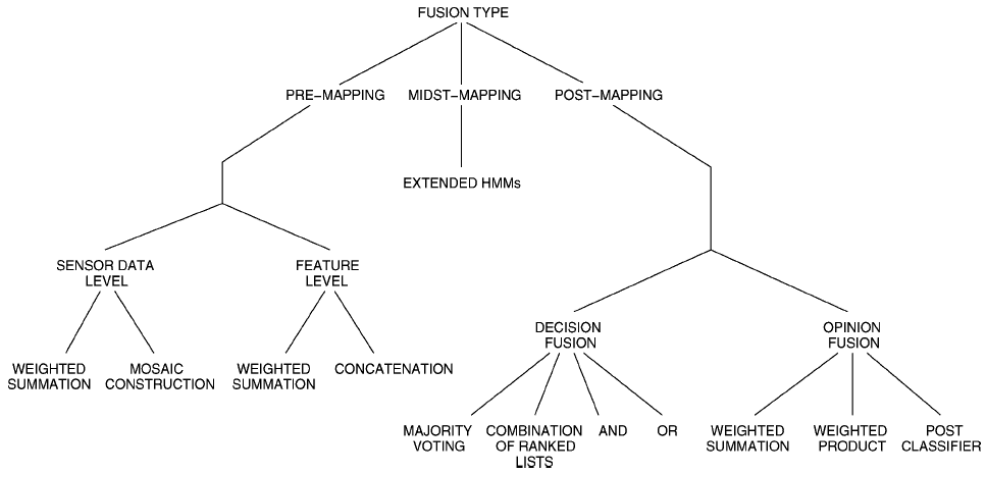

Figure 8: non-exhaustive tree ofmain fusion techniques [80].

II.H.1. Pre-mapping fusion

Pre-mapping fusion combines information in the sensor data space or in the feature space, before any classifier or expert; therefore the integration can be done at two different levels, depending on the space in question: sensor data level fusion and feature level fusion.

Sensor data level fusion integrates the raw signals captured from distinct sensing devices. One common technique is the *weighted summation*, in which multiple signals are scaled and added together; for example, several microphones can be used to record a stronger audio signal. Another important technique is the "*mosaic construction*", in which several signals are processed to obtain a richer one; for example, several single-channel audio recordings can be combined in a multi channel one, or distinct images of the same object can be processed to create super resolution images or mosaics. These methods require that the individual raw signals must be commensurate; if it is not the case, it is necessary to apply a mapping function before integration, in order to transform the input signals into a common interval.

In the *feature level fusion*, features are firstly extracted from separate signals and then combined together. One representative technique is again the *weighted summation*; if the numerical values are commensurate and the dimensions of feature spaces correspond, then feature vectors belonging to distinct signals are scaled and added together. Another technique is the *feature concatenation*, in which several feature vectors are juxtaposed to generate a single and extended one; as an example, this method can be used to combine audio and visual features. Unfortunately, the concatenation technique suffers from several drawbacks: first of all, there is no control on the information that is going to dominate the final decision, then the different features must be extracted synchronously (at the same frame rate), and finally the increase of dimension can turn out to be a problem in the classification stage. This last downside, which is known as the *curse ofdimensionality* issue, is somewhat avoided by using a suitable feature reduction technique like *principal component analysis (PCA)* [22], which extracts only the most salient information and decreases the dimension of the feature space.

II.H.2. Midst-mapping fusion

Midst-mapping fusion is a relatively new and more complex concept, since it combines information during the mapping from feature space into opinion or decision space. These techniques process several information streams concurrently, in order to provide a unified opinion or decision; they aim to exploit the temporal synergies and correlations between these sources of information, by avoiding the drawbacks of feature concatenation (detailed in Section II.H.1). The main techniques consist of extensions of *hidden Markov models (HMM)*, and are mostly used for audio-visual person recognition.

II.H.3. Post-mapping fusion

Post-mapping fusion combines information after the mapping from feature space into opinion or decision space; depending on the data type that is integrated, it is possible to distinguish between decision fusion and opinion fusion (also called score fusion).

In a *decision fusion* scenario each classifier provides an independent hard decision, and then an integration step combines these individual judgements in order to reach a final decision. One representative technique is the *majority voting*, in which the final decision is taken by considering the most popular one, among those presented by the individual classifiers. Another important method is the *ranked list combination*: in this case each classifier generates a ranked list of classes, spanning from the preferred choice to the least preferred one; after that these lists are combined by various means, possibly taking into account the reliability and discriminatory power of each classifier. Concerning the verification task, there are two straightforward operators for combining the results of a ranked list: the AND and the OR operators. The AND operator is equivalent to the unanimity rule, because all classifiers must agree on the same decision; it is a quite restrictive policy and no decision may be reached in case of disagreement. On the other hand, the OR operator is very relaxed, because the decision can be taken when just one classifier agrees. Finally, it has been theoretically demonstrated by Daugman [18], that the AND and OR operators are suboptimal, because they can enhance only one between FAR and FRR, worsening the other.

In an *opinion fusion* (also called *score fusion*) scenario each system is considered like an expert, which provides a numerical opinion on each possible decision, and then an integration step combines the scores and takes the final decision. The opinion values must be commensurate; without any loss of generality, we suppose to have all scores mapped in the interval: $o_{m,k} \in [0,1]$ for $\forall m,k$, with 0 as the lowest and 1 as the highest possible preferences. The *ranked list combination*, described in the decision fusion scenario, can be also considered as a special technique of opinion fusion; in fact, the rank of each class itself represents a measure of preference, but the absence of a score value reduces the information on the confidence of each opinion. In general, opinions are combined by using two main techniques, the weighted summation rule or the weighted product rule; after that, a final decision is reached by using the MAX operator on the final scores. This approach has a clear advantage on feature vector concatenation and all decision fusion techniques, because opinions can be weighted, easily taking into account the reliability and discriminatory power of each classifier. If we consider to have M experts, each of them providing a score $o_{m,k}$ for a given class k, then the final opinion obtained through the *weighted summation fusion* (also called *sum rule*) is the following:

$$f_k = \sum_{m=1}^{M} \alpha_m o_{m,k}$$

in which $\alpha_m \in [0,1]$ represents the normalised weight ($\sum_{m=1}^{M} \alpha_m = 1$) for expert m.

Another possibility is to use the *weighted product fusion* (also called *product rule*), which has been developed by considering a Bayesian framework and exploiting posterior probabilities as opinions. In fact, keeping the same mathematical formulation as before, the final opinion for class k is obtained as:

$$f_k = \prod_{m=1}^{M} \left(o_{m,k}\right)^{\alpha_m}$$

However, this latter technique presents some drawbacks: first of all, one expert can have a large influence on the final score of a class, especially if it has a very low opinion on that class, producing a final value close to zero. Then, in order to exploit the theoretical properties of the Bayesian framework, the individual posterior probabilities of each system must be strictly independent, an assumption usually not verified in practical applications. An alternative approach to the score fusion rule and the MAX operator can be to implement a *post-classifier* [22]; by considering the numerical opinions as "likelihood" values of each class, it is then possible to train a classifier in this "likelihood" space, let him integrate the different scores and reach a final decision. An important advantage of this approach is that the distinct opinions do not need to be commensurate, because the post-classifier automatically maps any heterogeneous "likelihood" space to a proper class label space, in which the final decision is taken. On the contrary, the main downside is that the dimensionality of the "likelihood" space is linearly dependent on the number of experts and classes, and can become huge: a multimodal system with M experts and K classes generates an opinion vector of size $M * K$. This is nevertheless reduced for the verification task, in which the number of possible classes is constant (with only clients and impostors), and the opinion vector is simply dependent on the number of experts, M.

II.H.4. Discussion on fusion strategies

It is generally believed that information fusion is potentially more effective if its integration is done as early as possible, and in literature it is common to observe that fusion at sensor or feature level offers higher improvement that at decision or opinion level. In fact, pre-mapping strategies take advantage of the richest and most genuine information (like sensor data and features), while the one used by post-mapping approaches (like similarity scores, class labels and opinions) is generated artificially, through a series of signal processing and machine learning techniques. In contrast, pre-mapping fusion appears to be the most challenging one, because the relationships among diverse sources of information in sensor or feature spaces are often not known or problematic, due to incommensurate or non compatible data. Moreover, in some proprietary systems the signal and feature data of the individual modalities is not accessible, in order to avoid industrial concurrence, so in the end post-mapping fusion becomes the only choice. In conclusion, all these issues have a direct reflection in the research literature, because there are very few publications on pre-mapping and midst-mapping fusion techniques, even if they hold the biggest potential for integrating useful discriminative information.

II.I. Concluding summary

In this chapter we introduced the discipline of biometrics, and its evolution towards multi-biometrics. We firstly defined what a biometric identifier and a biometric recognition are, by specifying their main properties and by describing their most important examples and applications. Then, we detailed the two main operational modes of a biometric system: verification (or authentication) and identification. After that, we illustrated the architecture of a typical recognition system, by explaining the steps of the three main modules: enrolment, recognition and adaptation. Afterwards, we examined the large domain of performance evaluation, where we principally focused on the multiple measures for assessing verification (or authentication) and identification results. We concluded the introduction on biometrics with a discussion on the limitations related to the accuracy and scalability of the systems, and on the privacy and security concerns associated to their utilisation. In the second part we analysed the domain of multi-biometrics, by specifying the different sources of biometric information that can be integrated in a multimodal system, and by defining the possible integration schemes. Finally, we introduced the scientific domain of information fusion, where we detailed the typical integration strategies applied in multi-biometrics.

Chapter III. Person recognition using facial video information: a state of the art

III.A. Introduction

For decades human face recognition [12][67][96] has been an active topic in the field of person recognition, or more generally in the field of object recognition. Most of algorithms have been proposed to deal with individual images, where usually both the enrolment and testing sets consist of a collection of facial pictures. Image-based recognition strategies have been exploiting only the physiological information of the face; in particular its appearance encoded in the pixel values of the images. Furthermore, the recognition performances of these approaches [69] have been severely affected by different kinds of variations, like pose, illumination and expression changes. Thus, researchers have started to look at video-based recognition, in which both the enrolment and recognition sets are facial video sequences representing the clients of the system.

Person recognition using facial video information has some advantages over image-based recognition. First of all, video frames can provide a huge amount of data compared to single pictures, and more robust and stable recognition can be achieved by integrating information and decisions from previous frames. Then, in addition to the physiological information already present in images, also the temporal one becomes available and can be exploited to improve the recognition task; consequently, nowadays researches have the possibility to analyse not only facial appearance but also head and facial motion, and human face starts to be considered as a hybrid biometric identifier (Section II.B.2), rather than only a physiological one. After that, more effective representations such as 3D face models [78] or super resolution images [89] can be generated from video sequences and used for recognition. Finally, video data allows learning and updating user models over time.

Currently, most person recognition techniques using videos are straightforward generalisations of image-based algorithms; in these systems, the feature extraction and classification are applied independently to each frame, then the similarity scores are integrated using post-mapping information fusion techniques (Section II.H.3), like the majority voting or the weighted summation rule. However, a few recent attempts that exploit the temporal information in videos have emerged to the scientific community, and these studies reveal the feasibility and benefit of considering the face as a hybrid biometric. Therefore, due to the fact that person recognition strategies using videos involve a heterogeneous mixture of techniques, we propose to divide them into the following two categories: those that neglect the temporal information, and those that exploit it even partially. We consider that a recognition approach neglects the temporal information, if the shuffling of frames in each video has no influence on the discriminatory power and the global performance of the system. In contrast, breaking the temporal consistency of video frames should have an evident impact on those techniques that exploit (even partially) the temporal information for recognition.

In this survey, we focus on those person recognition approaches that make use of facial video information. In particular, we analyse their feature extraction, model estimation and classification parts (Section II.D), and we overlook the sensing and pre-processing (face detection and segmentation) steps. It is worth noting that we do not consider those recognition strategies that are based on image, audio or 3D data. Finally, in the following survey we intentionally do not report a systematic quantitative comparison between the different techniques. In fact, the absence of common testing databases and the heterogeneous experimental conditions presented in the research literature offer results that are incommensurate; for this reason, proposing a quantitative comparison of scores would be meaningless.

III.B. Approaches neglecting the temporal information

III.B.1. Eigenfaces: extensions to video

Eigenfaces [87] is one of the essential basic techniques for person recognition by using facial appearance; it has been widely studied and largely applied to image and video data. It is out of the scope of this state of the art to detail all variants to the standard approach; here we focus on its applications to video, and the interested reader can refer to [12], [67] and [96] for the image case.

The eigenface technique is based on the notion of dimensionality reduction; in fact, Kirby and Sirovich [40] were the first to remark that the dimensionality of the *face space* is much smaller of that of a single face, considered as an arbitrary image. A first method to reduce the image space into a low dimensional feature space is to apply the *principal component analysis (PCA)* (also called the *Karhunen-Loeve transform (KLT)*) [22]: PCA computes a set of orthonormal vectors (the so called *eigenfaces*), which optimally represent the distribution of the training data in the root mean squares sense. A visual example of the most important eigenfaces is presented in Figure 9.

We consider to have a set of N vectorised sample images, $\{\mathbf{s}_n \in \mathrm{N}^M \mid n = 1,\ldots,N\}$, which take values on an M-dimensional image space and belong to one of the K classes (the individuals in the database). A possible approach to generate a mapping from the M-dimensional image space into an D-dimensional feature space is to use a *linear transformation*; in this case, each input image $\mathbf{s}_n$ can be approximated with its feature vector, $\mathbf{x}_n \in \Re^D$, by using the following linear projection:

$$\mathbf{x}_n = \mathbf{W}^T(\mathbf{s}_n - \boldsymbol{\mu})$$

for $n = 1,\ldots,N$, where $\mathbf{W} \in \Re^{M \times D}$ is the projection matrix with orthonormal columns, and $\boldsymbol{\mu} \in \Re^M$ is the mean image of all samples. In general, $D < M$, so a projected image can be reconstructed in the image space up to a certain error image, $\boldsymbol{\varepsilon}_n$; the reconstructed image is calculated as:

$$\tilde{\mathbf{s}}_n = \boldsymbol{\mu} + \mathbf{W}\mathbf{x}_n = \mathbf{s}_n + \boldsymbol{\varepsilon}_n$$

In the eigenface approach the optimal projection matrix, $\mathbf{W}^{(EIG)}$, is chosen by computing the principal components of the training data, or equivalently by maximising the determinant of the global scatter matrix of the projected samples. If the *global scatter matrix* is the following:

$$\mathbf{G} = \sum_{n=1}^{N} (\mathbf{s}_k - \boldsymbol{\mu})(\mathbf{s}_n - \boldsymbol{\mu})^T$$

then the optimal projection matrix is calculated as:

$$\mathbf{W}^{(EIG)} = \arg\max_{\mathbf{W}} |\mathbf{W}^T\mathbf{G}\mathbf{W}| = [\mathbf{w}_1,\ldots,\mathbf{w}_D]$$

The solution to the previous equation is the space spanned by the set of M-dimensional eigenvectors, $\{\mathbf{w}_d \in \Re^M \mid d = 1,\ldots,D\}$, corresponding the D largest eigenvalues of the scatter matrix:

$$\mathbf{G}\mathbf{w}_d = \lambda_d \mathbf{w}_d$$

The optimal projection matrix can be also calculated by exploiting the global covariance matrix, because the covariance and scatter matrices differ by just a scaling factor: $\boldsymbol{\Sigma} = \frac{1}{N}\mathbf{G}$. Moreover, considering that the number of images in the training set is habitually lower than the dimensionality of the image space ($N << M$), then the maximum possible number of eigenvectors (or eigenfaces) is: $D^{(MAX)} = N$.

In some cases, the feature vectors, $\mathbf{x}_n \in \Re^D$, are rescaled by a process called *whitening* [48]. In fact, the mean squared error that underlies the PCA preferentially weights low frequencies, those that correspond to the largest eigenvalues; therefore, the whitening is applied to counterbalance this phenomenon, through the normalisation of the scatter matrix for uniform gain control. Hence after the whitening process, each component associated to each eigenvector has a uniform unit variance:

$$\hat{x}_{n,d} = \frac{x_{n,d}}{\sqrt{\lambda_d}}$$

for $n = 1,\ldots,N$ and $d = 1,\ldots,D$, where $\hat{\mathbf{x}}_n \in \Re^D$ is the whitened feature vector.

The recognition task is usually done through a *nearest neighbour classifier*, in which the similarity measure is inversely proportional to distances in the reduced feature space; the most common distances are based on simple metrics, like L_1 (city-block), L_2 (Euclidean), cosine or Mahalanobis.

Figure 9: eigenfaces ofa set ofimages ofthe Stirling database [86].

In [81], Satoh proposed a straightforward extension of the traditional eigenface approach, by introducing a new similarity measure for matching video data. The similarity between distinct videos was obtained by considering the smallest distance between frame pairs (one from each video), in the reduced feature space. Taking into account the large variation of facial appearance in a given sequence, the choice of the closest frame pair showed limited robustness to outliers and obtained poor results.

Two similar extensions have been proposed by Huang and Trivedi [33], by using decision fusion techniques (Section II.H.3) to integrate opinions on each frame. The authors firstly applied the eigenface strategy to individual frames, to obtain a sequence of independent decisions on the identity of the user; after that, they reached a final decision by using the majority rule, which chose the most frequent identity in a video, or by adding a post classifier, which was implemented using discrete *hidden Markov models (HMMs)* with a maximum likelihood rule. In their experiments, the recognition system with the post classifier obtained slightly better results than the one exploiting the majority rule, and both showed improvements on the standard eigenface approach testing one frame per video.

In order to increase the performances of eigenspace-based strategies, the abundant video data has been exploited to train statistical models of the individual facial manifolds. Firstly, Torres and Vilà [86] employed the *subspace method* [64] with video data, which they named as the *self-eigenface* approach. To represent each individual facial manifold they generated multiple personal eigenspaces, one for each user, which were trained by selecting different views of the same person in a video sequence. For the classification task, the authors implemented a suitable similarity measure, which was based on the reconstruction error of a testing image after the projection on each individual subspace. The authors also exploited the colour information, which is habitually present in sequences, by creating separate eigenspaces for each colour component and for each individual; consequently, they modified the similarity measure by defining a total reconstruction error as a weighted sum of the reconstruction errors of each colour component.

Then, Satoh [81] extended the *CLAFIC method* [64] to face sequence matching, by proposing two different implementations. The first version was really close to the self-eigenface approach of Torres and Vilà: in both cases, an input frame was identified as the individual who generated the closest eigenspace to it. More precisely, the author trained multiple personal subspaces, and adopted a similarity measure proportional to the longest projection of a test in each subspace. For the second method, Satoh noticed that face images of a person compose a non-linear manifold, so he tested a CLAFIC-based variant in a nonlinear space. In particular, he developed a CLAFIC adaptation of the *kernel-based nonlinear subspace method* [53], which consisted of: a nonlinear transformation of feature spaces defined by kernel functions, and an application of the subspace method in the transformed high-dimensional spaces.

In [93], Yamaguchi et al. applied the *mutual subspace method* [64] to the problem of person recognition using facial video sequences. The main difference between this approach and the other subspace methods was that it exploited an eigenspace approximation not only for client modelling, but also for any test sequence; this way, the authors implemented a "space-to-space" matching, and considered as similarity measure the angle between one input and one reference subspaces. Recently, Nishiyama et al. [61] further improved this strategy, by constraining the "space-to-space" matching onto multiple special subspaces, built to enhance the discrimination between classes. In fact, both the input and reference eigenspaces were projected onto each matching subspace, in order to calculate partial similarity scores (the angles); then, the global similarity values were obtained through a weighted sum of the partial ones. This last approach was able to provide very good recognition scores: 96.8% of CIR on a database containing 500 users; nevertheless, it exploited the video just as a source of data, neglecting the temporal information.

III.B.2. Fisherfaces: extensions to video

Fisherfaces [3] is another state of the art technique for person recognition using facial appearance. Similarly to the eigenface approach, presented in Section III.B.1, fisherfaces is also based on the notion of face space reduction into a low dimensional feature space. The optimal projection is calculated by applying the *Fisher's linear discriminant (FLD)* (also called *linear discriminant analysis (LDA)*) [22], which is a class specific linear method that tries to "shape" the scatter, in order to make it more reliable for classification. While in the eigenface strategy the scatter being maximised is due to between-class scatter (useful for classification) and within-class one (unwanted information); in the fisherface approach the scatter being maximised is the ratio between the between-class scatter and the within-class. Figure 10 shows a visual representation of the first fisherfaces obtained through this method.

We consider having the same framework as in the eigenface approach (section III.B.1): a linear projection from the image space to the feature subspace, then a matching step with a nearest neighbour classifier using distances. Following the same notation, the *between-class scatter matrix* is defined as:

$$\mathbf{G}_B = \sum_{k=1}^{K} N_k (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^T$$

in which $\boldsymbol{\mu} \in \Re^M$ is the mean image of all samples, $\boldsymbol{\mu}_k \in \Re^M$ is the mean image of class k, and N_k is the cardinality (number of samples) of class k. The *within-class scatter matrix* is the following:

$$\mathbf{G}_W = \sum_{k=1}^{K} \sum_{\mathbf{s}_n \in k} (\mathbf{s}_n - \boldsymbol{\mu}_k)(\mathbf{s}_n - \boldsymbol{\mu}_k)^T$$

If $\mathbf{G}_W \in \Re^{M \times M}$ is non-singular, the FLD chooses the optimal projection, $\mathbf{W}^{(FLD)} \in \Re^{M \times D}$, by maximising the ratio of the determinant of the between-class scatter matrix of the projected samples on the determinant of the within-class scatter matrix of the projected samples:

$$\mathbf{W}^{(FLD)} = \arg\max_{\mathbf{W}} \frac{|\mathbf{W}^T \mathbf{G}_B \mathbf{W}|}{|\mathbf{W}^T \mathbf{G}_W \mathbf{W}|} = [\mathbf{w}_1, \ldots, \mathbf{w}_D]$$

where $\{\mathbf{w}_d \in \Re^M \mid d = 1, \ldots, D\}$ is the set of generalised eigenvectors of $\mathbf{G}_B$ and $\mathbf{G}_W$, corresponding to the largest generalised eigenvalues $\{\lambda_d \mid d = 1, \ldots, D\}$:

$$\mathbf{G}_B \mathbf{w}_d = \lambda_d \mathbf{G}_W \mathbf{w}_d$$

for $d = 1,\ldots,D$. Considering that there are at most $K-1$ non-zero generalised eigenvalues, due to the definition of the scatter matrices, and that in general the number of classes is lower than the total number of training images, $K < N << M$, then the highest achievable dimension for the feature space is: $D^{(MAX)} = K-1$. In practical cases, it is not possible to compute the optimal projection by using the previous criterion, because of the singularity of the within-class scatter matrix; in fact the rank of $\mathbf{G}_W \in \Re^{M\times M}$ is at most $N-K$, and usually $N << M$, so $\mathbf{G}_W$ is always singular.

One strategy to overcome this problem, called *fisherfaces* [3], is to reduce the image space before applying the FLD; by using a PCA transform, the face space is decreased to $N-K$ and the resulting $\mathbf{G}_W$ is non-singular. The optimal projection, $\mathbf{W}^{(FIS)} \in \Re^{M\times D}$, is then:

$$\mathbf{W}^{(FIS)T} = \mathbf{W}^{(FLD)T}\mathbf{W}^{(PCA)T}$$

where $\mathbf{W}^{(PCA)} \in \Re^{M\times(N-K)}$ and $\mathbf{W}^{(FLD)} \in \Re^{(N-K)\times D}$ are projection matrices with orthonormal columns.

Figure 10: fisherfaces of a set of images of the FERET database.

In [81], Satoh also proposed a straightforward extension of the traditional fisherface approach, by employing the same video similarity measure developed for the eigenface case (Section III.B.1). Again, the similarity between distinct videos was calculated by considering the smallest distance between frame pairs (one from each video), in the reduced feature space. This strategy experienced the same weaknesses to outlier frames as before, but it obtained better recognition results because the fisherface method is known to be more discriminating than the eigenface one.

III.B.3. Active appearance models

Active appearance models (AAMs) [17][23] are statistical models of the face that combine shape and intensity variation in an unified framework. They are capable of estimating full photo-realistic models by using a reduced number of parameters and their rapid and accurate optimisation algorithm makes them valuable in a tracking and recognition scenario.

The statistical model of shape variation is obtained by extracting the face shape with *active shape models* [16]; a set of key landmark points are firstly located in an input image, then aligned in a common coordinate system, and finally represented in an optimal subspace computed using the *principal component analysis (PCA)* (also called the *Karhunen-Loeve transform (KLT)*) [22]. If we call $\mathbf{q} \in \Re^{M^{(q)}}$ the *aligned shape vector*, then it is approximated as following:

$$\mathbf{q} = \boldsymbol{\mu}^{(q)} + \mathbf{W}^{(q)}\mathbf{v}^{(q)}$$

where $\boldsymbol{\mu}^{(q)} \in \Re^{M^{(q)}}$ is the mean shape, $\mathbf{W}^{(q)} \in \Re^{M^{(q)} \times D^{(q)}}$ is the projection matrix with orthonormal columns, and $\mathbf{v}^{(q)} \in \Re^{D^{(q)}}$ are the shape projection coefficients.

The statistical model of appearance is calculated using normalised facial images, which are initially warped to match the mean shape and then processed to reduce the illumination variation. Again, the *shape-normalised image vector* $\mathbf{s} \in \mathrm{N}^{M^{(s)}}$ is linearly approximated through PCA:

$$\mathbf{s} = \boldsymbol{\mu}^{(s)} + \mathbf{W}^{(s)}\mathbf{v}^{(s)}$$

where $\boldsymbol{\mu}^{(s)} \in \Re^{M^{(s)}}$ is the mean normalised appearance, $\mathbf{W}^{(s)} \in \Re^{M^{(s)} \times D^{(s)}}$ is the projection matrix with orthonormal columns, and $\mathbf{v}^{(s)} \in \Re^{D^{(s)}}$ are the appearance projection coefficients.

The final active appearance model is obtained by jointly representing the shape and appearance models in the optimal PCA subspace. If we consider the concatenated vector, $\mathbf{g} \in \Re^{\left(D^{(q)}+D^{(s)}\right)}$:

$$\mathbf{g} = \begin{bmatrix} \mathbf{B}^{(q)}\mathbf{v}^{(q)} \\ \mathbf{v}^{(s)} \end{bmatrix} = \begin{bmatrix} \mathbf{B}^{(q)}\mathbf{W}^{(q)T}\left(\mathbf{q} - \boldsymbol{\mu}^{(q)}\right) \\ \mathbf{W}^{(s)T}\left(\mathbf{s} - \boldsymbol{\mu}^{(s)}\right) \end{bmatrix}$$

in which $\mathbf{B}^{(q)} \in \Re^{D^{(q)} \times D^{(q)}}$ is a diagonal matrix for shape scaling, then the *combined shape-appearance projection coefficients* are:

$$\mathbf{x} = \mathbf{W}^{(x)T}\mathbf{g}$$

where $\mathbf{W}^{(x)} \in \Re^{\left(D^{(q)}+D^{(s)}\right) \times L}$ is the combined orthonormal projection matrix. Figure 11 shows a visual example of the principal modes of variation of the shape-appearance parameters, $\mathbf{x} \in \Re^{L}$.

Due to the linear nature of the AAMs, the aligned shape vector, $\mathbf{q} \in \Re^{M^{(q)}}$, and the shape-normalised image vector, $\mathbf{s} \in \mathrm{N}^{M^{(s)}}$, can be directly expressed from the shape-appearance parameters, $\mathbf{x} \in \Re^{L}$:

$$\begin{aligned} \mathbf{q} &= \boldsymbol{\mu}^{(q)} + \mathbf{W}^{(q)}\mathbf{B}^{(q)}\mathbf{W}_q^{(x)}\mathbf{x} \\ \mathbf{s} &= \boldsymbol{\mu}^{(s)} + \mathbf{W}^{(s)}\mathbf{W}_s^{(x)}\mathbf{x} \end{aligned} \quad \text{where } \mathbf{W}^{(x)} = \begin{bmatrix} \mathbf{W}_q^{(x)} \\ \mathbf{W}_s^{(x)} \end{bmatrix}$$

The estimation of the model parameters, from a starting approximation to a precise fit, is solved through a high dimensional optimisation algorithm, which considers as cost function the root mean square error between the reconstructed and the original images. The optimal solution is obtained through a rapid search strategy, where the huge space of possible solutions is constrained by linearly modelling the relationship between the variations of the shape-appearance parameters and those of the reconstructed images, $\partial \mathbf{x} = \mathbf{A}\partial \mathbf{s}$.

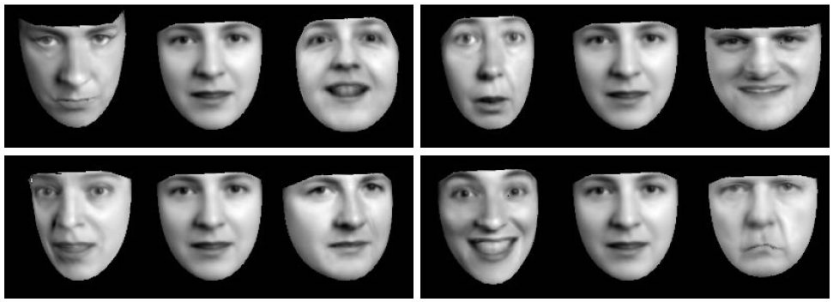

Figure 11: first four modes ofshape-appearance variations [17].

Edwards et al. [24][25] successfully applied active appearance models to the problem of person recognition, by separating the inter-class variability from the intra-class one. In fact, computing the *Fisher's linear discriminant (FLD)* (also called *linear discriminant analysis (LDA)*) [22] in the shape-appearance space, they isolated the parameters related to identity, $\mathbf{u} \in \Re^{L^{(u)}}$, from those of non-identity (pose, expressions...), $\mathbf{r} \in \Re^{L^{(r)}}$:

$$\mathbf{x} = \mathbf{U}\mathbf{u} + \mathbf{R}\mathbf{r}$$

where $\mathbf{U} \in \Re^{L \times L^{(u)}}$ and $\mathbf{R} \in \Re^{L \times L^{(r)}}$ are mutually orthogonal projection matrices, relative to identity and non identity spaces. For video processing the authors developed an iterative tracking algorithm based on Kalman filtering, in which the process model for the identity parameters was assumed constant, and the one for non-identity values was first order (constant velocity). In the end, the recognition system compared the different identity parameters robustly estimated from video sequences, by exploiting the Euclidean distance as similarity measure.

In [46] Li et al. implemented a variant of AAMs, by developing a multi-view dynamic face model to extract shape-and-pose-free normalised facial textures. In fact, their statistical model of shape variation was different than *active shape models* [16], because they computed a sparse 3D point distribution model using the 2D positions of facial landmarks, rather than estimating a 2D shape model. Then, the authors extracted nonlinear discriminative features [47], by applying the *kernel discriminant analysis (KDA)* on the shape-and-pose-free facial images (rather than using the combined shape-appearance vectors); this way, the shape information was employed only in the normalisation step and not for recognition. Li et al. also implemented a partitioning strategy based on pose information (tilt and yaw) [47], in order to compare similar views and reduce the intra-class variability. In particular, for each predefined view of a client, the related discriminative features were approximated using a plane; the pose information was then exploited for matching testing frames with facial models of corresponding planes. The authors adopted the Euclidean distance measure for estimating the frame similarity, and computed a weighted summation of individual distances to obtain the global video score.

III.B.4. Radial basis function neural networks: extensions to video

Radial basis function neural networks (RBFNNs) [6] have been applied to biometric person recognition tasks mostly because of their computational simplicity, robust generalisation properties (for example across views and facial orientations), goodness at handling sparse high-dimensional data, and guarantee to obtain a globally optimal solution.

RBFNNs have a feed forward architecture, with one input layer, one hidden layer and one output layer, as shown in Figure 12. The input layer has N_I units, and an N_I-dimensional input vector, it is fully connected to N_H hidden units; the hidden layer is also fully connected to N_C output units, which represent the N_C classes.

The activation functions in the hidden layer are generally Gaussian kernels, with mean vectors (centres), $\boldsymbol{\mu}_j \in \Re^{N_I}$, and covariance matrices, $\boldsymbol{\Sigma}_j = \sigma_j^2 \mathbf{I} \in \Re^{N_I \times N_I}$ (here considered as diagonal); the number and spread of these centres influence the smoothness of the mapping. The activation values of the hidden units are then given by:

$$g_{i,j} = e^{\frac{-\|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2}{2\sigma_j^2}}$$

for $i = 1, \ldots, N_I$ and $j = 1, \ldots, N_H$. This value is related to the proximity between the test sample, $\mathbf{x}_i \in \Re^{N_I}$, and the centre vector, $\boldsymbol{\mu}_j$. The σ_j^2 parameters determine the width and the scale of the activation functions and are estimated from the distance between centres.

These same hidden units are fully connected to the output units through a series of weights: $\{\alpha_{j,k} \mid j = 1, \ldots, N_H, k = 1, \ldots, N_C\}$. The global response of the k-th output unit for the input vector $\mathbf{x}_i$ is the following:

$$y_{i,k} = \sum_{j=0}^{N_H} \alpha_{j,k} g_{i,j}$$

for $k = 1,\ldots,N_C$, where $g_{i,0} = 1$ is the bias unit.

The training of a RBFNN consists of estimating the model parameters, $\boldsymbol{\Theta} = \{\boldsymbol{\mu}_j, \sigma_j, \alpha_{j,k}\}$, for the problem in question; by applying the pseudo-inverse method, the matrix of weights is obtained through a standard least squares solution, which allows exact calculations and instantaneous training. The classification task is done by computing the output vector $\mathbf{y}_i \in \Re^{N_C}$, and then choosing the highest activated output unit.

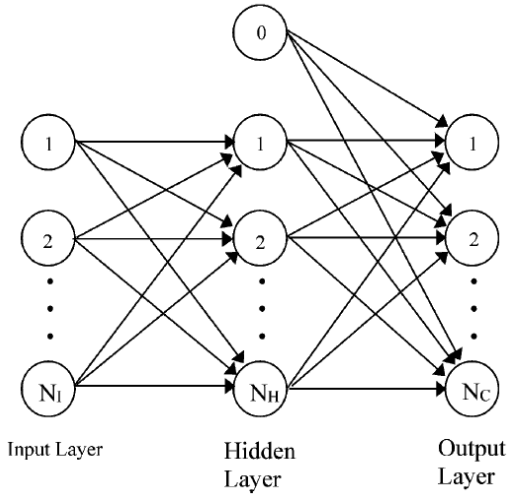

Figure 12: radial basis function neural network [31].

RBFNNs have been firstly applied to video sequences by Howell and Buxton in [32], even if they were just considering video frames as abundant test data. They developed their networks by using two feature spaces: one represented by the zero crossing information from the difference of Gaussian filtering of images, and the other by applying Gabor wavelet analysis. Then, the authors employed each training vector as a centre for their hidden units, and computed the $\{\sigma_j\}$ values by using the average of their Euclidean distances. Finally, they implemented a confidence measure in order to check the quality of video frames, and discard those unsuitable for classification: considering that training vectors that are different from network centres are mapped to low output values, they chose as a confidence measure the ratio between the highest over the second highest output, and discarded the frame if below a threshold value.

In a successive work [31], Hock Koh et al. improved the framework of Howell and Buxton, firstly by smoothing the confidence measures on a time window through a median filtering, then by implementing a measure of video similarity that combined the individual decisions on frames with the majority vote rule. The feature space was obtained by applying a radial grid sampling on the input frames: they first located a few sampling points by centring a uniformly spaced radial grid on the nose tip, then for each point they calculated the mean value over a circular patch. They also modified the RBFNN training, where they computed the centres by taking the mean vectors of multiple training images (not only one vector), previously clustered in a supervised manner by a K-means algorithm. The authors reported that their approach had a good tolerance over variations in facial scale and orientation ($\pm 25°$).

III.B.5. Elastic graph matching: extensions to video

Elastic graph matching (EGM) is a person recognition technique which has its roots in the neural network community. It has been firstly introduced by Lades et al. [43], then it has been improved by Wiskott et al. in their *elastic bunch graph matching (EBGM)* version [92].

The original EGM approach builds a face graph for each user model, by applying a rectangular grid, Ψ, on a training image; the lattice used is much coarser than the pixel one. The facial information is captured at each position (i, j) of the grid through the feature vector field $X = \{\mathbf{x}_{i,j} \mid (i, j) \in \Psi\}$, in which each feature vector, $\mathbf{x}_{i,j}$, summarises the local properties of the face and is called a *jet*. In general, jets are calculated by using the absolute value of Gabor wavelet coefficients, but other descriptors have been employed, like morphological feature vectors. An analogous approach is applied on a test image; in this case, however, the vector field $Y = \{\mathbf{y}_{u,v} \mid (u, v) \in \Gamma\}$ is computed on a finer grid, Γ, as shown in Figure 13.

Afterwards, to be able to compare related jets, the EGM method needs to find the best mapping between the face graph of an enrolled model and that of a test image; we indicate with M^* the optimal mapping among all possible mappings, $\{\mathrm{M}\}$, between the vector fields X and Y. The quality of a given match, M, is evaluated through a cost function, $Q(\ )$, that favours the similarity of associated jets and penalises the spatial deformation of the lattice:

$$Q(\mathrm{M}) = Q^{(MTC)}(\mathrm{M}) + \rho Q^{(DEF)}(\mathrm{M})$$

where $Q^{(MTC)}$ is the cost of jet matching, $Q^{(DEF)}$ is the cost of grid deformations, and ρ is a weighting parameter that controls the *rigidity of the mapping*. The overall cost of jet matching, $Q^{(MTC)}$, is computed by adding the individual cosine distances between corresponding jets; similarly, the overall cost of grid deformations, $Q^{(DEF)}$, is obtained by averaging the node deformations, which are calculated using the Euclidean distance.

Unfortunately, the number of all possible mappings is extremely large and no exhaustive search is possible; on the other hand, it is not necessary to find the optimal mapping, $\mathbf{M}^*$, but a close approximation to it will be sufficient for the recognition task. The solution to the matching problem is obtained through a two step optimisation:

1. *Rigid matching* ($\rho \to \infty$): first of all, the model graph is rigidly shifted around the test one in a sparse scanning, providing a rough head localisation.
2. *Deformable matching*: then, the model graph is varied in size and aspect ratio and the nodes are stretched with random local perturbations, in order to obtain a precise alignment.

In the end, recognition is achieved by computing similarity scores based on the overall cost of jet matching, without using the cost of grid deformations.

The major improvement in the EBGM approach [92] is the association between graph nodes and facial landmarks, like the pupils, the corners of the mouth and the tip of the nose, which are called *fiducial points*; the face graph becomes object-oriented, presenting a well defined grid structure, in which the same nodes correspond to the same facial landmarks.

In the EBGM case, the best mapping between distinct face graphs is greatly simplified, since it is assured by the presence of common fiducial points; nevertheless, the algorithm needs to locate these points on each new image. Initially, the graph matching procedure is enhanced by the introduction of a general representation of the face, called the *face bunch graph*; its purpose is to provide a wide-ranging description of the human face by bundling together in a bunch several feature vectors, which refer to the same fiducial point. For example, a fiducial point related to the eye should contain jets computed on diverse conditions: when the eye is open and close, when the user wears glasses, for men and women, Asians and Europeans, etc. Then, the face bunch graph is used for the localisation of the graph on a face image; it exploits a cost function similar to that of original approach but more accurate, due to the introduction of the phase information in the cost of jet matching. As a final point, the two-step optimisation and the recognition parts are identical to the EGM ones.

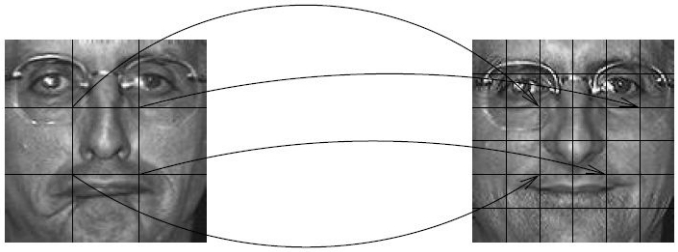

Figure 13: example ofmapping between a template (or model) image and a query (or test) image.

Steffens et al. [83] developed a complete system from video acquisition to final recognition, in which they extended the EBGM technique to image sequences. The authors improved the original method by implementing a coarse-to-fine matching approach: in order to speed up the selection of video frames used for recognition, they built multiple face bunch graphs of different complexity and sizes, which were applied sequentially. More precisely, their system firstly run a preliminary quality check on the captured frames using the coarser bunch graph, in order to extract the best two frontal face images; these two frames were further normalised with histogram equalisation and background removal, and then precisely matched with the more complex face graphs. Finally, only the best mapping was considered for recognition; in fact, even if both similarity scores were computed, only the highest was retained for the final decision.

III.B.6. Hierarchical discriminative regression trees

Hierarchical discriminative regression trees (HDRTs) [34] are decision trees that have been developed for classification and regression tasks. In person recognition applications, they generate a mapping from image space to identity space, and their branches represent conjunctions of features that lead to classification. HDRTs have the advantage that they are fast for training and testing, they are scalable and can handle large databases, they use multivariate nonlinear splits (and linear ones as special case), and that they are constructed incrementally (online).

First of all, the classification problem is cast into a regression one in order to employ HDRTs. For this purpose, each image vector $\mathbf{x}_n \in \mathrm{N}^M$ belonging to class k is converted into the sample pair $(\mathbf{x}_n, \boldsymbol{\mu}_k)$, where $\boldsymbol{\mu}_k \in \Re^M$ is the continuous outcome for class k and it is calculated by taking the average of all image vectors belonging to that class. Then, the training algorithm incrementally builds a HDRT from a series of input sample pairs, $(\mathbf{x}_n, \boldsymbol{\mu}_k)$, which are doubly clustered in both input and output spaces.

At each node, the partitioning in output space provides a virtual class label (the $\boldsymbol{\mu}$-cluster) that is used for determining to which $\mathbf{x}$-cluster the arriving sample belongs. The clustering is obtained by using Euclidean distance and every $\boldsymbol{\mu}$-cluster is represented by only its mean value, which is incrementally updated using amnesic average [91]. In parallel, the partitioning in the input space is used to structure the regression tree: at each node, every $\mathbf{x}$-cluster approximates the sample population for the vectors which belong to it. Similarly to the partitioning in the output space, no samples are stored and the $\mathbf{x}$-clusters are represented using only their means and covariance matrices, which are updated incrementally with their amnesic versions.

Unfortunately, the dimension of the input space is very high and the image vectors have excessive redundant information, so the clustering process and the distance calculations are too demanding. For this reason, the algorithm projects the input samples into a discriminative subspace; this step is necessary to reduce the size of the partitioning space, but it also generates feature vectors containing less irrelevant and noisy information. Considering that in each node there are no more than C $\mathbf{x}$-clusters, the linear discriminative subspace that passes through the centres of these clusters is represented by $C-1$ orthonormal basis vectors, obtained through a Gram-Schmidt ortho-normalisation process. Figure 14 shows an example of a HDRT for person recognition where every block represents a tree node, with the $\mathbf{x}$-cluster centres in the first row and the orthonormal basis vectors in the second one.

Moreover, the probability for a sample, $\mathbf{x}$, to belong to a given $\mathbf{x}$-cluster is approximated using a distance metric similar to a multidimensional Gaussian density function; this distance is necessary to determine which cluster should be recursively searched, until the corresponding child node is found. To better deal with large, small and unbalanced sample cases, the algorithm implements a *size dependent negative-log-likelihood* distance measure [34], which allow a progressive smooth transition among Euclidean, Mahalanobis and Gaussian negative-log-likelihoods, based on the number of samples available. It's worth noting that if all $\mathbf{x}$-clusters were modelled using a standard Gaussian distribution, then the tree structure would implement a hierarchical version of a *Gaussian mixture model (GMM)*, where the shallow levels would be modelled with large Gaussians and the deep ones with smaller Gaussians.

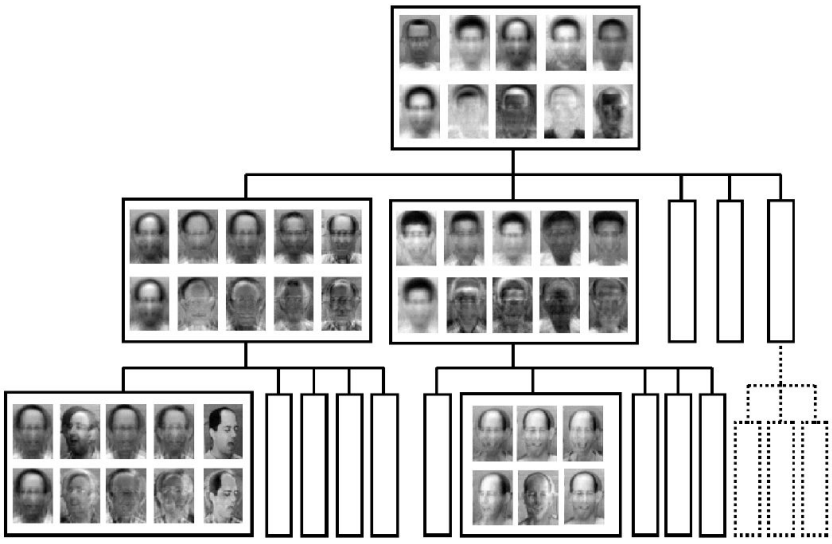

Figure 14: illustration of the hierarchical discriminative regression tree for person recognition [34].

HDRT have been applied to person recognition using videos by Weng et al. [91]. Their system exploited facial sequences only as a source of data; in fact, the frames were considered independently both for tree generation and performance evaluation. The input videos were initially pre-processed by applying a zero-mean-unit-variance normalization and by radially weighting the pixels of the face, assigning more importance to the central ones. In the end, the recognition system obtained very good results, analogous to a LDA-based nearest neighbour classifier, and definitively better than any alternative approach using regression trees.

III.B.7. Unsupervised pair wise clustering

Clustering methods automatically partition a data set into subsets (the so called *clusters*), so that the patterns in each subset share some common traits, often *proximity* according to some defined distance measure. *Pair wise clustering algorithms* are a special case of clustering methods, in which the partitioning is based on pair wise relations between individual patterns, rather than centralised relations between samples and a few cluster representatives (like centroids, for example). Pair wise clustering can be visualised through a graph (as in Figure 15), in which each node represents a pattern and the edges correspond to proximity values.

It is well known that facial video sequences captured in unconstrained dynamic scenes are affected by numerous variations, like different views, scale, illumination and facial expressions, and that they form complex non-linear manifolds in face image space. Therefore, clustering using individual frames is problematic, because the intra-class variations can be larger than inter-class ones, and the resulting partitioning may discriminate views rather than identities. In a similar way, centroids can be meaningless or difficult to define, because in an *unsupervised scenario* there is no explicit category information available and the number of clusters (identities) is not known in advance. However, when using pair wise clustering two video sequences are not related directly to each other or to a representative centroid, but they can be linked together through a third sequence or a connected group of sequences, thus forming an *associative chain* (as illustrated in Figure 15) which reduces the effect of intra-class variations.

A key element of pair wise clustering algorithms is the computation of the *proximity matrix*, $\mathbf{P} = \{p_{i,j} \mid i, j = 1, \ldots, J\}$, which expresses the distance between all pairs of sequences, Φ_i and Φ_j. Its calculation is dependent on the choice of the distance measure between facial images and the one between video sequences. In fact, the distance between facial images is used to estimate the proximity between two video frames; commonly, simple measures are preferred like: the Euclidean distance (L_2-norm), the city-block distance (L_1-norm) or the L_0^*-norm. Given two (vectorised) images, $\mathbf{x} \in \Re^M$ and $\mathbf{y} \in \Re^M$, the L_0^*-norm calculates the number of pixel locations that differ more than a predefined threshold value, θ:

$$L_0^*(\mathbf{x}, \mathbf{y}) = \sum_{|x_i - y_i| > \theta} |x_i - y_i|^0$$

Then, the distance between facial images is used in the computation of the distance between video sequences, which inversely represent the proximity measure between two videos. One possible choice is to adopt the *minimal distance*, which is the distance between the two nearest frames:

$$d^{(MIN)}\left(\Phi_i,\Phi_j\right)=\min_{\mathbf{x}\in\Phi_i,\mathbf{y}\in\Phi_j} L\left(\mathbf{x},\mathbf{y}\right)$$

where $L(\ )$ specifies a suitable norm. Otherwise, the *modified Hausdorff distance* can be used:

$$d^{(MHS)}\left(\Phi_i,\Phi_j\right)=\frac{\left(f^{th}_{\mathbf{x}\in\Phi_i}\min_{\mathbf{y}\in\Phi_j} L\left(\mathbf{x},\mathbf{y}\right)\right)+\left(f^{th}_{\mathbf{y}\in\Phi_j}\min_{\mathbf{x}\in\Phi_i} L\left(\mathbf{x},\mathbf{y}\right)\right)}{2}$$

in which $f^{th}_{\mathbf{a}\in\Phi_l}(\ )$ is the f[th] quantile function over the set of frames $\mathbf{a}$ in Φ_l. After that, the distances between video sequences are rearranged into a proximity matrix (or transformed into an equivalent *affinity matrix* [75]).

Once the pair wise relations are estimated, the clustering algorithm partitions data by optimising a local criterion (like a *cluster consistency rule* in [74] or the *structural cluster stability* in [75]), in order to obtain an optimal trade-off between over- and under-segmentation. In particular, the clustering process alternates merging and splitting phases, which enforce the cluster validity conditions on each subset, until a stopping criterion is met.

Finally, person recognition is obtained by considering the clustering result when inserting a new sample (a test video sequence): if it is added to a valid partition, its identity is assumed as the one of that partititon; otherwise the test is rejected or added as a new client.

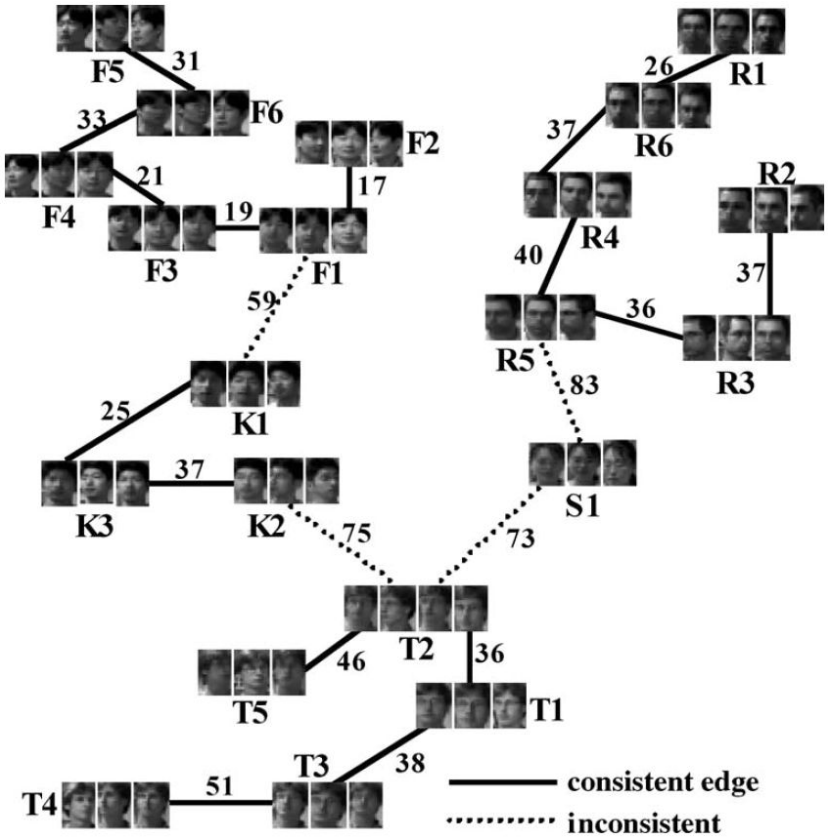

Figure 15: graphical representation ofpair wise clustering applied to person recognition using videos [74]. For nodes, letters specify distinct individuals while numbers indicate different sequences; for edges, the values express distances.

In [74], Raytchev and Murase presented an unsupervised pair wise clustering algorithm, which incrementally built a graph structure by chaining together similar views in video sequences. In its batch version, the clustering graph was formed by firstly computing the *minimal spanning tree* using the distances in the proximity matrix as edge weights. Then, the connecting edges were suitably labelled to specify consistent associations (those that bound patterns in each cluster), and finally the graph structure was optimised by using local statistical information, in order to eliminate spurious associations of different identities (the so called *chaining effect*). An incremental version of the algorithm was also developed, mostly to update the partitioning or to test new videos during the recognition phase.

Raytchev and Murase also developed another unsupervised recognition approach [75], proposing two novel pair wise clustering algorithms based on opposing interaction forces between patterns: attraction and repulsion. One method, called *CAR1*, optimised a local criterion (the *structural cluster stability*) by alternating merging and splitting steps; the other one (*CAR2*) was based on a global criterion, which looked for an optimal balance between attraction and repulsion. The interaction forces between nodes were calculated using a particular affinity matrix, consisting of positive and negative similarity values. The authors compared different approaches on a small video database (33 subjects) with relevant pose and illumination variations: the *CAR1* algorithm obtained the best overall recognition results (over *CAR2*, their previous approach [74], and two concurrent pair wise clustering alternatives) and acceptable clustering quality. On the other hand, all unsupervised recognition strategies performed worse than supervised ones, which have the advantage of exploiting the category information in the enrolment phase.

III.C. Approaches exploiting the temporal information

III.C.1. Discriminant analysis on facial optical flow

The temporal information in video sequences enables the analysis of facial motion and its application as a biometric identifier for person recognition. In fact, it is possible to extract the movement of the face by estimating its *optical flow* (see Figure 16), and then exploit it as a feature for classification. A recognition system based on facial motion has the advantage of being less sensitive to variations in facial appearance, due for example to different illumination conditions, makeup, beard cuts and haircuts. On the other hand, the discriminatory power of this biometric identifier seems inferior to that of facial appearance, maybe because this research domain is quite unexplored and the recognition techniques are still immature.

A typical gradient-based approach for optical flow estimation requires the optimisation of an energy function, which contains one image and one smoothness constraint:

$$E = \iint \left[\left(\frac{\partial \mathbf{\Phi}}{\partial r}\mathbf{U} + \frac{\partial \mathbf{\Phi}}{\partial c}\mathbf{V} + \frac{\partial \mathbf{\Phi}}{\partial t} \right)^2 + \alpha \left(\left|\nabla \mathbf{U}\right|^2 + \left|\nabla \mathbf{V}\right|^2 \right) \right] dr\,dc$$

where $\mathbf{\Phi}_t = \{\phi_{r,c,t}\}$ is the image brightness function for frame t, $(\mathbf{U}_t, \mathbf{V}_t) = \{(u_{r,c,t}, v_{r,c,t})\}$ are the optical flow fields, and α is a weighting parameter between constraints. The optimisation problem can be converted into a linear system of convex-quadratic functions of wavelet scaling coefficients, as explained in [13]. In the end, for each input video the algorithm computes a sequence of optical flow fields: $\{(\mathbf{U}_t, \mathbf{V}_t) \mid t = 1, \ldots, T\}$, where T is the length of the sequence.

A behavioural biometric identifier like facial motion presents some issues which have to be addressed. First of all, there is a need for a temporal segmentation of dynamic sequences, in order to locate and associate similar gestures to be matched. Then these video chunks should be normalised by synchronising their speed and length, in order to provide a common representation for analogous gestures, and therefore being able to calculate commensurate feature vectors. Finally, it is necessary to select which part of facial motion must be retained for computing features for recognition.

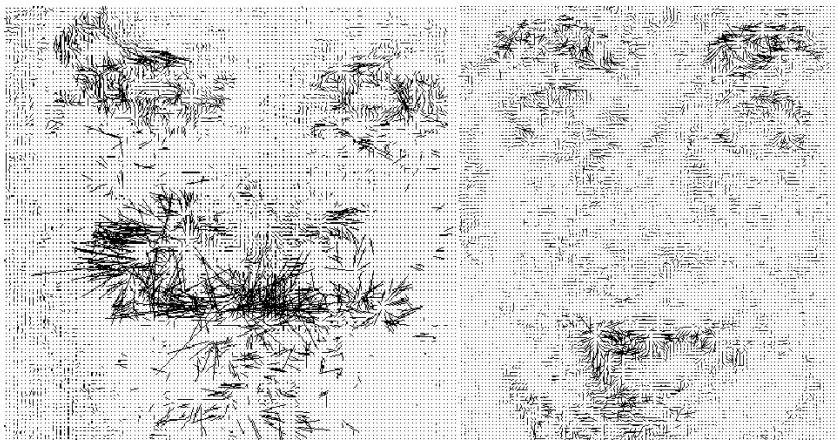

Figure 16: example offacial motion represented using optical flow.

In [14], Chen et al. developed a person recognition system by applying the fisherface approach (Section III.B.2) on a different biometric identifier: facial motion. In fact, for each video their algorithm firstly calculated a sequence of optical flow fields, and subsequently concatenated them (frame by frame) to form a unique high dimensional vector. Then, these motion vectors are projected into a discriminative feature subspace, obtained by applying PCA and LDA on the training dataset. Finally, their system recognised identities by implementing a nearest neighbour classifier working on distances. It is important to notice that in the framework proposed by Chen et al., the issues of temporal segmentation and video chunk normalisation were not explicitly addressed; in fact, their algorithm considered having sequences of commensurate facial motion, so it directly analysed the entire video clips. On the other hand, every frame was semi-automatically pre-processed before the optical flow computation, in order to align head sizes and locations; after that, only the lowest half of optical flow fields was used to extract features for recognition, so that these were mostly related to mouth motion. In the end, this recognition system did not perform better than the original fisherface approach, but it resulted more robust to illumination changes.

III.C.2. Hidden Markov models: extensions to video

Hidden Markov models (HMMs) are a powerful tool to model temporal motion information; for this reason, they have been used in speech recognition [27], gesture and expression recognition. Moreover, HMMs have been successfully applied to person recognition using facial appearance, by spatially associating regions of the face to HMM states; for a detailed review on this research topic, the interested reader can refer to [12], [67] and [96]. In this section, we will focus on those HMM approaches that are modelling temporal information in video data, as shown in Figure 17.

An HMM [50][73] is a statistical model in which the reference system is assumed to be a Markov process with unknown parameters, and the challenge is to determine these *hidden parameters* from observable data, the so called *observations*. More precisely, an HMM is composed by two stochastic processes; one is an unobservable Markov chain with L *states*, $\Omega = \{\omega_l \mid l = 1,\ldots,L\}$, an *initial state probability distribution* , $\boldsymbol{\pi}_0 = [\pi_{0,1},\ldots,\pi_{0,L}]$, and a *state transition probability matrix*,

$$\mathbf{A} = \{a_{i,j} \equiv p(q_t = \omega_j \mid q_{t-1} = \omega_i) \mid 1 \le i,j \le L, 1 \le t \le T\}$$

with constraints $\sum_{j=1}^{L} a_{i,j} = 1$ for $1 \le i \le L$. The second stochastic process is a set of probability density functions, $\mathrm{B} = \{b_l(\mathbf{x}) \mid 1 \le l \le L\}$, of the observation, $\mathbf{x} \in \Re^M$. For a continuous HMM, the probability density function associated with each state l is approximated using a *Gaussian mixture model (GMM)* (see Section IV.B.4):

$$b_l(\mathbf{x}) = \sum_{c=1}^{C} \alpha_{l,c} \aleph(\mathbf{x} \mid \boldsymbol{\mu}_{l,c}, \boldsymbol{\Sigma}_{l,c})$$

where C_l is the number of components, $\alpha_{l,c}$ is the mixture weight for the c-th component, and $\aleph(\mathbf{x} \mid \boldsymbol{\mu}_{l,c}, \boldsymbol{\Sigma}_{l,c})$ is a Gaussian probability density function with mean vector $\boldsymbol{\mu}_{l,c} \in \Re^M$ and covariance matrix $\boldsymbol{\Sigma}_{l,c} \in \Re^{M \times M}$. In short, an HMM can be defined by its parameter set: $\boldsymbol{\Theta} = \{\mathbf{A}, \mathrm{B}, \boldsymbol{\pi}_0\}$.

When applying HMMs to person recognition using facial appearance, researchers have tested different kinds of observations computed from image features: pixel values, eigen-coefficients and *discrete cosine transform (DCT)* coefficients. Though, the use of video data and the application of HMMs on temporal motion information requires a compact representation for the observation vectors; for this reason eigen-coefficients are preferred and *principal component analysis (PCA)* (also called the *Karhunen-Loeve transform (KLT)*) [22] is used for dimensionality reduction of the image space, as it computes a set of orthonormal vectors which optimally represent the distribution of data in the root mean squares sense. In the end, each video frame is projected into an eigenspace (Section III.B.1) and forms a feature vector; we note the sequence of observations related to a given video as: $X = \{\mathbf{x}_t \in \Re^M \mid 1 \le t \le T\}$, where T is the length of the sequence.

During the enrolment phase, each subject is modelled by an L-state fully connected HMM, which learns the statistics of the training sequences and of the temporal dynamics belonging to that individual. The training process [73] estimates the HMM parameter set, Θ_k, for each user k: in the initialisation part, the observations are separated into L classes and a first estimate of the parameter set is obtained; then, the Baum-Welch procedure updates the parameters of the model in order to maximise the resulting likelihood, $p(X \mid \Theta_k)$.

In the recognition step, the sequence of observations of an input video is analysed over time by the HMM of each client. After all, either the likelihood score or the posterior probability score, which are calculated by applying the Viterbi algorithm, are used as similarity measure for recognition.

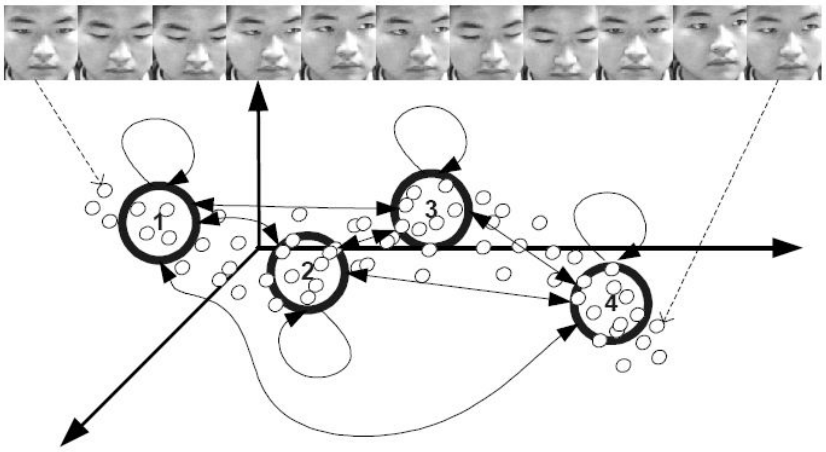

Figure 17: example of a hidden Markov model temporally applied to video sequences [50].

In [33], Huang and Trivedi were the first to develop a person recognition system by using HMMs for modelling the temporal motion information in video sequences; though, their work was not very convincing. In fact, the experimental configuration that obtained the best results employed HMMs with a single state and a single Gaussian component; this framework was equivalent to using a single multidimensional Gaussian approximation, and for this reason the benefits of temporal correlation were lost.

Afterwards, Liu and Cheng [50] successfully applied HMMs for temporal video recognition by improving the basic implementation of Huang and Trivedi. In fact, to avoid singularities on the estimation of the covariance matrices, the authors modified the training algorithm this way: each covariance matrix, $\mathbf{\Sigma}_{j,c} \in \Re^{M \times M}$, was gradually adapted from a global diagonal one (a general model) by using its class-dependent data. Liu and Cheng also proposed an online version of their recognition system, by implementing an adaptive strategy for the HMMs. More precisely, each test sequence successfully recognised was used to update the model parameters of the client in question (except for the covariance matrices), by applying a *maximum a posteriori (MAP)* adaptation technique. In order to discard incorrect or uncertain testing videos, the likelihood difference values were used as confidence measures. In conclusion, the system exploiting adaptive HMMs performed better than the one without adaptation, and both obtained higher recognition scores than the eigenface approach with majority voting (Section III.B.1).

III.C.3. Stochastic tracking and recognition through particle filtering

Stochastic tracking and recognition approaches are based on a unified probabilistic framework, in which individuals are simultaneously tracked and recognised by estimating the posterior probability density function of a *time series state space model (TSSSM)*. Tracking is formulated as a Bayesian inference problem, and it is solved as a probability density propagation problem (due to the temporal nature of tracking itself); recognition is obtained by applying the *maximum a posteriori (MAP)* rule on the posterior probabilities. A TSSSM with non-linear dynamics and non-Gaussian noise model is adopted, whose state and probability estimations are numerically computed using sequential Monte Carlo methods [21][49], in particular the *sequential importance sampling (SIS)* algorithm.

A *time series state space model (TSSSM)* [97] applied to person recognition is governed by three fundamental equations. The *motion equation* defines the kinematic behaviour of the system; in its general form, the tracking motion vector at a given instant t is computed as:

$$\boldsymbol{\theta}_t = G(\boldsymbol{\theta}_{t-1}, \mathbf{u}_t) \text{ for } t > 1$$

where G represents the kinematic function and $\mathbf{u}_t$ is the noise in the motion model, whose distribution determines the motion state transition probabilities, $p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1})$. The tracking motion vector, $\boldsymbol{\theta}_t$, is usually parameterised using an affine motion model, with four deformation parameters and two translation ones; alternatively, 3D parameters can be used. Concerning the kinematic behaviour, a first-order Markov chain is commonly adopted by using an additive function like: $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \mathbf{u}_t$. The *identity equation* determines the temporal evolution of the identity variable, k_t. It is usually assumed constant, so it has the form of: $k_t = k_{t-1}$ for $t > 1$. Accordingly, the identity transition probabilities are simplified as:

$$p(k_t \mid k_{t-1}) = \delta(k_t - k_{t-1})$$

The *observation equation* represents the link between kinematics and identity. By assuming that the transformed observation is a noise-corrupted version of an image template (the user model), then the observation equation is defined as:

$$T_{\boldsymbol{\theta}_t}(\mathbf{z}_t) = \Gamma_{k_t} + \mathbf{v}_t \text{ for } t > 1$$

in which $T_{\boldsymbol{\theta}_t}(\mathbf{z}_t)$ is the transformed version of the observation $\mathbf{z}_t$, Γ_{k_t} is the image template for the identity k_t, and $\mathbf{v}_t$ is the observation noise whose distribution determines the observation likelihood, $p(\mathbf{z}_t \mid \boldsymbol{\theta}_t, k_t)$. The resulting transformation $T_{\boldsymbol{\theta}_t}(\ )$ applied on the observation is composed by a geometric transformation, mostly an affine with parameters $\boldsymbol{\theta}_t$, and a photometric one, like histogram equalisation or zero-mean-unit-variance. The observation likelihood, $p(\mathbf{z}_t \mid \boldsymbol{\theta}_t, k_t)$, is measured in several ways: either by a truncated Laplacian or a truncated Gaussian, either by computing the prediction error using simple metrics, like L_1 (city-block) or L_2 (Euclidean).

Assuming the mutual independence between all noise variables, and the prior knowledge on the distributions of $p(\boldsymbol{\theta}_0 \mid \mathbf{z}_0)$ and $p(k_0 \mid \mathbf{z}_0)$, the goal of the algorithm is to compute the *posterior probability of the identity variable*, $p(k_t \mid \mathbf{z}_0, \ldots, \mathbf{z}_t)$, which is a probability mass function (due to the discrete nature of the identity variable) and is obtained by marginalising the joint posterior probability, $p(\boldsymbol{\theta}_t, k_t \mid \mathbf{z}_0, \ldots, \mathbf{z}_t)$, over $\boldsymbol{\theta}_t$. More precisely, the formula of the posterior probability of the identity variable is:

$$p(k_t = l \mid \mathbf{Z}_{0:t}) = p(l \mid \mathbf{Z}_{0:t}) \int_{\boldsymbol{\theta}_0} \ldots \int_{\boldsymbol{\theta}_t} p(\boldsymbol{\theta}_0 \mid \mathbf{z}_0) \prod_{s=1}^{t} \frac{p(\mathbf{z}_s \mid \boldsymbol{\theta}_s, l) p(\boldsymbol{\theta}_s \mid \boldsymbol{\theta}_{s-1})}{p(\mathbf{z}_s \mid \mathbf{Z}_{0:t})} d\boldsymbol{\theta}_t \ldots d\boldsymbol{\theta}_0$$

where $\mathbf{Z}_{0:t}$ is the compact notation for $\mathbf{z}_0, \ldots, \mathbf{z}_t$.

The numerical solution of the previous theoretical framework is achieved by using the *sequential importance sampling (SIS)* algorithm, which is a particular case of particle filtering that belongs to the general family of sequential Monte Carlo methods [21][49][97]. At each time step, the SIS algorithm approximates the joint probability distribution using a set of weighted particles, and propagates it to the next time step. In the end, the marginal posterior distribution of the identity variable is used as a similarity measure, and the MAP rule provides recognition results.

Figure 18: illustration ofsimultaneous tracking and recognition using particle filtering [98].

In [45], Li and Chellappa were the first to develop a generic approach for stochastic tracking and verification using particle filtering. They implemented a simplified TSSSM with no identity variable, in which only the tracking motion vector was estimated and propagated. They also proposed two facial representations for the observations, $\mathbf{z}_t$: the common intensity images of the face, and an elastic graph matching representation of the facial landmarks (section III.B.5). Unfortunately, they did not provide any evaluation of their techniques.

Then, Zhou et al. [97] improved the approach of Li and Chellappa, by including both the tracking motion vector and the identity variable in the TSSSM. They also considered several observation likelihoods, $p(\mathbf{z}_t \mid \boldsymbol{\theta}_t, k_t)$, and introduced a more complex one by explicitly modelling: the appearance changes within videos using a truncated Laplacian, and the intra-personal appearance variations using a *probabilistic subspace density*, proposed by Moghaddam in [58]. More interestingly, the authors developed a probabilistic learning approach to automatically build user models from video frames. In fact, during the enrolment phase, the algorithm incrementally selected *exemplar* frames of an individual, and used them as mixture centres of a probabilistic distribution for that client. For the recognition phase, they modified the TSSSM and the observation likelihood accordingly, by adding the exemplar variable in the state space model. This last approach obtained the best results and very good identification rates on the small (29 subjects) Motion of Body video database [29].

Successively, Zhou et al. [98] refined their previous recognition system by deriving an adaptive version. They modified the observation likelihood by modelling: the appearance changes within videos using an adaptive appearance model, the intra- and inter-personal appearance variations using a *probabilistic subspace density* [58], and up weighting frontal view frames using another probabilistic subspace density. Then the authors proposed an adaptive motion model, which consisted of: an adaptive velocity model, predicted using a first-order linear approximation, an adaptive noise component, function of the prediction error, and an adaptive number of particles (in the SIS algorithm). Moreover, they included an occlusion handling technique based on robust statistics, which stopped the automatic adaptations during occluded frames. The results obtained by this system were the best of the stochastic approaches reviewed in this section; in fact, this adaptive version achieved perfect tracking and recognition on the small Motion of Body video database [29].

III.C.4. Tracking and recognition using probabilistic appearance manifolds

Historically, tracking and recognition were two independent components of a person recognition system using video data; though, novel strategies have been developed to integrate these tasks into a single framework. One solution is to employ a TSSSM (as detailed in Section III.C.3), otherwise it is possible to simultaneously track and recognise individuals by using the *probabilistic appearance manifold* approach [44]; this technique is an extension to video tracking and recognition of the concept of *appearance manifold*, introduced by Murase and Nayar in [60].

If we consider the complex nonlinear appearance manifold of person k, Ψ_k, then it can be decomposed into a collection of L_k disjoint sub-manifolds:

$$\Psi_k = \left\{ \Gamma_{k,1} \cup \ldots \cup \Gamma_{k,L_k} \right\}$$

Next, each sub-manifold, $\Gamma_{k,l}$, can be approximated using a low dimensional linear subspace, $\Omega_{k,l}$; this subspace can be obtained, for example, by applying the *principal component analysis (PCA)* (also called the *Karhunen-Loeve transform (KLT)*) [22]. In particular, for each client k, the learning algorithm firstly partitions his training video frames into L_k disjoint subsets, by clustering different views of the individual using the K-means algorithm. Then, the images in each subset are considered as samples drawn from each sub-manifold, $\Gamma_{k,l}$, and used to compute its linear approximation, $\Omega_{k,l}$. This way, the appearance model is able to cope with different poses and viewpoints present in videos; nevertheless, other variations like shape and illumination changes are not directly modelled and their occurrences are handled as episodic.

Once the disjoint sub-manifolds are determined, the temporal ordering of video frames is analysed to learn the connectivity relations. In fact, for each individual k, the likelihood $p\left(\Gamma_{k,i} \mid \Gamma_{k,j}\right)$ of observing a transition between sub-manifolds i and j is estimated by counting the actual transitions in the training videos. Afterwards, the transition probabilities are rearranged in a transition matrix, $\mathbf{P}_k \in \Re^{L_k \times L_k}$. Figure 19 illustrates the probabilistic appearance manifold approach: it shows an example of appearance manifold approximation with subspaces, and their relative transition probabilities.

The simultaneous tracking and recognition task, which determines the face location and personal identity for each frame, Φ_t, is formulated as a *maximum a posteriori (MAP)* estimation problem. We consider a tracking parameter vector, $\mathbf{u}_t$, that includes the centre, size and orientation of a rectangular region, and a cropping function, $f\left(\mathbf{u}_t, \Phi_t\right)$, that retrieves the sub-image of frame Φ_t enclosed in the rectangular region defined by $\mathbf{u}_t$. This way, the tracking and recognition result for each frame, $\left(\mathbf{u}_t^*, k_t^*\right)$, is obtained by solving the following optimisation problem:

$$\left(\mathbf{u}_t^*, k_t^*\right) = \arg\min_{\mathbf{u},k} d\left(f\left(\mathbf{u}, \Phi_t\right), \Psi_k\right)$$

where $d(\ )$ is a suitable distance metric between a sub-image and an appearance manifold, Ψ_k.

There are practical difficulties to solve the previous optimisation problem, because the domain of optimisation, $(\mathbf{u},k)$, can be extremely large and there are no closed formulas for applying efficient search strategies based on gradients. For this reason, Lee et al. [44] proposed to minimise each variable independently, by transforming the original formulation into two sub-optimisation problems corresponding to tracking and recognition respectively:

$$\mathbf{u}_t^* = \arg\min_{\mathbf{u}} d\left(f(\mathbf{u},\Phi_t),\Psi_{k_{t-1}^*}\right)$$
$$k_t^* = \arg\min_{k} d\left(f(\mathbf{u}_t^*,\Phi_t),\Psi_k\right)$$

Finally, in order to compute the distance between a sub-image, $\Lambda = f(\mathbf{u}_t,\Phi_t)$, and an appearance manifold, it is necessary to find the point $x_k^* \in \Psi_k$ which is closest to the sub-image. Unfortunately, finding x_k^* can be problematic, because Ψ_k has a very coarse and sparse representation. The distance calculation can be simplified by taking advantage of the linear approximations of the appearance manifold:

$$d(\Lambda,\Psi_k) = \sum_{l=1}^{L_k} p(\Gamma_{k,l} \mid \Lambda) d(\Lambda,\Gamma_{k,l}) \approx \sum_{l=1}^{L_k} p(\Gamma_{k,l} \mid \Lambda) d(\Lambda,\Omega_{k,l})$$

where $p(\Gamma_{k,l} \mid \Lambda)$ is the probability that $\Gamma_{k,l}$ contains a point at minimal distance to Λ, and $d(\Lambda,\Omega_{k,l})$ is the Euclidean distance from the face subspace $\Omega_{k,l}$.

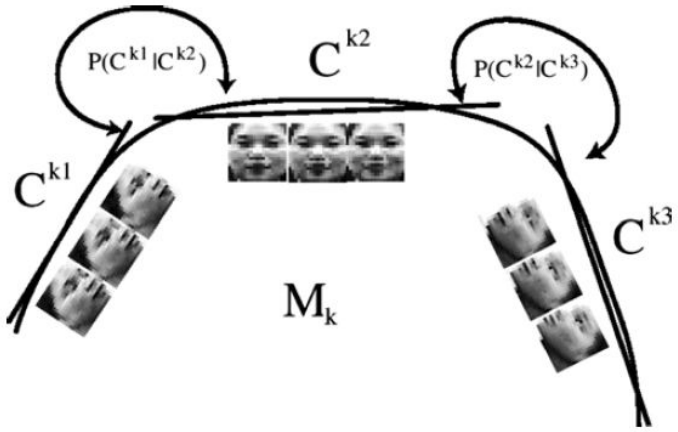

Figure 19: example ofan appearance manifold approximation (M) with subspaces (C), and the relative transition probabilities (P) [44].

In [44], Lee et al. developed the *probabilistic appearance manifold* approach for tracking and recognition using video sequences. The authors applied Bayesian inference to include the temporal coherence of human motion in the distance calculation, $d(\Lambda_t, \Psi_k)$; in fact, they replaced the conditional probability, $p(\Gamma_{k,l} \mid \Lambda_t)$, by using the joint conditional probabilities, $p(\Gamma_{k,l} \mid \Lambda_t, \ldots, \Lambda_0)$, which were recursively estimated using the transitions between sub-manifolds, $p(\Gamma_{k,i} \mid \Gamma_{k,j})$. In the experimental results obtained using a small database (20 individuals), the proposed approach: outperformed standard image-based recognition techniques, showed better robustness and stability than a majority voting strategy or a similar system without temporal coherence, and was able to detect identity changes and to handle large pose variations.

III.D. Concluding summary

In this chapter we proposed a detailed state of the art on person recognition using facial video information. We saw that image-based recognition strategies have been exploiting only the physiological information of the face; in particular its appearance encoded in the pixel values of the images. Next, we emphasised the advantages of person recognition using video sequences compared to image one: the huge amount of data, the presence of the temporal information, the possibility to have more effective representations, and to learn and update user models over time. Then, we classified the existing approaches proposed in the scientific literature between those that neglect the temporal information, and those that exploit it even partially. Concerning the first category, we detailed the extensions to video data of: eigenfaces, fisherfaces, AAMs, RBFNNs, EGM, HDRTs and pair wise clustering methods. After that, we focused on the strategies exploiting the temporal information, in particular those analysing: facial motion with optical flow, or the evolution of facial appearance over time with HMMs or with various probabilistic tracking and recognition approaches.

We conclude this chapter by underlying a few important points. First of all, only recently the attention of the scientific community has been attracted towards the use of facial video information for person recognition. Then, the research on this domain has been mostly focused on developing straightforward extensions of image-based approaches, which exploit only the spatial information in video sequences; furthermore, most of temporal strategies take only advantage of the evolution of facial appearance over time. Finally, the use of the face as a hybrid identifier, for example by exploiting facial appearance and motion for recognition, is still a largely unexplored topic.

Chapter IV. Video person recognition using unconstrained 2D head motion

IV.A. Introduction

Our study on the literature related to person recognition approaches using facial video sequences (Chapter III) has revealed that there are really few works exploiting the temporal information, and that none of them is using the unconstrained head motion as a biometric identifier; in fact, apart from the recognition approach based on facial motion (Section III.C.1), all other techniques take advantage of the evolution of facial appearance over time. These elements encouraged us to explore the use of this neglected video temporal information, by proposing a strategy that exploits the unconstrained head motion for person recognition, and we have found that this information possesses enough discriminatory power to be considered as a valuable biometric for the development of new applications.

We decided to focus on the head motion information, because we believe that the way an individual moves his head is somewhat characteristic, and that the dynamic patterns could be used to discriminate people. We are supported in this claim by the study of Knight and Johnston [41], which revealed that under non-optimal image conditions (like negative images) "moving faces are significantly better recognised than still faces". Moreover, we chose to exploit natural head motion in an unconstrained recognition scenario, because generating a personal moving signature with the head seemed awkward and impractical, and most of real video data is in an unconstrained format, like in video surveillance applications. In other words, our system has no prior knowledge on the explicit gestures that each user is doing in every sequence, like there are no defined passphrases in a text-independent speaker recognition application. Hence, to solve the issue of temporal synchronisation between unpredictable gestures, we learn the personal motion information using *Gaussian mixture models (GMMs)*, which are well suited to our unconstrained recognition scenario.

In our biometric system, the head motion information is extracted by tracking a few facial landmarks in the 2D image plane. It is a convenient choice for our experimental setup, where there is no camera motion, there are no zooms or changes in scale, and that the depth variation due to the in-depth movement of the user is insignificant, because the camera is far. Therefore, all motion information that can be extracted through a tracking in the image plane is relative to the behaviour of the individuals. In addition, instead of computing a dense optical flow field, which is computationally expensive, we take advantage of the fact that the head is a semi-rigid object, and we represent its motion by following a few facial landmarks in video frames. We preferred this solution rather than a 3D head tracking from video sequences, because the latter is a complicate and computational expensive process and it is not required in our experimental scenario. In fact, for tracking in the 3D space we should have: dealt with the generation of the 3D models from video data, solved the registration problem, estimated the 3D pose from video frames, and removed any facial deformation to improve the matching accuracy. Hence, considering the complexity of 3D head tracking, we preferred to concentrate our efforts on the feature extraction and classification steps, by developing our simpler tracking strategy in the 2D image plane.

The remainder of this chapter is organised in two main sections: one theoretical part that details the structure of our person recognition system using head motion, and one experimental part that thorough fully evaluates the performances of our approach in various conditions.

IV.B. Proposed method

The architecture of the person recognition approach using unconstrained 2D head motion is illustrated in Figure 20, and closely resembles the one for the general biometric system, which has been introduced in Section II.D.

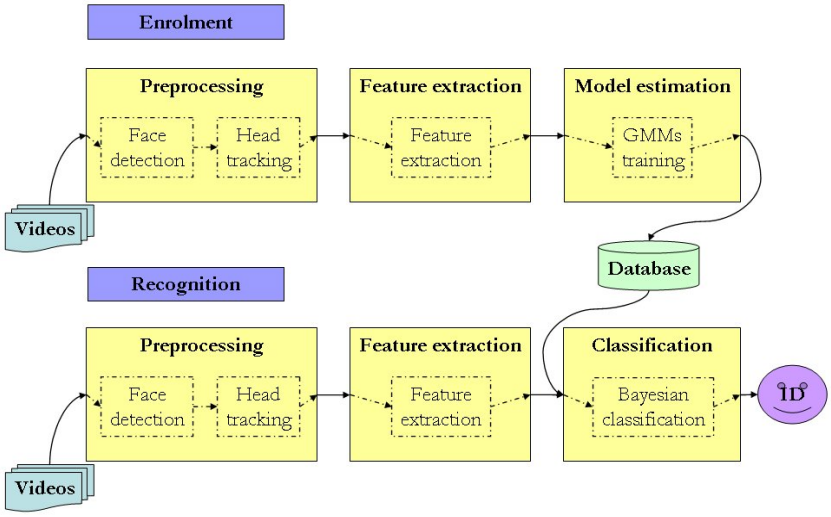

Figure 20: architecture ofthe person recognition system that exploits unconstrained head motion.

A video sequence is firstly pre-processed, by detecting and tracking a few facial landmarks of interest over time, in order to recover the global 2D head motion information of the individual. Those tracking signals are then normalised and transformed into features that provide a better discriminative representation. After that, the enrolment module estimates each client model by using a *Gaussian mixture model (GMM)* approximation; in the end, full person recognition (both identification and verification) is achieved through *Bayesian decision* (also called *Bayesian inference*). The five main steps of our system are detailed in the following sections.

IV.B.1. Pre-processing: face detection

The face detection step is semi-automatic: a *graphical user interface (GUI)* displays the first frame of each video and an operator must click on the facial landmarks of interest, which are selected and located for tracking. In fact, the face detection step chooses F facial landmarks, and then computes their $2F$ coordinate values (Cartesian coordinates) that are stored in the first tracking vector:

$$\mathbf{s}_1 = \left[r_{1,1}, c_{1,1}, \ldots, r_{F,1}, c_{F,1}\right]^T \in \mathrm{N}^{2F}$$

where $r_{f,1}$ and $c_{f,1}$ are respectively the vertical (row) and horizontal (column) components of the f -th landmark.

We prefer to have an active human interaction in the face detection process, in order to guarantee a perfect initialisation for the tracking step by precisely locating the facial landmarks of interest. In fact, the automatic face detection approaches proposed in literature [30] still demonstrate significant wrong detection and false positive rates, which can greatly influence our recognition system; we are comforted in our choice by the experimental evaluation proposed in Section IV.C.5, about the importance of tracking accuracy on final recognition scores.

IV.B.2. Pre-processing: head tracking

Given the locations of the F facial landmarks in the first frame, $\mathbf{s}_1 \in \mathrm{N}^{2F}$, a fully automatic tracking algorithm traces these reference points until the end of the sequence. More precisely, for each frame, Φ_t, the head tracker estimates the 2D image locations of the selected landmarks, $\mathbf{s}_t \in \mathrm{N}^{2F}$, which are concatenated one after another to form the tracking matrix:

$$\mathbf{S} = \left[\mathbf{s}_1, \ldots \mathbf{s}_T\right] \in N^{2F \times T}$$

where T is the video length (defined as the total number of frames). As a result, each column of the tracking matrix contains the coordinates of all the landmarks in a given frame; in contrast, each row of $\mathbf{S}$ denotes the vertical or horizontal tracking signal for a given landmark. A visual example of the tracking signals over time can be seen in Figure 21.

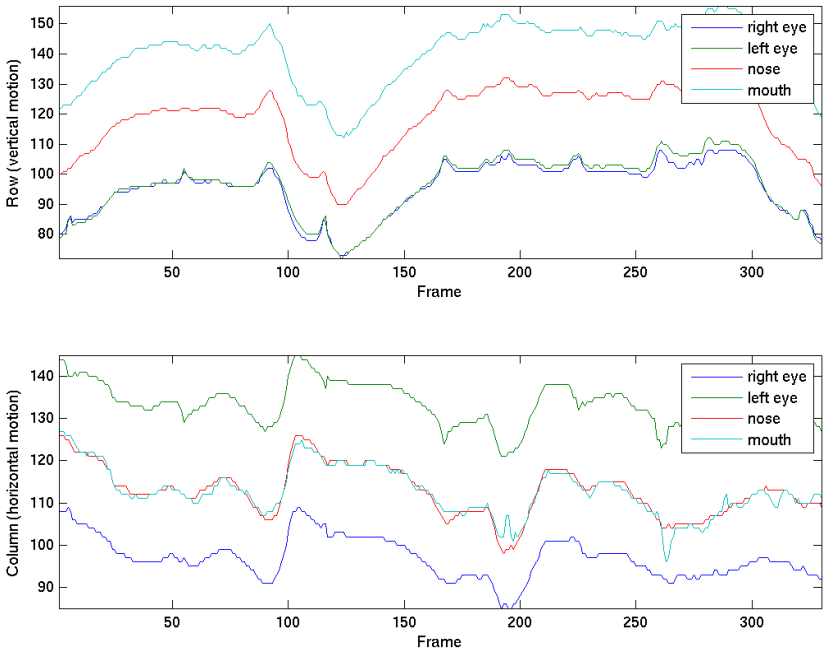

Figure 21: example of the tracking signals over time.

The tracking of each facial landmark is based on the principle of *template matching*: after having created a sub image, called the *template*, the algorithm moves the template over each allowed position in a candidate image, and computes a *similarity score* between the template and the region corresponding to that position. At last, the region (or position) that reveals the highest similarity with the template is returned as the best match.

In our implementation of the template matching strategy, the similarity score in the t-th frame is calculated by adding up the similarity values of the distinct RGB colour components:

$$S(\Lambda_t, \Gamma_t) \equiv \sum_{i=1}^{3} S(\Lambda_{t,i}, \Gamma_{t,i})$$

where $\Lambda_{t,i}$ is the i-th colour channel of a candidate region Λ_t, and $\Gamma_{t,i}$ is the i-th colour channel of the template Γ_t. Then, the similarity scores of the individual colour components are computed by taking the negative value of one among these simple distance metrics:

- *City-block distance* (L_1): $d^{(L_1)}\left(\Lambda_{t,i},\Gamma_{t,i}\right)\equiv\sum_{r=1}^{R}\sum_{c=1}^{C}\left|\lambda_{t,i,r,c}-\gamma_{t,i,r,c}\right|$.
- *Euclidean distance* (L_2): $d^{(L_2)}\left(\Lambda_{t,i},\Gamma_{t,i}\right)\equiv\sqrt{\sum_{r=1}^{R}\sum_{c=1}^{C}\left(\lambda_{t,i,r,c}-\gamma_{t,i,r,c}\right)^2}$.
- *Cosine distance*: $d^{(\cos)}\left(\Lambda_{t,i},\Gamma_{t,i}\right)\equiv\frac{1}{2}-\frac{\sum_{r=1}^{R}\sum_{c=1}^{C}\lambda_{t,i,r,c}\gamma_{t,i,r,c}}{2\sqrt{\left(\sum_{r=1}^{R}\sum_{c=1}^{C}\lambda_{t,i,r,c}^2\right)\left(\sum_{r=1}^{R}\sum_{c=1}^{C}\gamma_{t,i,r,c}^2\right)}}$.

We also extend the template matching approach by including a simple *template update* strategy: the present template, Γ_t, is calculated by a weighted sum (colour component by colour component) of the initial template, Γ_1, and all previous best matches, Λ_t^*:

$$\Gamma_t\equiv\alpha\Lambda_{t-1}^*+(1-\alpha)\Gamma_{t-1}=\alpha\sum_{j=1}^{t-1}(1-\alpha)^{t-1-j}\Lambda_j^*+(1-\alpha)^{t-1}\Gamma_1$$

for $t=2,\ldots,T$, in which $\alpha\in[0,1]$ is a weighting constant. It is easy to notice that this template update strategy includes the extreme cases of:

- *No update*, when $\alpha=0$ and $\Gamma_t=\Gamma_1$ for $\forall t$.
- *Full update*, when $\alpha=1$ and $\Gamma_t=\Lambda_{t-1}^*$ for $\forall t$.

In our implementation, we massively reduce the computational load of the head tracking step by constraining the search for the best match to a small neighbourhood. In fact, for each facial landmark we take advantage of its spatio-temporal continuity in consecutive frames, and we restrict the search space to a small window centred on the previous best match.

Finally, in order to improve the robustness of the tracking and to reduce the impact of intra-video illumination and colour variations, we pre-process each sequence by applying a *histogram equalisation* or a *contrast stretching* to each frame (colour component by colour component) [28].

IV.B.3. Feature extraction

The feature extraction step isolates the discriminative information that characterise the individual and discards the irrelevant one; it can be considered as a nonlinear transformation, $f(\ )$, applied on a tracking matrix, $\mathbf{S}\in N^{2F\times T}$:

$$\mathbf{X} = f(\mathbf{S})$$

where $\mathbf{X} \in \Re^{D \times N}$ is the resulting feature matrix, composed by N feature vectors of dimension D, $\mathbf{x}_n \in \Re^D$, typically indexed at discrete time n.

In our implementation, the feature extraction step is divided in two phases: the geometrical normalisation of the tracking signals, and the calculation of the feature vectors. The former part centres and scales the tracking signals; this way, after clearing the features from any dependence on absolute head location and size, the head motion information is isolated and the inter-video variation is reduced. We adopt one among these geometrical normalisations:

- *Centring using zero mean*: each tracking signal is centred on its average position by the following transformation: $v_{f,t} = s_{f,t} - \mu_f$ for $f = 1,\ldots,2F$ and $t = 1,\ldots,T$, where μ_f is the mean value of the f-th signal, $\mu_f = \frac{1}{T}\sum_{t=1}^{T} s_{f,t}$.

- *Centring using zero mean and scaling by imposing a unit variance*: each tracking signal is centred on its average position and the range of each signal is normalised to have a unit variance by the following transformation: $v_{f,t} = \frac{s_{f,t} - \mu_f}{\sigma_f}$ for $f = 1,\ldots,2F$ and $t = 1,\ldots,T$, where μ_f is the mean value and σ_f is the standard deviation, $\sigma_f = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left(s_{f,t} - \mu_f\right)^2}$.

- *Centring using zero mean and scaling depending on head size*: each tracking signal is centred on its average position, and its range is normalised based on the average eye distance in each video shot. This constraint on uniform eye distance conveys more discriminative information than the previous one on uniform variance; however, it is less robust to noise because of its dependence on the precision of the tracked signals (of the eyes).

The feature matrix, $\mathbf{X} \in \Re^{D \times N}$, is generated by concatenating one or more of the following distinct features:

- *Head positions*: the location of the head over time is included using the normalised tracking signals: $x_{d,n} = v_{f,t}$ for $f = 1,\ldots,2F$ and $t = 1,\ldots,T$.

- *Velocities*: the velocity of the head over time is calculated by taking the first derivatives of the normalised tracking signals: $x_{d,n} = v_{f,t} - v_{f,t-1}$ for $f = 1,\ldots,2F$ and $t = 2,\ldots,T$ (the border effect is solved as: $x_{d,1} = x_{d,2}$ for $\forall d$).

- *Accelerations*: the acceleration of the head over time is calculated by taking the second derivatives of the normalised tracking signals: $x_{d,n} = v_{f,t+1} - 2v_{f,t} + v_{f,t-1}$ for $f = 1,\ldots,2F$ and $t = 2,\ldots,T-1$ (the border effects are solved as: $x_{d,1} = x_{d,2}$ and $x_{d,T} = x_{d,T-1}$ for $\forall d$).

In our study we tested other features, computed using polar coordinates (like head positions and its derivatives) or the frequency domain (like spectral energies), but they empirically showed less discriminatory power and were abandoned. After the concatenation of one or more of the previous parameters, the number of feature vectors is equal to the video length, $N = T$. On the other hand, the dimension of the feature space depends on the concatenation strategy adopted; for example, when using only one type of features (like head positions) we have: $D = 2F$.

Finally, the feature space can be reduced by applying the *principal component analysis (PCA)* (also called the *Karhunen-Loeve transform (KLT)*) [22] (refer to section III.B.1) to all vectors in the feature matrix, $\mathbf{X} \in \Re^{D \times N}$. In fact, the dimensionality of the feature space D is an important parameter for the training of GMMs, and in some cases it may be convenient to reduce it, as explained in the next section (IV.B.4).

IV.B.4. Model estimation: GMM training

In order to register new users in our recognition system it is necessary to characterise their *personal models* (also called *class models*), by using the features extracted from the enrolment data set. For this purpose, we adopt a probabilistic approach that estimates the distribution of feature vectors of each client in the feature space; in other words, for each individual (or class), k , we aim to represent his class conditional *probability density function (PDF)* of feature vectors: $p(\mathbf{x}_n \mid k)$. It is worth noting that finding a proper PDF is a crucial task and can have a critical impact on recognition results.

As a result, we decide to approximate each class conditional PDF by employing *finite mixture models*, in particular *Gaussian mixture models (GMMs)*. First of all, GMMs have been frequently used as a generic probabilistic model for approximating multivariate densities, and are capable of representing arbitrary densities. Moreover, GMMs can be well suited to our unconstrained recognition problem, in which there is no prior knowledge on user motion, because they are intrinsically unconstrained. In fact, GMMs estimate only the underlying distribution of motion features, and are insensitive to the temporal synchronisation between different gestures; though, this unconstrained nature is also a disadvantage because the higher levels of information, like the knowledge of each gesture, are ignored. Finally, GMMs are computationally inexpensive and are based on a well understood statistical framework.

A GMM is a finite mixture model of Gaussian distributions (also called normal distributions). A *non-singular multivariate normal distribution* of a random variable, $\mathbf{x} \in \Re^D$, is defined as:

$$\aleph(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \frac{1}{(2\pi)^{D/2}\sqrt{|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

where $\boldsymbol{\mu} \in \Re^D$ is the mean vector, and $\boldsymbol{\Sigma} \in \Re^{D \times D}$ is the non-singular covariance matrix.

Then, a *Gaussian mixture model probability density function (GMM-PDF)* is a weighted sum of C normal distributions:

$$p(\mathbf{x} \mid \boldsymbol{\Theta}) \equiv \sum_{c=1}^{C} \alpha_c \aleph(\mathbf{x} \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

in which $\boldsymbol{\Theta} = \{\alpha_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c \mid c = 1, \ldots, C\}$ is the parameter list, and $\alpha_c \in [0,1]$ is the weight of the c-th Gaussian component. In addition, each α_c corresponds to the *a priori probability* that an observation $\mathbf{x}$ has been generated by the c-th normal source, and its value is normalised such as: $\sum_{c=1}^{C} \alpha_c \equiv 1$. In a GMM modelling, the total number of Gaussian components C does not need to be guessed accurately: it is just a parameter defining the complexity of the approximating distribution. However, if C is too small, there is not an adequate amount of components to learn the feature distribution precisely enough; on the other hand, when C is too large the modelling is excessively complex: this may lead either to an over fitted classifier, either to singularities in the covariance matrices once the amount of training data becomes insufficient. An example of GMM approximation and its equiprobability surfaces is illustrated in Figure 22.

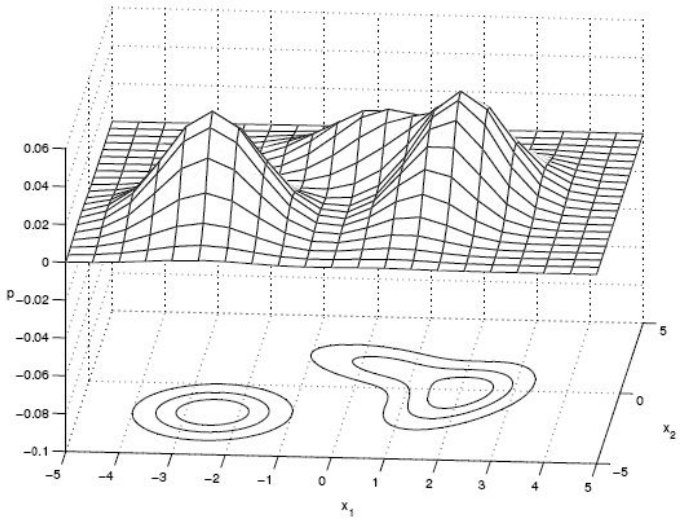

Figure 22: example ofGaussian mixture model (GMM) approximation and its equiprobability surfaces.

If we assume statistical independence between the K classes that correspond to the clients of our system, then the overall estimation of GMM parameters can be divided into K separate estimation problems. Hence, for each client k, his model parameters $\boldsymbol{\Theta}_k$ are obtained by solving a *maximum likelihood problem* through the *expectation-maximisation (EM)* algorithm [5][22][66]. To simplify the notation, for the rest of this section we will focus on a single estimation problem and we will drop the index k on clients and classes.

We consider having a set of N training feature vectors, $X = \{\mathbf{x}_n \in \Re^D \mid n = 1,\ldots,N\}$, that are identically distributed because supposedly drawn from a common distribution $p(\mathbf{x}_n \mid \boldsymbol{\Theta})$. To evaluate the quality of the modelling, we define the *(incomplete-data) likelihood function* as:

$$L(\boldsymbol{\Theta} \mid X) \equiv \prod_{n=1}^{N} p(\mathbf{x}_n \mid \boldsymbol{\Theta})$$

which represents the likelihood of the parameters $\boldsymbol{\Theta}$, given the training data X. It is worth noting that the likelihood $L(\boldsymbol{\Theta} \mid X)$ is a function of the parameters, while the data is fixed. We also define the *(incomplete-data) log-likelihood function,* because it is computationally more practical:

$$M(\boldsymbol{\Theta} \mid X) \equiv \ln L(\boldsymbol{\Theta} \mid X) \equiv \sum_{n=1}^{N} \ln p(\mathbf{x}_n \mid \boldsymbol{\Theta})$$

The optimal parameter set $\boldsymbol{\Theta}^*$ is obtained by maximising the likelihood function or the log-likelihood one:

$$\boldsymbol{\Theta}^* = \arg\max_{\boldsymbol{\Theta}} L(\boldsymbol{\Theta} \mid X) \cong \arg\max_{\boldsymbol{\Theta}} M(\boldsymbol{\Theta} \mid X)$$

In fact, due to the monotonicity property of the logarithm function, it is theoretically equivalent to maximise $L(\,)$ or $M(\,)$. Unfortunately, the analytical approach for solving the maximum likelihood problem is intractable for GMMs with unknown and unrestricted covariance matrices and means; the solution is then to apply an optimisation strategy, such as the *expectation-maximisation (EM)* algorithm.

The EM algorithm is a general iterative method that calculates the maximum likelihood estimate of the parameters of an underlying distribution from a given data set, X, when the data is incomplete or has missing values. For finite mixture models as GMMs, the optimisation of the likelihood function is analytically intractable, unless we assume the existence of values for additional but missing (or hidden) parameters, Y. If we consider the training feature set X, which is called the *incomplete data set* and is generated by some distribution $p(\mathbf{x}_n \mid \boldsymbol{\Theta})$, and the *hidden data set* Y, then we can assume that a *complete data set* exists, $Z = \{X, Y\}$, supposedly drawn from a joint density function $p(\mathbf{z}_n \mid \boldsymbol{\Theta})$. With this new PDF, we can define the *complete-data log-likelihood* function:

$$M(\boldsymbol{\Theta} \mid Z) \equiv M(\boldsymbol{\Theta} \mid X, Y) \equiv \ln L(\boldsymbol{\Theta} \mid X, Y) \equiv \sum_{n=1}^{N} \ln p(\mathbf{z}_n \mid \boldsymbol{\Theta})$$

Furthermore, $M(\boldsymbol{\Theta} \mid X, Y)$ is a random variable, because the missing information Y is: unknown, random and probably governed by an underlying distribution.

The *expectation step (E-step)* of the EM algorithm finds the expected value of the complete-data log-likelihood, with respect to the unknown data Y, given the observed data X, and the current parameter estimates $\boldsymbol{\Theta}^{(i-1)}$:

$$Q\left(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(i-1)}\right) = E_Y\left[M(\boldsymbol{\Theta} \mid X, Y) \mid X, \boldsymbol{\Theta}^{(i-1)}\right] = E_Y\left[\ln p(X, Y \mid \boldsymbol{\Theta}) \mid X, \boldsymbol{\Theta}^{(i-1)}\right]$$

In this formula, the expectation makes $Q\left(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(i-1)}\right)$ a deterministic function that can be maximised; in fact, X and $\boldsymbol{\Theta}^{(i-1)}$ are constants, $\boldsymbol{\Theta}$ is a standard variable, and Y is a random variable but it is marginalised by the expectation.

The *maximisation step (M-step)* of the EM algorithm maximises the expectation with respect to $\boldsymbol{\Theta}$:

$$\boldsymbol{\Theta}^{(i)} = \arg\max_{\boldsymbol{\Theta}} Q\left(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(i-1)}\right)$$

The two previous steps are repeated until a *stopping criterion* is met, which is generally based on absolute or relative improvements of Q and $\boldsymbol{\Theta}$, and the total number of iterations. In fact, the EM algorithm is guaranteed to increase the log-likelihood value at each iteration, until it converges to a local maximum of the likelihood function, but it can eventually lead to singular estimates of the covariance matrices.

The EM algorithm has to be initialised in some way, since it starts from an early guess of the model parameters, $\boldsymbol{\Theta}^{(0)}$. It is an important step, because the choice of $\boldsymbol{\Theta}^{(0)}$ determines where the algorithm converges, or hits the boundary of the parameter space producing singular meaningless results. Some solutions for the initialisation use multiple random starts or a clustering algorithm like the *K-means* or the *fuzzy K-means* [4].

When choosing GMMs as finite mixture models, the missing information, $Y = \{y_n \mid n = 1, \ldots, N\}$, is the knowledge of which component produced each feature vector $\mathbf{x}_n$; in other words, $y_n = c$ if $\mathbf{x}_n$ has been generated by the c-th Gaussian component. Then, the complete-data log likelihood for GMMs becomes:

$$M(\boldsymbol{\Theta} \mid X, Y) \equiv \sum_{n=1}^{N} \ln\left[\alpha_{y_n} \aleph\left(\mathbf{x}_n \mid \boldsymbol{\mu}_{y_n}, \boldsymbol{\Sigma}_{y_n}\right)\right]$$

Substituting this expression into the general EM formulation, and after some calculations [5][66], we obtain the *EM equations relative to GMMs*. For the E-step:

$$w_{n,c}^{(i)} \equiv p\left(y_n = c \mid \mathbf{x}_n, \boldsymbol{\Theta}^{(i)}\right) = \frac{\alpha_c^{(i)} \aleph\left(\mathbf{x}_n \mid \boldsymbol{\mu}_c^{(i)}, \boldsymbol{\Sigma}_c^{(i)}\right)}{\sum_{c=1}^{C} \alpha_c^{(i)} \aleph\left(\mathbf{x}_n \mid \boldsymbol{\mu}_c^{(i)}, \boldsymbol{\Sigma}_c^{(i)}\right)}$$

which is the a posteriori probability that $y_n = c$ after having observed $\mathbf{x}_n$ (or equivalently the probability that $\mathbf{x}_n$ has been generated by the c-th component). Next, the M-step equations:

$$\alpha_c^{(i+1)} = \frac{1}{N}\sum_{n=1}^{N} w_{n,c}^{(i)}$$

$$\boldsymbol{\mu}_c^{(i+1)} = \frac{\sum_{n=1}^{N} w_{n,c}^{(i)} \mathbf{x}_n}{\sum_{n=1}^{N} w_{n,c}^{(i)}}$$

$$\boldsymbol{\Sigma}_c^{(i+1)} = \frac{\sum_{n=1}^{N} w_{n,c}^{(i)} \left(\mathbf{x}_n - \boldsymbol{\mu}_c^{(i+1)}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_c^{(i+1)}\right)^T}{\sum_{n=1}^{N} w_{n,c}^{(i)}}$$

The initialisation of the EM algorithm for the estimation of the GMM parameters is done in two phases. Firstly, the training data is clustered into C partitions, by applying the *K-means* method or the *fuzzy K-means* [4] one. After that, the initial parameter set, $\boldsymbol{\Theta}^{(0)} = \left\{\alpha_c^{(0)}, \boldsymbol{\mu}_c^{(0)}, \boldsymbol{\Sigma}_c^{(0)} \mid c = 1, \ldots, C\right\}$, is calculated by: taking the cluster means, uniform or cluster covariance matrices, and uniform or cluster weights.

We also need to guess the minimum number of training feature vectors, $N^{(\min)}$, that are recommended for a reliable estimation of the GMM parameters. Firstly, we recall that the *number of free parameters in a GMM*, with C Gaussian components and D-dimensional real feature vectors $\mathbf{x}_n \in \Re^D$, is:

$$\eta = C * \left(\frac{1}{2}D^2 + \frac{3}{2}D\right) + C - 1$$

Then, as a rule of thumb, we empirically require a minimum number of feature vectors as in [66]: $N^{(\min)} > 3\eta$. It is worth noting that the number of recommended vectors increases linearly with the number of Gaussian components (the complexity of the modelling), and quadratically with the dimensionality of the feature space.

Finally, in addition to the standard EM algorithm, we implemented two variants for the estimation of GMM parameters: the *Figueiredo-Jain algorithm* and the *Greedy EM algorithm*. The Figueiredo-Jain algorithm [26] automatically adjusts the number of components, by annihilating those that are not supported by the data or are becoming singular. This way, it better avoids the boundary of the parameter space and can start with an arbitrary number of initial components. Alternatively, the Greedy EM algorithm [90] begins with a single Gaussian and then adds components into the mixture one by one. It basically repeats two steps: it inserts the component that mostly increases the likelihood into the mixture, and then it runs the EM algorithm to update the parameters.

IV.B.5. Classification: Bayesian classification

The classification task of our system is achieved by applying the probability theory and the *Bayesian decision rule* (also called *Bayesian inference*) [66], so that the classifier chooses the most probable class, or equivalently the option with the lowest risk (expected cost).

In our framework, we remember that a given test is represented by a video sequence. Then, we aim to compute the *video posterior probability*, $p(k \mid \mathbf{X})$, which we define as the probability that all feature vectors extracted from a video $\mathbf{X} \in \Re^{D \times N}$ belong to class k:

$$p(k \mid \mathbf{X}) \equiv p(k \mid \mathbf{x}_1, \ldots, \mathbf{x}_N)$$

By applying the *Bayes' rule*, the posterior probability $p(k \mid \mathbf{X})$ becomes:

$$p(k \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid k)p(k)}{p(\mathbf{X})} = \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_N \mid k)p(k)}{p(\mathbf{x}_1, \ldots, \mathbf{x}_N)}$$

First of all, the divisor:

$$p(\mathbf{X}) \equiv p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \sum_{k=1}^{K} p(\mathbf{x}_1, \ldots, \mathbf{x}_N \mid k)p(k) = M_{\mathbf{X}}$$

is merely a scaling factor $M_{\mathbf{X}}$, to assure that the posterior probabilities $p(k \mid \mathbf{X})$ are really probabilities (their sum is one). Hence, we can simplify the previous expression as:

$$p(k \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid k)p(k)}{M_{\mathbf{X}}} = \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_N \mid k)p(k)}{M_{\mathbf{X}}}$$

Afterwards, the *a priori probability* $p(k)$ represents the probability of occurrence of each class k, and it is usually estimated from the training database. Finally, in order to calculate the video posterior probability $p(k \mid \mathbf{X})$, we have to express the joint class conditional PDF $p(\mathbf{X} \mid k)$ as a function of the class conditional PDFs of feature vectors $p(\mathbf{x}_n \mid k)$, which are our user models estimated during the enrolment. This task can be problematic, unless we assume that the feature vectors $\mathbf{x}_n$ are independent from each other; this way, the joint class conditional PDF $p(\mathbf{X} \mid k)$ takes the form of:

$$p(\mathbf{X} \mid k) \equiv p(\mathbf{x}_1, \ldots, \mathbf{x}_N \mid k) \cong \prod_{n=1}^{N} p(\mathbf{x}_n \mid k)$$

and the video posterior probability becomes:

$$p(k \mid \mathbf{X}) \cong \frac{p(k)}{M_{\mathbf{X}}} \prod_{n=1}^{N} p(\mathbf{x}_n \mid k)$$

The *similarity score for the identification task* $S^{(ID)}(\mathbf{X}, \boldsymbol{\Theta}_k)$ (Section II.C.2), is derived from the video posterior probability $p(k \mid \mathbf{X})$ by computing the *log-posterior probability*, because it is analytically and numerically more practical, and the properties of the similarity function do not change thanks to the monotonicity of the logarithm. Hence, $S^{(ID)}(\mathbf{X}, \boldsymbol{\Theta}_k)$ takes the form of:

$$S^{(ID)}(\mathbf{X}, \boldsymbol{\Theta}_k) = \ln p(k \mid \mathbf{X}) = \sum_{n=1}^{N} \ln p(\mathbf{x}_n \mid k) + \ln p(k) - \ln M_{\mathbf{X}}$$

Finally, the *similarity score for the verification task* $S^{(VER)}(\mathbf{X}, \boldsymbol{\Theta}_k)$ (Section II.C.1), is the *log-posterior probability ratio*:

$$S^{(VER)}(\mathbf{X}, \boldsymbol{\Theta}_k) = \ln\left[\frac{p(k \mid \mathbf{X})}{p(\bar{k} \mid \mathbf{X})}\right] = \sum_{n=1}^{N} \ln p(\mathbf{x}_n \mid k) - \sum_{n=1}^{N} \ln p(\mathbf{x}_n \mid \bar{k}) + 2 \ln p(k) - 1$$

where $p(\bar{k} \mid \mathbf{X})$ is the posterior probability of the *alternative hypothesis* $\bar{k}$, and $p(\mathbf{x}_n \mid \bar{k})$ is the *impostor model* (the class conditional PDF for $\bar{k}$). In other words, $p(\bar{k} \mid \mathbf{X})$ expresses the probability that all feature vectors extracted from a video $\mathbf{X} \in \Re^{D \times N}$ do not belong to class k, and $p(\mathbf{x}_n \mid \bar{k})$ represents the probability that the alternative hypothesis $\bar{k}$ can generate $\mathbf{x}_n$.

Unfortunately, the estimation of the impostor model $p(\mathbf{x}_n \mid \bar{k})$ is usually problematic, because it should represent the space of all possible alternatives to k, which is huge and requires a massive amount of training data. Inspired by the *speaker verification* domain and the work of Rosenberg et al. [77], we approximate the impostor model by using the set of other client models $p(\mathbf{x}_n \mid k)$, which are called *background models* or *cohorts*. More precisely, $p(\mathbf{x}_n \mid \bar{k})$ is estimated by taking the average of the L best client models on a given test (a video in our case):

$$p(\mathbf{x}_n \mid \bar{k}) \cong \frac{1}{L} \sum_{l=1}^{L} p(\mathbf{x}_n \mid k^{(l)})$$

where $k^{(l)}$ is the client model that produces the l-th highest video posterior probability $p(k \mid \mathbf{X})$.

IV.C. Experimental results

Due to the absence of standard video databases suited for our approach, we assess the performance of our person recognition system on our video database of Italian TV speakers: please refer to Section VIII.A for a discussion on existing data sets, a description of our database, and the structure of the enrolment and recognition subsets. It is worth noting that all experimental results and relative comments are related to our small video database of Italian TV speakers, so that they should not be considered as absolute general conclusions.

In the following sections, we firstly introduce the *default configuration*, which obtains the best recognition results overall. Next, we evaluate the precision of the tracking and the discriminative power of our system in different experimental conditions, by varying: signals, features and GMM estimations. After that, we compare our results with the state of the art eigenface technique, and we analyse the degradation of performance due to inaccurate and noisy tracking. Finally, we also evaluate the discriminatory power of our method in a gender recognition application.

IV.C.1. Default configuration

We denote the parameter configuration that attains the best overall recognition performance as the *default configuration*, and we use it as a reference throughout the experiments; a summary of the parameters for the default configuration of the recognition system using head motion is presented in Table 1.

PRE-PROCESSING	Facial landmarks	4 (eyes, nose & mouth)
	Colour space	RGB (red, green & blue)
	Video pre-processing	Histogram equalisation
	Distance metric for similarity scores	City-block distance
	Template size	19 pixel rows (or height)
		25 pixel columns (or weigth)
	Search window	12 x 12 pixels
	Template update	None ($\alpha = 0$)
FEATURE EXTRACTION	Geometrical normalisation of tracking signals	Centring using zero mean
	Head features	Normalised head positions
	PCA reduction	No
	Dimensionality of feature space	8
MODEL ESTIMATION	GMM parameter estimation	Expectation-maximisation (EM)
	Gaussian components	4
	Initialisation	K-means
		Uniform weights
		Cluster means
		Uniform covariances
CLASSIFICATION	Number of background (cohort) models	2

Table 1: summary of the parameters for the default configuration of the recognition system using head motion.

In the default configuration, the head motion of each individual is represented through 8 tracking signals of 4 facial landmarks: the two eyes, the nose and the mouth. To improve the robustness of the tracking and reduce the intra-video variation, all frames are pre-processed using a histogram equalisation, colour component by colour component. During the head tracking step, the algorithm generates a starting template of 19 pixel rows and 25 pixel columns for each landmark, and uses no update strategy ($\alpha = 0$); then, the similarity scores of each colour component are based on the city-block distance measure. After that, the feature extraction consists of centring the tracking signals, by applying a zero mean transformation, and using the normalised head positions as features for recognition; in the end, the dimensionality of the feature space is still 8 (not reduced with PCA). Then, the client models are approximated using GMMs with 4 Gaussian components, and their parameters are estimated through the EM algorithm, which is initialised with: cluster means (computed using K-means), uniform weights and covariances. Finally, the impostor models for verification are approximated by taking the average of the best 2 background (or cohort) models.

We note that, in order to simplify the understanding and comparison between different graphs, in the following experiments we always express the results relative to the default configuration with a blue colour line.

IV.C.2. Precision of the head tracking

In this section we analyse the precision of the tracking signals, automatically extracted by our landmark tracker. We provide quantitative results for only one typical sequence, because we had to manually click on all facial landmarks in each frame of the video to generate the *ground truth* (1320 precise clicks), and it was unfeasible to process the whole database (274560 clicks!); though, the same effects have been observed on a bigger part of the data set by visual inspection.

The errors between the signals produced by the head tracker, $\mathbf{S} = [\mathbf{s}_1, \ldots \mathbf{s}_{2F}]^T \in N^{2F \times T}$, and the ground truth, $\mathbf{G} = [\mathbf{g}_1, \ldots \mathbf{g}_{2F}]^T \in N^{2F \times T}$, are calculated using the following measures:

- *Average absolute error*: $\varepsilon = \frac{1}{2FT} d^{(L_1)}(\mathbf{s}, \mathbf{g}) = \frac{1}{2FT} \sum_{f=1}^{2F} \sum_{t=1}^{T} \left| s_{f,t} - g_{f,t} \right|$.
- *Cumulative frame abs. error*: $e_t = \sum_{i=1}^{t} d^{(L_1)}(\mathbf{s}_i, \mathbf{g}_i) = \sum_{i=1}^{t} \sum_{f=1}^{2F} \left| s_{f,i} - g_{f,i} \right|$.

We first analyse the performance of the different pre-processing filters. We observe that histogram equalisation provides the best results, with an average absolute error of 0.7591 (pixels per point); on the other hand, contrast stretching is even worse than no filtering at all, with an average absolute error of 1.2451 and 0.8720 (pixels per point) respectively.

The evolution of the frame absolute error over time is shown in Figure 23. We notice that the cumulative error of the default configuration (with histogram equalisation) has a constant slope, which means that the error is uniformly distributed all over the sequence. On the contrary, without pre-processing or with contrast stretching, the error increases in the second half of the video when the intra-video appearance variation becomes significant and the initial templates differ more than the actual matches.

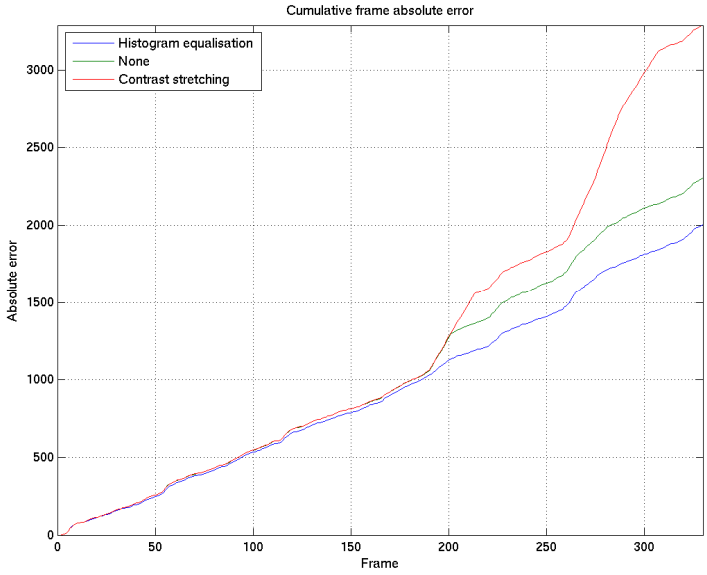

Figure 23: cumulative frame absolute error for various video pre-processing filters.

Then, we examine the importance of the choice of the distance measure, for the computation of the similarity scores; the city-block and Euclidean metrics perform better than the cosine distance, presenting an average absolute error of: 0.7591, 0.8693 and 1.0470 (pixels per point) respectively.

Figure 24 plots the cumulative frame absolute error for the three distance measures; the city-block distance appears to be more robust and tolerant to intra-video appearance changes than the two alternative metrics. We explain this behaviour by considering that the Euclidean and cosine distances possess a higher sensitivity towards appearance variations; in fact, when the initial templates start to differ from the actual matches due to local misalignments caused by pose changes or facial deformations, then: the number of outlier pixels increases, and their importance in the calculation of the matching error is more amplified by the 2-norm (used in Euclidean and cosine) rather than the 1-norm (used in the city-block). As a result, there are more occasional bad matches during the tracking process, the cumulative frame absolute errors show multiple leaps, and the resulting signals loose accuracy and get noisy.

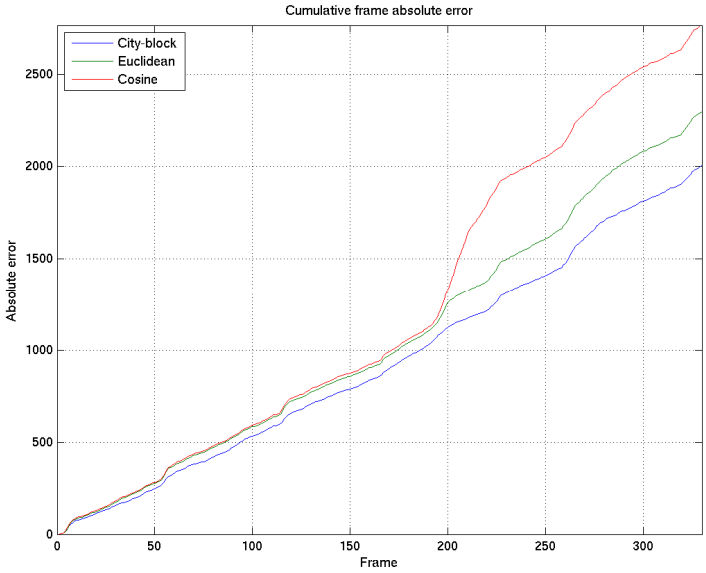

Figure 24: cumulative frame absolute error for various distance measures.

Afterwards, we study the effect of our template update strategy on the precision of the tracking signals; considering that the average absolute errors for partial and full updates are 1.8292 and 4.0598 (pixels per point) respectively, we can conclude that our template update strategy has a catastrophic impact on the accuracy of the tracking.

To have a better insight on this phenomenon, we visually inspected the modified templates over time, and we discovered that the more the templates were updated, the more they drifted away from the selected landmarks. In fact, although the template centres were initially aligned with the facial landmarks of interest, they constantly oscillated around their exact locations, because of the joint effect of small inaccuracies occurred during matching and repeated imprecise updates, and eventually glided away from the landmarks. We can also observe this effect by looking at the evolution of the frame absolute error in Figure 25, in which the progressive misalignment between each template and its corresponding landmark causes a constant increase on the slope of the error plot.

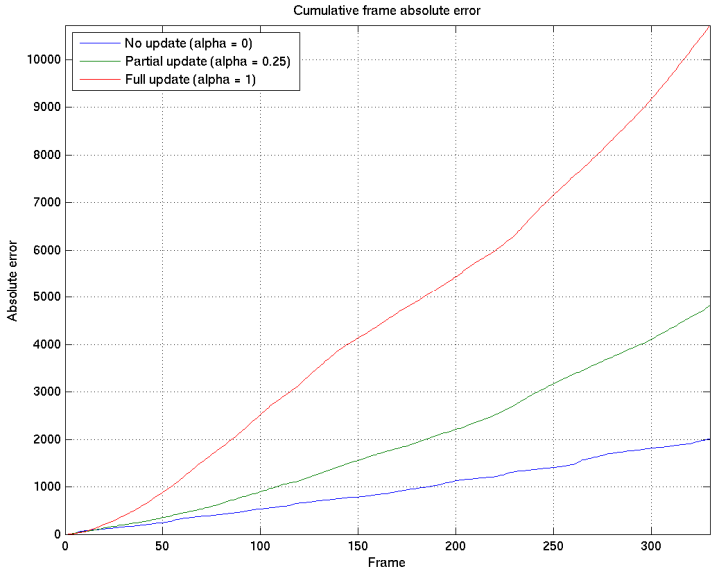

Figure 25: cumulative frame absolute error for various template updates.

Finally, we do not report qualitative recognition results as a function of the tracking parameters (pre-processing, distance measures, etc.), but we have empirically observed that the discriminatory power of the tracking signals is directly influenced by their precision; in addition, we are going to focus on the importance of the tracking accuracy in Section IV.C.5, where we propose an experimental evaluation of the negative effect of noisy signals on the final recognition scores.

IV.C.3. Recognition results in diverse experimental conditions

In this section, we assess the performance of our person recognition system by testing it in diverse experimental conditions, with various landmarks, signals, features and GMM estimations. We present and comment a selection of experiments, preferring those that better illustrate the properties of our approach, and help understanding the choices towards the best configuration.

Concerning the measures of performance, we express the identification results by reporting the *correct identification rates (CIRs)*, and by plotting the *cumulative (correct) match scores (CMSs)* as a function of the M best matches retained (Section II.E.2). For the verification scenario, we report the *equal error rates (EERs)* and we show the *receiver operating characteristic (ROC)* curves, which offer a global description of the system from low to high security applications (Section II.E.1).

We firstly analyse the impact of the number of facial landmarks and tracking signals on the discriminatory power of the feature space. The best results are obtained by the default configuration, which selects 4 landmarks (the eyes, the nose and the mouth) and represents the head motion using 8 signals: its CIR is the highest (90.4%) and its EER is the lowest (3.0%). When using only 2 or 3 facial landmarks (4 and 6 signals respectively), the results are sensibly worse, with CIRs of 72.1% and 76.9% and EERs of 10.5% and 6.5% respectively.

Figure 26 provides a complete overview of the recognition results: there is no doubt that the more tracking points and signals are used, the finer is the head motion representation and the more discriminating are the features extracted from them. We are also aware that the 3D movement of a rigid object can be represented by 3 points (6 signals) in the 3D space; though, we empirically observe that we need more than 3 facial landmarks for a proper estimation of the head motion in the 2D image space. If we consider that the human head is a semi-rigid object, with local deformations in the lower part of the face, then the resulting motion is more complex than that of a rigid object, and its modelling may require more than 6 parameters. A second element to take into account is the projection of the original 3D moving head into the 2D image plane; as a result, even if in our database the camera is fixed and there are minor depth changes, the motion estimation in the 2D image plane may require more tracking points than in the 3D space. Finally, as we have seen in Section IV.C.2, the automatic tracking algorithm does not provide exact signals, so using more facial landmarks can compensate for tracking errors.

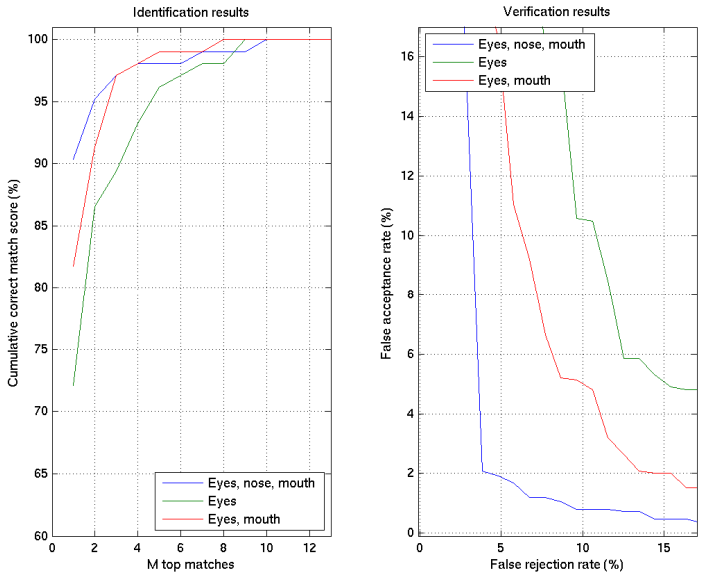

Figure 26: recognition results with a different number offacial landmarks and tracking signals.

Then, we study the effect of geometrical normalisations of the tracking signals on identification and verification results. The best discriminative features are extracted after centring the signals by using a zero mean transformation, with or without the scaling based on head size; when no scaling is used (default configuration), the CIR is 90.4% and the EER is 3.0%, otherwise the CIR is still 90.4% but the EER is 3.7%.

The plots for the CMSs and the DETs are shown in Figure 27 and confirm the numerical results above: it is not really advantageous to apply a scale normalisation based on head size, since a simple centring is sufficient. We explain this effect by recalling that in our database there are no zooms, and the head size is pretty similar in every video, so such a normalisation is most likely unnecessary; on the other hand, the calculation of the average eye distance is affected by tracking errors, and this inaccuracy probably causes the small degradation in performance of the red curves. Without any doubt, the worst results are those for the zero mean and unit variance normalisation: the CIR is 68.3% and the EER is 10.8%. These poor scores are caused by the imposition of a uniform variance, which is an excessive constraint in our experimental conditions and clearly alters the discriminatory power of the tracking signals.

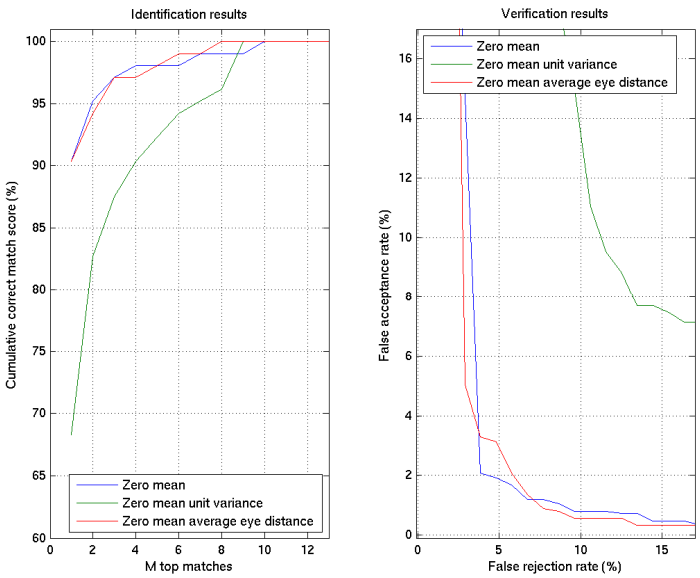

Figure 27: recognition results with different geometrical normalisations.

Afterwards, we consider some GMM approximations of different complexity, by increasing the number of Gaussian components from one to five, and we evaluate the effectiveness of their client modelling. At the beginning, the recognition rates improve with the complexity of the approximating distribution: the CIR increases from 76.9% of the single Gaussian to 90.4% when using three or four components (as in the default configuration), and the EER decreases from 7.6% to 3.8% and 3.0% for three and four components respectively. But once the optimal number of Gaussians is attained, a further increase of complexity slowly makes the scores worse: in fact with five components the CIR is 85.6% and the EER is 4.5%.

The evolution of the classification results due to the increase in complexity of the GMM modelling is illustrated in Figure 28. It is clear that selecting only one or two Gaussians does not provide very good outcomes, because the distribution of features is undoubtedly multimodal and one or two components cannot approximate their PDF precisely enough. On the other hand, when using too many components (five or more) the classifier gets over fitted to the enrolment subset, which reduces its ability to generalise to unknown data; furthermore, a complex modelling increases the occurrence of singular estimations in the covariance matrices, especially when the training data is small as in our case. In conclusion, choosing three or four Gaussian components for our system appears to be the best compromise between complexity, accuracy and generalisation power.

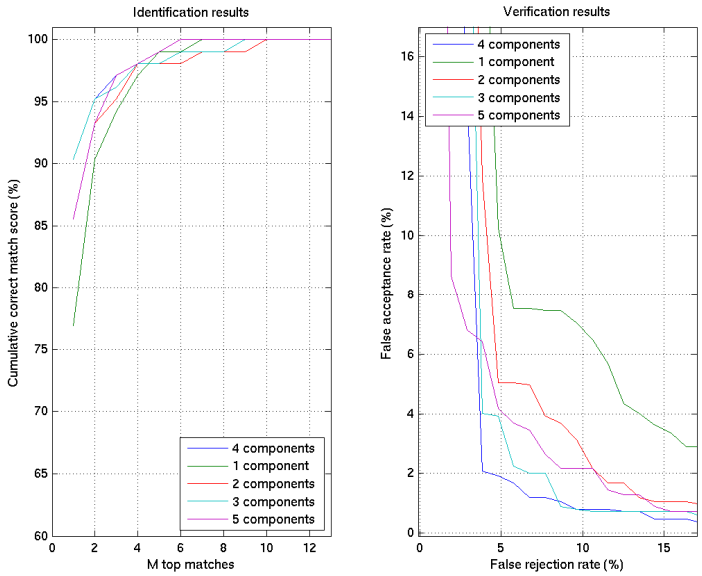

Figure 28: recognition results with a different number ofGaussian components.

Finally, we examine the consequences of applying diverse GMM estimation techniques on classification performances. The default configuration employs an EM algorithm initialised with: cluster means (obtained through K-means) and uniform weights and covariances. We also implement two other EM variants that apply a K-means or a fuzzy K-means clustering algorithm, and have the following starting conditions and final results:

1. *EM, K-means (2)*: initialised with cluster means, weights and covariances; its CIR is 85.6% and its EER is 4.6%.
2. *EM, fuzzy K-means*: initialised with cluster means, then uniform weights and covariances; its CIR is 89.4% and its EER is 4.1%.

In addition, we also estimate the GMM parameters by using the Figueiredo-Jain algorithm, which achieves a CIR of 88.5% and an EER of 5.6%.

The identification and verification results, illustrated in Figure 29, show that the EM algorithms initialised with uniform weights and covariances (EM, K-means (1) or EM, fuzzy K-means) perform equally well; what is more, their result are better then when using cluster means, weight and covariances (EM, K-means (2)), or more elaborate estimation techniques like the Figueiredo-Jain (and Greedy EM, which is not reported here). Looking at the closeness of the best curves (sometimes even coincident), we deduce that the clustering algorithm is not a determinant factor for the GMM modelling and final classification results; however, we observe that the simpler K-means is slightly more performing than its fuzzy version, but at the expense of less stability in its clustering outcomes over multiple estimations. Moreover, the comparison between the EM and FJ algorithms confirms the experimental results of Paalanen et al. [66], which are the following: "the standard EM algorithm outperforms FJ and GEM if a good prior knowledge exists about the number of components." In fact, with the previous experiments on the complexity of the GMM modelling, we have determined the optimal number of components, so once the EM is taking advantage of this information the automatic estimation of Figueiredo-Jain and Greedy EM does not obtain better results.

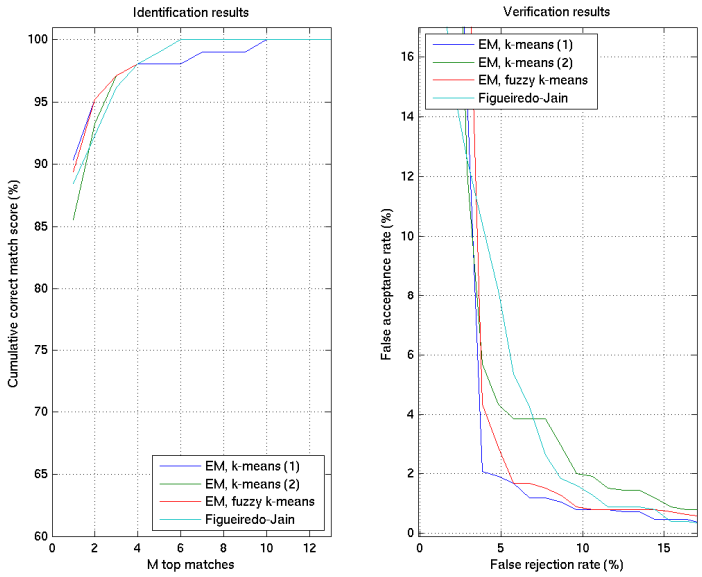

Figure 29: recognition results with different estimation techniques ofGMM.

IV.C.4. Comparison with the eigenface technique

As we have seen in Chapter III, to the best of our knowledge there are no other approaches that make use of natural head motion for person recognition; for this reason, we cannot compare our experimental results with concurrent techniques that exploit the same biometric identifier. Nevertheless, to give an idea of the discriminatory power of our person recognition strategy using head motion, we relate it with a state of the art recognition technique based on facial appearance: *eigenfaces* [87]. We remind that the standard eigenface approach linearly projects the facial images to a feature subspace computed by applying the *principal component analysis (PCA)*, and that the classification in this face space is obtained through a nearest neighbour classifier using distances as similarity measures. However, we invite the reader to refer to [87] and Section III.B.1 for a detailed description on the eigenface method.

In our implementation of the eigenface approach, we firstly pre-process all images with a *histogram equalisation*, colour component by colour component, to reduce the mismatches due to illumination variations. Next, we represent the data set by using the *NTSC colour space* (which consists of: luminance, hue and saturation) [28], because it empirically provides more discriminative signals than the RGB does. Once the colour components are rearranged into large vectors, we apply the PCA to the enrolment subset to compute a reduced face space of dimension 243, and we calculate the feature vectors by *whitening* the projection coefficients in the eigenspace. Then, the client models are registered into the system using their centroid vectors, which are calculated by taking the average of the feature vectors in the enrolment subset; in the end, recognition is achieved using a nearest neighbour classifier with *cosine distances* in (the whitened) face space. A summary of the parameters for our eigenface implementation is proposed in Table 2.

		NORMALISED DATABASE	RAW DATABASE
DATABASE	Database name	Image datab. of Italian TV speakers	Image datab. of Italian TV speakers
	Normalisation	Accurate: in-plane rotated and aligned	None
PRE-PROCESSING	Image size	32 pixel rows (or height)	48 pixel rows (or height)
		32 pixel columns (or weigth)	61 pixel columns (or weigth)
	Resizing interpolation method	Nearest neighbour	Nearest neighbour
	Image pre-processing	Histogram equalisation	Histogram equalisation
	Colour space	NTSC (luminance, hue & saturation)	NTSC (luminance, hue & saturation)
	Vertical mirroring	No	No
FEATURE EXTR.	Image space reduction method	Centered PCA	Centered PCA
	Subspace dimension	243	243
	Whitening of feature vectors	Yes	Yes
MODEL ESTIM.	Client model generation method	Centroid vector (average of features)	Centroid vector (average of features)
CLASSIFICATION	Similarity measure	Based on cosine distance	Based on cosine distance

Table 2: summary of the parameters for our eigenface implementation.

The recognition results for our system are calculated using the default configuration and the video database of Italian TV speakers (Section VIII.A): as we have already seen, the CIR is 90.4% and the EER is 3.0%. For evaluating the performance of the eigenface approach, we prefer to employ an appropriate version of the video data set, called the image database of Italian TV speaker, which is derived from the video one by sub sampling and normalising frames, and it is detailed in Section VIII.B. With this database the eigenface approach is tested in its optimal condition, due to the manual accurate normalisation of video frames, and the results are excellent: 100.0% of CIR and 0.0% of EER (perfect recognition). Nevertheless, the image database of Italian TV speakers creates a too favourable and unrealistic situation, so we also evaluated the eigenface technique on a raw version of the dataset without normalisation; in this *somewhat unfavourable condition*[1], the CIR decreases to 69.2% and the EER increases to 10.8%. However, it is worth noting that our recognition system is not in its optimal working condition either, because the tracking signals are corrupted by the noise of the automatic tracking process, which significantly reduces their discriminatory power.

[1] We consider that the eigenface technique applied to our not normalised database is working in a *somewhat unfavourable condition*, because we suppose that in real applications the eigenface recognition system incorporates an automatic face detection step, which should provide a better face normalisation (face warping, alignment, etc.) than the one in our raw data set.

We have a complete overview of the recognition results by looking at Figure 30, whose graphs clearly confirm that the performance of our recognition system is in between those for the favourable and unfavourable eigenface approaches. This is an interesting outcome, because it demonstrates that natural head motion possesses enough discriminatory power to be used as a possible biometric in recognition applications, and it corroborates the assertion that the face can be considered as a hybrid biometric identifier. On the other hand, we are aware that eigenfaces are far from being the most performing technique working on facial appearance [69], and that our system is not as accurate as the best methods presented in this research domain. For this reason, we currently do not regard head motion as a practical alternative to facial appearance; nevertheless, we are convinced that this information can be successfully integrated in multimodal recognition systems, and that it has the potential to become a worthy biometric in video surveillance applications.

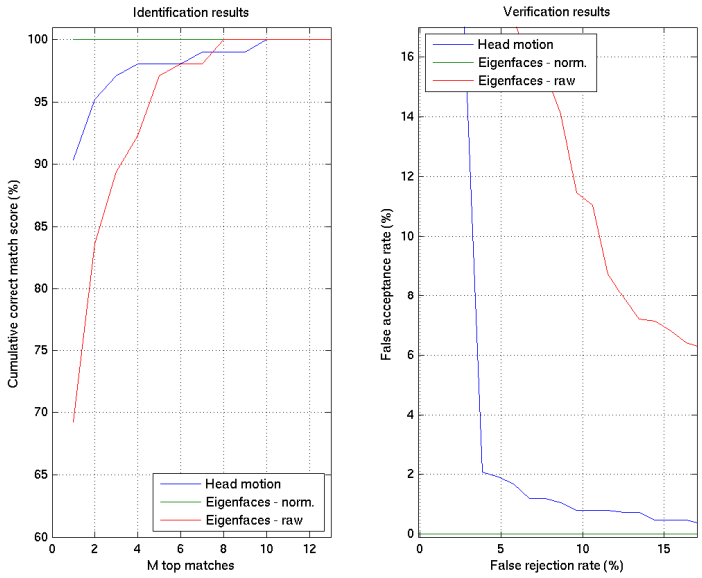

Figure 30: comparison ofperson recognition results between: the proposed method and eigenfaces.

IV.C.5. Recognition results with artificial noisy tracking signals

Due to the absence of ground truth tracking points for the whole database, we cannot assess the discriminatory power of the head motion information in that ideal scenario, or fully evaluate the robustness of our person recognition system to noisy tracking signals. However, we have already analysed the precision of our automatic landmark tracker in Section IV.C.2, and we have found that the extracted points present a significant error compared to the ground truth ones: their average absolute error in the default configuration is 0.7591 pixels per point. Hence, here we want to investigate and quantitatively evaluate the influence of the tracking accuracy on recognition results, by artificially adding a Gaussian noise with zero mean and variable standard deviation to all tracking signals (in both the enrolment and recognition subsets). For this purpose, we define the *noise strength*, ξ, as the ratio between the standard deviation of the Gaussian noise, $\sigma^{(n)}$, and the average standard deviation of the tracking signals, $\overline{\sigma}^{(s)}$, expressed in percentages:

$$\xi = \frac{\sigma^{(n)}}{\overline{\sigma}^{(s)}} * 100$$

where the average standard deviation of the tracking signals, $\overline{\sigma}^{(s)}$, is constant and it is computed as:

$$\overline{\sigma}^{(s)} \equiv \frac{1}{2F}\sum_{f=1}^{2F}\sigma_f^{(s)} \equiv \frac{1}{2F}\sum_{f=1}^{2F}\sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left(s_{f,t}-\mu_f\right)^2}$$

The experiments run using the default configuration show that the recognition rates are clearly affected by tracking noise. In fact, the CIR suddenly decreases from 90.4% to 85.6% (for $\xi = 5\%$), then 81.7% ($\xi = 10\%$) and 77.9% ($\xi = 15\%$); in parallel, the EER rapidly increases from 3.0% to 3.9% (for $\xi = 5\%$), then 6.0% ($\xi = 10\%$) and 8.7% ($\xi = 15\%$). After that, a further rise in the noise strength does not translate into equivalent performance degradation: for example, when the noise strength is 50%, the CIR is still 60.6% and the EER is 13.9%.

By looking at Figure 31, we can visually evaluate the incidence of tracking noise on identification and verification scores; in particular, we notice that the highest degradation in performance occurs for low values of the noise strength (0% - 15%). One possible explanation for this effect is the following: it is highly possible that the finer head motion information is the most characteristic one, so a small amount of noise has a big incidence on recognition results, because it quickly corrupts the most discriminative part of the tracking signals. On the other hand, the overall movement of the head is less distinctive than its finer motion, but definitively more robust to noise; for this reason a subsequent increase in the noise strength does not affect the performance of the system as much as before.

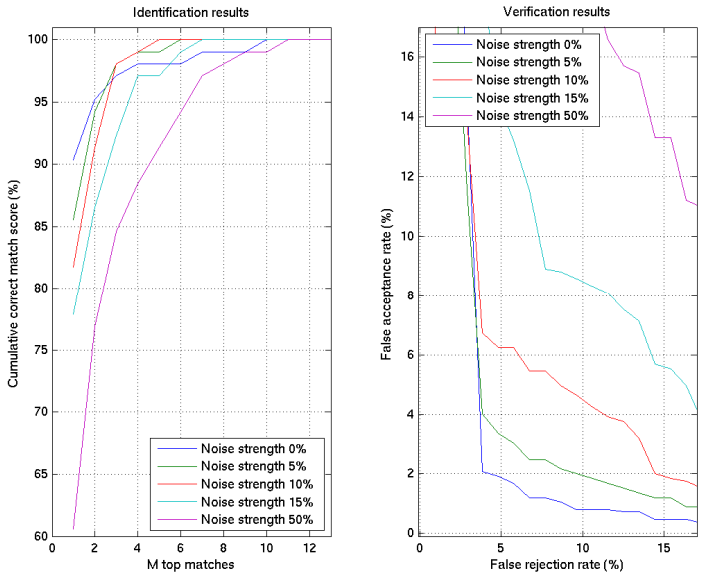

Figure 31: recognition results with different strength values for the noise added to the tracking signals.

Finally, to give an idea of the unfavourable working condition of our system, we estimate the noise power that is present in the signals automatically extracted by the landmark tracker. For this experiment, we consider the default configuration and the sequence for which we have the ground truth information; then, we calculate its noise standard deviation ($\sigma^{(n)} = 0.9548$) and we finally end up with a noise strength of: 15.5%! Therefore, we presume that the potential discriminatory power of the head motion information is superior of what appears in our experimental results.

IV.C.6. Gender recognition results

To conclude these experimental sections, we would like to evaluate the performance of our approach when applied to a different scenario: a gender recognition application; in particular, we want to compare the gender discriminatory power of natural head motion with that of facial appearance, similarly to what we did in Section IV.C.4 for identity. Therefore, we consider an equivalent experimental set-up that is adapted to the gender recognition task, with the same eigenface implementation and databases of Italian TV speakers. It is worth noting that due to the particular nature of this recognition problem, which consists of only two classes, the CIRs and EERs are directly related: $\eta^{(CIR)} = 100\% - \xi^{(EER)}$.

In this scenario, our system does not obtain good gender recognition scores: its CIR is 84.6% so its EER is 15.4%. In addition, natural head motion appears less discriminating than facial appearance, because both the favourable and unfavourable conditions of the eigenface approach outperform our system. As we expected, when using the normalised image database, eigenfaces obtains excellent recognition results: it achieves perfect recognition, with a CIR of 100.0% and an EER of 0.0%. Though, facial appearance performs better even in its unfavourable condition, when tested with the database of raw images, showing a CIR of 89.4% and an EER of 10.6%.

We have a complete overview of the gender recognition results by looking at Figure 32, whose graphs clearly illustrate how the eigenface approach outperforms our system. Moreover, if we compare the person verification results of Figure 30 with the gender ones in Figure 32, we notice that there is evident performance degradation for our system, while the eigenface approach has similar rates. All these elements motivate us to conclude that, apart from tracking noise, the natural head motion information is not such a discriminative identifier for gender recognition; moreover, keeping in mind the good results obtained for person recognition, we also deduce that natural head motion is more an individual rather than a sexual characteristic.

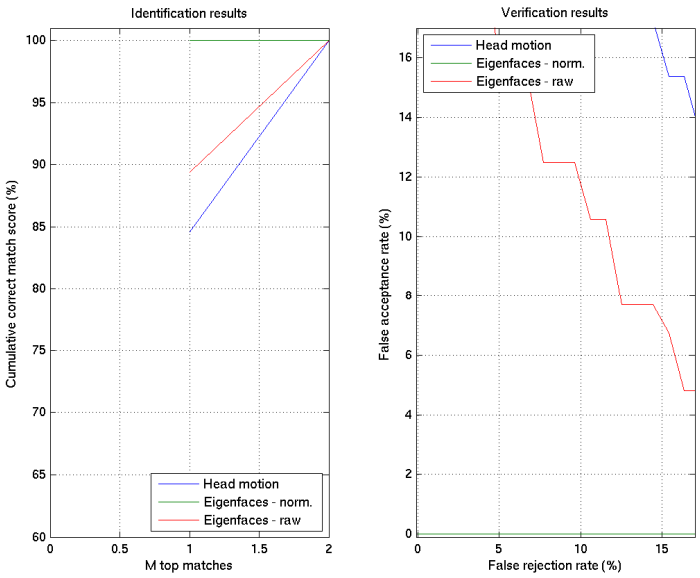

Figure 32: comparison of gender recognition results between: the proposed method and eigenfaces.

IV.D.Concluding summary

In this chapter we presented a novel person recognition system that exploited the unconstrained head motion information, extracted by tracking a few facial landmarks in the image plane. In particular, we detailed how each video sequence was firstly pre-processed by semi-automatically detecting the face, and then automatically tracking the facial landmarks over time using a *template matching* strategy. Then, we described the geometrical normalisations of the extracted signals, the calculation of the feature vectors, and how these were successively used to estimate the client models through a *Gaussian mixture model (GMM)* approximation. In the end, we achieved person identification and verification by applying the probability theory and the *Bayesian decision rule* (also called *Bayesian inference*).

We assessed the performance of our system through multiple experiments, whose results corroborate the following conclusions. First of all, the tracking signals that are automatically extracted by our system are not very accurate; for this reason, their potential discriminatory power and the performance of our recognition approach are significantly reduced. Nevertheless, our biometric system achieves good person recognition results, which are in between those obtained with a favourable and an unfavourable eigenface implementation. Hence, we deduce that natural head motion possesses enough discriminatory power to be used as a possible biometric in recognition applications, but it is not yet a practical alternative to facial appearance. Finally, considering the poor scores in a gender recognition scenario, we conclude observing that natural head motion seems a more individual rather than a sexual characteristic.

Chapter V. Multimodal integration of head motion with mouth motion and facial appearance

V.A. Introduction

Our research on the use of natural head motion for person recognition (Chapter IV) attested that video data contains more valuable biometric information than just the well known facial appearance, and corroborated the assertion that human face can be considered as a hybrid identifier (Section II.B.2). Nevertheless, it is a common trend in literature (Chapter III) to exploit only a part of the biometric information embedded in video sequences, mainly the physiological one related to facial appearance, and as far as we know it has never been proposed a hybrid person recognition system, which makes use of the physiological and behavioural aspects of the face at the same time. These facts encouraged us to develop a novel person recognition approach using as much biometric video information as possible; in particular, we successfully integrated head motion with mouth motion and facial appearance, through a unified probabilistic framework.

We firstly decided to study the discriminative properties of unconstrained facial motion, due to its close relationship with the head one; in fact, their analogous dynamic nature facilitated the integration of a few mouth parameters in our previous recognition system (detailed in Chapter IV), by enriching its feature space with this valuable temporal information. Then we considered taking advantage of the facial appearance information also present in video sequences, which is one of the traditional biometric identifiers for person recognition and it has been largely studied during the last decades [12][67][96]. Unfortunately, the different nature of facial appearance and head and mouth motion prevented us from directly integrating this spatial information in our temporal system; therefore, we were obliged to develop two parallel recognition subsystems, with independent feature spaces, user models and classifiers. As a result, we had to constrain both the spatial and temporal subsystems to adopt the same probabilistic classification framework, in order to facilitate the multimodal integration of motion and appearance in the fusion module.

The remainder of this chapter is organised in two main sections: one theoretical part that details the structure of our multimodal recognition system, and one experimental part that thorough fully evaluates the performances of our approach in various conditions.

V.B. Proposed method

The architecture of the multimodal extension of our person recognition system is illustrated in Figure 33.

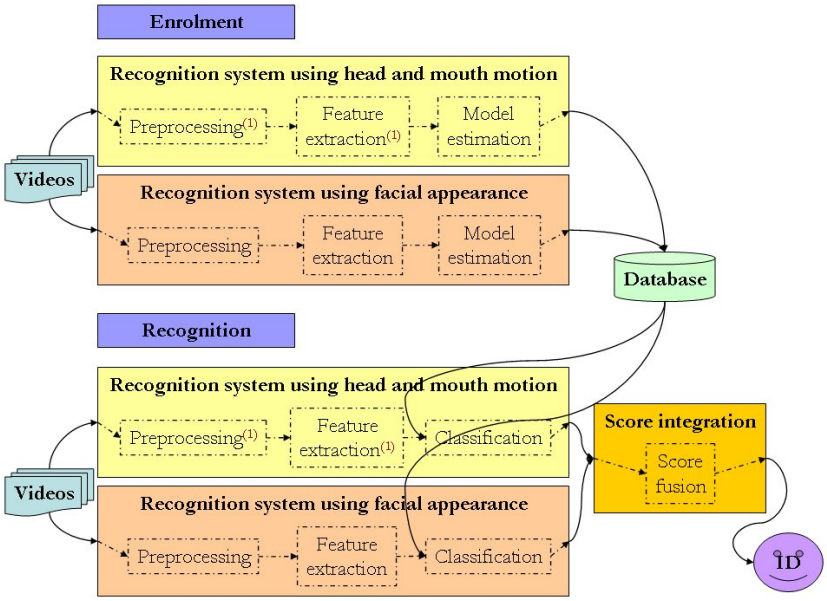

Figure 33: architecture ofthe multimodal extension ofour person recognition system.

The multimodal recognition system is composed by two parallel complementary subsystems and a score integration step. The first recognition subsystem (identified by light yellow boxes in Figure 33) is exploiting the temporal video information and is based on unconstrained head and mouth motion; in particular, it closely resembles the approach presented in Chapter IV, with the addition of some mouth parameters to enrich the discriminatory power of the extracted features. The second recognition subsystem (identified by tan boxes in Figure 33) works with the spatial information and exploits facial appearance; more precisely, it is a probabilistic extension of the original eigenface technique presented in [87] by Turk and Pentland. For a consistent integration of this heterogeneous biometric information (motion and appearance) into a unified recognition approach, both subsystems share the same probabilistic framework: a *Gaussian mixture model (GMMs)* approximation to represent the biometric features of each client, and *Bayesian inference* to calculate the similarity between tests and models. In the end, the similarity scores of the two parallel subsystems are combined in the last step (identified by a gold box in Figure 33), which operates the final identification and verification decisions after a suitable *opinion fusion* (or *score fusion*). The two subsystems and the integration step are detailed in the following three subsections.

V.B.1. Integration with mouth motion[2]

To integrate some additional parameters related to local mouth dynamics, the recognition system based on 2D unconstrained head motion has to be modified in its pre-processing and feature extraction steps, which are described in Section IV.B and are identified by a red "(1)" in Figure 33. In fact, besides the head detection and tracking phases, the pre-processing should also include a lip segmentation part; next, not only the head but also the mouth parameters have to be calculated in the feature extraction step, and then combined into common feature vectors.

The *lip segmentation algorithm* applies a series of image processing techniques [28] to locate the outer lip contour in every video frame, Φ_t. First of all, it crops a sub image corresponding to the mouth region, Λ_t, from the original frame, by exploiting the position of the mouth given by the head tracker, $\mathbf{s}_t^{(mt)}$.

Then, it applies the colour conversion that has been proposed by Canzler and Dziurzyk [9] for lip enhancement: the purpose of this transformation is to reduce the contribution of the blue component, because it plays a reduced role in lip segmentation based on skin colour. Thus, the transformed lip region, Γ_t, is the following:

$$\Gamma_t = \frac{2\Lambda_{t,G} - \Lambda_{t,R} - 0.5\Lambda_{t,B}}{4}$$

where $\{\Lambda_{t,i} \mid i = R,G,B\}$ are the RGB colour components of the mouth region, Λ_t.

Afterwards, the lip segmentation algorithm detects the edges of the transformed lip region, Γ_t; it firstly computes the horizontal and vertical gradients with the Sobel approximation of the derivatives, and then it generates the *binary edge map* by applying Otsu's thresholding [65], which chooses the threshold value that minimises the inter-class variance of the black and white pixels in the binary image.

[2] The work on the extraction of mouth parameters has been done in collaboration with Usman Saeed.

The last part of the segmentation process consists of several additional steps to isolate and enhance the shape of the outer lip contour. In fact, the binary edge map can contain spurious contours due to the presence of the nose tip, tongue or teeth, so it may be convenient to apply some *morphological operators* to delete them as: dilate the image, fill the holes, and remove 8-connected components that are linked with the boundaries of the edge map. In addition, as discussed by Bourel et al. in [7], the resulting lip contour can be prone to the following two problems: it can be missed altogether or it can be extracted incompletely. In the former case, typically another facial landmark other than the lip is segmented, like the nose tip or the tongue; this problem can be easily detected and corrected, by checking some geometrical constraints such as: the lip cannot be linked to the boundary of the edge map, or the area inside the segmented lip contour should not be smaller than 1/3 of the average one in that sequence. However, when the lip is not segmented in its entirety, it usually means that the lower part is missing; this problem is more difficult to detect and can be recovered only partially by using a temporal smoothing filter.

Finally, the outer lip contour is regularised by computing the *convex hull* of the enhanced edge map with the quickhull algorithm [2]. In fact, the convex hull, which is the minimal convex subset of points that contains the whole set, provides a more efficient representation of the contour, by smoothing it and filling its holes with a homogeneous sampling of the curve. The final result of the lip segmentation algorithm is the *regularised binary edge map*, Ψ_t, and a few examples of these extracted contours are shown in Figure 34.

Figure 34: illustration of the segmented outer lip contours.

The feature extraction step is also modified for being able to calculate both head and mouth parameters, and integrate them into combined feature vectors; the generation of features related to the head motion information is already detailed in Section IV.B.3, so here we focus on the computation of the additional parameters related to mouth dynamics.

The *mouth feature matrix*, $\mathbf{X}^{(mt)} \in \Re^{D^{(mt)} \times N}$, is generated by concatenating one or more of the following features:

- *Centred major axis of the outer lip contour* : the length of the major axis of the lip contour, $u_{1,t} \in \Re^T$, partially characterises the mouth motion over time: $x_{d,n} = u_{1,t} - \mu_1$ for $t = 1,\ldots,T$, where μ_1 is the mean value, $\mu_1 = \frac{1}{T}\sum_{t=1}^{T} u_{1,t}$.

- *Centred minor axis of the outer lip contour* : the length of the minor axis of the lip contour, $u_{2,t} \in \Re^T$, also characterises the mouth motion over time: $x_{d,n} = u_{2,t} - \mu_2$ for $t = 1,\ldots,T$, where μ_2 is the mean value, $\mu_2 = \frac{1}{T}\sum_{t=1}^{T} u_{2,t}$.

In our study we tested other features based on the eccentricity value or the perimeter length of the lip contour, but they empirically showed less discriminatory power and were abandoned.

Finally, the *head feature matrix*, $\mathbf{X}^{(hd)} \in \Re^{D^{(hd)} \times N}$, and the mouth feature matrix, $\mathbf{X}^{(mt)} \in \Re^{D^{(mt)} \times N}$, are integrated into an *extended feature matrix*, $\mathbf{X}$, by applying the *feature concatenation* fusion strategy (Section II.H.1):

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(hd)} \\ \mathbf{X}^{(mt)} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^{(hd)} & \cdots & \mathbf{x}_N^{(hd)} \\ \mathbf{x}_1^{(mt)} & \cdots & \mathbf{x}_N^{(mt)} \end{bmatrix} \in \Re^{D \times N}$$

where $D = D^{(hd)} + D^{(mt)}$. We notice that the number of feature vectors is equal to the video length, $N = T$, while the dimension of the extended feature space D depends on the features selected; for example, when using (normalised) head positions and (centred) major and minor axes: $D = 2F + 2$, where F is the number of facial landmarks. In the end, each extended feature matrix, $\mathbf{X} \in \Re^{D \times N}$, retains the whole head and mouth discriminative information extracted from the corresponding video sequence; then, these extended features are successively used for model estimation and classification, as detailed in Sections IV.B.4 and IV.B.5.

V.B.2. Probabilistic extension of the eigenface method

For a consistent integration of the facial appearance information in our multimodal person recognition system, we developed a probabilistic extension of the original eigenface technique [87], which is depicted in Figure 33 with tan blocks. In particular, the pre-processing and feature extraction steps are kept pretty close to the standard eigenface approach, while the model estimation and classification steps are adapted to share the same probabilistic framework of the other recognition subsystem that exploits head and mouth motion. In the following we summarise the steps of our subsystem, and we invite the reader to refer to: [87] and Section III.B.1 for a better insight on eigenfaces, and Sections IV.B.4 and IV.B.5 for an extensive description of the probabilistic framework.

The pre-processing step applies some image processing techniques [28] to a set of N colour pictures of size $R \times C$ belonging to a video sequence, $\left\{\Phi_n \in \mathrm{N}^{R \times C \times 3} \mid n = 1, \ldots, N\right\}$, and generates a set of transformed vectors, $\left\{\mathbf{s}_n \in \mathrm{N}^{3RC} \mid n = 1, \ldots, N\right\}$, in which the image pixels are arranged in long vectors (a process called *image vectorisation*). The first transformation is a contrast enhancement like a *histogram equalisation* or a *contrast stretching* (colour component by colour component), that is useful to reduce the impact of inter-image illumination and colour variations. Then, this step converts the image signal into the most discriminative representation, chosen among these ones: RGB (red, green and blue), HSV (hue, saturation and value), NTSC (luminance, hue and saturation), YCbCr (luminance and chrominance components) or greyscale values. In addition, the amount of data can be optionally incremented by mirroring each image along its vertical axis (a process called *vertical mirroring*), before that the image vectorisation takes place.

Afterwards, the feature extraction step isolates the discriminative information that characterises the individual and discards the irrelevant one, by transforming the vectorised data set, $\left\{\mathbf{s}_n \in \mathrm{N}^{3RC} \mid n = 1, \ldots, N\right\}$, into the corresponding *feature matrix*: $\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_N\right] \in \Re^{D \times N}$. First of all, it applies a linear transformation from the high dimensional *image space*, N^{3RC}, to a lower dimensional space (called the *face space*), $\Re^D$, which is much smaller: $D << 3RC$. More precisely, each vectorised image $\mathbf{s}_n \in \mathrm{N}^{3RC}$ is approximated with its projection in the face space $\mathbf{v}_n \in \Re^D$ by the following *linear transformation*:

$$\mathbf{v}_n = \mathbf{W}^T \left(\mathbf{s}_n - \boldsymbol{\mu}\right)$$

where $\mathbf{W} \in \Re^{3RC \times D}$ is a projection matrix with orthonormal columns, and $\boldsymbol{\mu} \in \Re^D$ is the mean image vector of the whole training set:

$$\boldsymbol{\mu} = \frac{1}{JN} \sum_{j=1}^{J} \sum_{n=1}^{N} \mathbf{s}_{j,n}$$

in which J is the total number of sequences in the training set, and $\mathbf{s}_{j,n} \in \mathrm{N}^{3RC}$ is the n-th vectorised image belonging to video Φ_j.

The *optimal projection matrix* $\mathbf{W}$ is computed using the *principal component analysis (PCA)* (also called the *Karhunen-Loeve transform (KLT)*) [22], which has the property of optimally representing the distribution of data in the root mean squares sense; the details on the calculation of $\mathbf{W}$ can be found in [87] and in Section III.B.1.

Once the image data set is projected into the face space, the vectors in the feature matrix, $\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_N\right] \in \Re^{D \times N}$, are generated by choosing either:

1. The *projections in face space*: in this case, $\mathbf{x}_n = \mathbf{v}_n$ for $n = 1, \ldots, N$.

2. The *whitened projections in face space*: the *whitening* process (Section III.B.1) rescales the projection coefficients $\mathbf{v}_n$ to counterbalance the overweighting of the low frequencies as following: $x_{d,n} = \frac{v_{d,n}}{\sqrt{\lambda_d}}$ for $d = 1,\ldots,D$ and $n = 1,\ldots,N$, in which λ_d is the d-th largest eigenvalue.

The model estimation step adopts the same probabilistic approach of the parallel subsystem using head and mouth motion for recognition. In fact, the distribution of the feature vectors of each client is modelled with a GMM, which approximates the class *conditional probability density function* of each user, k, in feature space:

$$p(\mathbf{x}_n \mid k) \cong p(\mathbf{x}_n \mid \boldsymbol{\Theta}_k) \equiv \sum_{c=1}^{C_k} \alpha_{k,c} \aleph(\mathbf{x}_n \mid \boldsymbol{\mu}_{k,c}, \boldsymbol{\Sigma}_{k,c})$$

in which $\aleph(\ )$ is a non-singular multivariate normal distribution, $\boldsymbol{\Theta}_k = \{\alpha_{k,c}, \boldsymbol{\mu}_{k,c}, \boldsymbol{\Sigma}_{k,c} \mid c = 1,\ldots,C\}$ is the parameter list for the k-th model, and $\alpha_{k,c} \in [0,1]$ is the weight of the c-th Gaussian component. We do not provide more details here, so the reader is invited to refer to Section IV.B.4 for a discussion on GMM modelling and an extensive description of its parameter estimation. Though, it is important to notice that the facial appearance manifold needs a high dimensional feature space to be exhaustively represented, and that high dimensional distributions are difficult to approximate due to limited amount of training data; this is a big issue indeed because we remind that, for the estimation of GMM parameters, the minimum number of recommended feature vectors increases quadratically with the dimensionality of the feature space, D.

Finally, the classification step closely resembles to the one in the temporal recognition system (Section IV.B.5); in fact, it also computes the similarity scores by applying the probability theory and the *Bayesian decision rule* (also called *Bayesian inference*). In our implementation, we select only one key frame to test a given video sequence; hence, the related feature matrix contains only one feature vector, $\mathbf{X} = \mathbf{x} \in \Re^D$, and the *video posterior probability* is equal to the *frame posterior probability*, which is calculated using the *Bayes' rule*:

$$p(k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid k) p(k)}{p(\mathbf{x})}$$

First of all, the divisor:

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x} \mid k) p(k) = M_{\mathbf{x}}$$

is merely a scaling factor $M_{\mathbf{x}}$, to assure that the posterior probabilities $p(k \mid \mathbf{x})$ are really probabilities (their sum is one). Then, the *a priori probability* $p(k)$ represents the probability of occurrence of each class k, and it is usually estimated from the training database.

Afterwards, the *similarity score for the identification task* $S^{(ID)}(\mathbf{x}, \mathbf{\Theta}_k)$ (Section II.C.2), is derived from the image/video posterior probability $p(k \mid \mathbf{x})$ by computing the *log-posterior probability*, because it is analytically and numerically more practical, and the properties of the similarity function do not change thanks to the monotonicity of the logarithm. Hence, $S^{(ID)}(\mathbf{x}, \mathbf{\Theta}_k)$ takes the form of:

$$S^{(ID)}(\mathbf{x}, \mathbf{\Theta}_k) = \ln p(k \mid \mathbf{x}) = \ln p(\mathbf{x} \mid k) + \ln p(k) - \ln M_{\mathbf{x}}$$

Finally, the *similarity score for the verification task* $S^{(VER)}(\mathbf{x}, \mathbf{\Theta}_k)$ (Section II.C.1), is the *log-posterior probability ratio*:

$$S^{(VER)}(\mathbf{x}, \mathbf{\Theta}_k) = \ln\left[\frac{p(k \mid \mathbf{x})}{p(\bar{k} \mid \mathbf{x})}\right] = \ln p(\mathbf{x} \mid k) - \ln p(\mathbf{x} \mid \bar{k}) + 2\ln p(k) - 1$$

where $p(\bar{k} \mid \mathbf{x})$ is the posterior probability of the *alternative hypothesis* $\bar{k}$, and $p(\mathbf{x} \mid \bar{k})$ is the *impostor model* (the class conditional PDF for $\bar{k}$). Following the same approach of Section IV.B.5, we approximate the impostor model by using the L best client models $p(\mathbf{x} \mid k^{(l)})$, which are called *background models* or *cohorts*:

$$p(\mathbf{x} \mid \bar{k}) \cong \frac{1}{L}\sum_{l=1}^{L} p(\mathbf{x} \mid k^{(l)})$$

where $k^{(l)}$ is the client model that produces the l-th highest posterior probability $p(k \mid \mathbf{x})$.

V.B.3. Integration with facial appearance

The score integration step, which is identified by a gold box in Figure 33, combines the similarity scores of the two parallel subsystems by applying a suitable *opinion fusion* (or *score fusion*) strategy (Section II.H.3); after that, it takes the final identification and verification decisions using this extended measure of similarity.

We remind that the identification and verification similarity measures of both recognition subsystems are: the *video log-posterior probability* and the *video log-posterior probability ratio*, respectively. More precisely, the similarities between the j-th test and the k-th client model are:

$$\eta_{j,k}^{(ID)} \equiv S^{(ID)}(\mathbf{X}_j^{(mtn)}, \mathbf{\Theta}_k) = \ln p(k \mid \mathbf{X}_j^{(mtn)})$$
$$\eta_{j,k}^{(VER)} \equiv S^{(VER)}(\mathbf{X}_j^{(mtn)}, \mathbf{\Theta}_k) = \ln\left[\frac{p(k \mid \mathbf{X}_j^{(mtn)})}{p(\bar{k} \mid \mathbf{X}_j^{(mtn)})}\right]$$

for the recognition subsystem using unconstrained head and mouth motion (Section IV.B.5), and:

$$\rho_{j,k}^{(ID)} \equiv S^{(ID)}\left(\mathbf{x}_j^{(app)}, \mathbf{\Theta}_k\right) = \ln p\left(k \mid \mathbf{x}_j^{(app)}\right)$$
$$\rho_{j,k}^{(VER)} \equiv S^{(VER)}\left(\mathbf{x}_j^{(app)}, \mathbf{\Theta}_k\right) = \ln\left[\frac{p\left(k \mid \mathbf{x}_j^{(app)}\right)}{p\left(\bar{k} \mid \mathbf{x}_j^{(app)}\right)}\right]$$

for the one working with facial appearance (Section V.B.2).

Accordingly, the score integration step calculates the multimodal similarity scores between the j-th test sequence Φ_j, and the k-th client model by applying two different versions of the *weighted summation fusion* (also called *sum rule*), which has the following general formula:

$$\xi_{j,k}^{(i)} \equiv S^{(i)}\left(\Phi_j, \mathbf{\Theta}_k\right) = a_{j,k}\eta_{j,k}^{(i)} + b_{j,k}\rho_{j,k}^{(i)}$$

where $a_{j,k}$ and $b_{j,k}$ are the weighting values and i specifies the identification or verification case. It is worth noting that, due to the properties of the logarithm function, the weighted summation fusion of the log-posterior probabilities is equivalent to the *weighted product fusion* (also called *product rule*) of the posterior probabilities.

The first strategy is the *equal weighting* of modalities: it is obtained by taking the average of the separate similarity scores, or equivalently by setting the weights as:

$$a_{j,k} = b_{j,k} = 0.5$$

for $\forall j,k$. This choice has an interesting probabilistic interpretation; in fact, if we assume that the features related to facial motion $\mathbf{X}_j^{(mtn)}$ and those to facial appearance $\mathbf{x}_j^{(app)}$ are statistically independent, then the *multimodal similarity scores for the identification task* are equal to the *joint log-posterior probabilities* of $\mathbf{X}_j^{(mtn)}$ and $\mathbf{x}_j^{(app)}$:

$$\xi_{j,k}^{(ID)} \equiv S^{(ID)}\left(\Phi_j, \mathbf{\Theta}_k\right) = \frac{1}{2}\log p\left(k \mid \mathbf{X}_j^{(mtn)}, \mathbf{x}_j^{(app)}\right) + \frac{1}{2}p(k)$$

except for an irrelevant translating factor, the a priori probability $p(k)$, which is not dependent on the test itself and it is already known before the recognition process.

The other fusion strategy is the *adaptive weighting* proposed by Chang et al. in [11], which automatically estimates the weights of each modality in a given test j as:

$$a_{j,k} = \frac{1}{A_{j,k}}\left(\frac{\eta_{j,k^{(1st)}}^{(i)} - \eta_{j,k^{(2nd)}}^{(i)}}{\eta_{j,k^{(1st)}}^{(i)} - \eta_{j,k^{(3rd)}}^{(i)}}\right)$$
$$b_{j,k} = \frac{1}{A_{j,k}}\left(\frac{\rho_{j,k^{(1st)}}^{(i)} - \rho_{j,k^{(2nd)}}^{(i)}}{\rho_{j,k^{(1st)}}^{(i)} - \rho_{j,k^{(3rd)}}^{(i)}}\right)$$

where $\eta^{(i)}_{j,k^{(l)}}$ and $\rho^{(i)}_{j,k^{(l)}}$ are the l best scores for the j-th test, and $A_{j,k}$ is a scaling factor to make the sum of weights equal to 1: $A_{j,k} = a_{j,k} + b_{j,k}$ for $\forall j,k$. The principle of this adaptive weighting is to evaluate the distribution of the similarity scores in each subsystem, and consider it as a measure of confidence for that subsystem; it follows that the more reliable is a subsystem, the higher should be its contribution on the final similarity score. In particular, the adaptive weighting suggested by Chang et al. calculates the ratio between the distance of the best score from: the second best one, and from the third best one; we also tried to replace $\eta^{(i)}_{j,k^{(3rd)}}$ and $\rho^{(i)}_{j,k^{(3rd)}}$ with the mean values of the respective scores (in each test), but we did not notice any significant difference from just using the third best ones.

V.C. Experimental results

Due to the absence of standard video databases suited for our approach, we assess the performance of our multimodal person recognition system on our database of Italian TV speakers; hence, we use the video version of the data set for the subsystem using unconstrained head and mouth motion information, and its image version for the one working on facial appearance. The interested reader can find a discussion on existing data sets, a description of our video database, and the structure of the enrolment and recognition subsets in Section VIII.A; for the image version of the database, the details on its generation and normalisation are explained in Section VIII.B. It is worth noting that all experimental results and relative comments are related to our small video database of Italian TV speakers, so that they should not be considered as absolute general conclusions.

In the following sections, we firstly introduce the *default configuration*, which obtains the best recognition results overall. Next, we evaluate the performances of our system in different experimental conditions, by varying: sources of biometric information and fusion strategies. After that, we compare our results with the state of the art eigenface technique, and finally we evaluate the discriminatory power of our method in a gender recognition application.

V.C.1. Default configuration

We denote the parameter configuration that attains the best overall recognition performance as the *default configuration*, and we use it as a reference throughout the experiments; considering that the multimodal recognition system is composed by two subsystems and one integration step, as illustrated in Figure 33, we specify the best parameter set for all of them.

The subsystem using head and mouth motion keeps the same default configuration of the one detailed in Section IV.C.1, except for the addition of two mouth features: the centred major and minor axes of the outer lip contour. Hence, an updated summary of the parameters for the default configuration of this subsystem is presented in Table 3.

PRE-PROCESSING	Facial landmarks	4 (eyes, nose & mouth)
	Colour space	RGB (red, green & blue)
	Video pre-processing	Histogram equalisation
	Distance metric for similarity scores	City-block distance
	Template size	19 pixel rows (or height)
		25 pixel columns (or weigth)
	Search window	12 x 12 pixels
	Template update	None ($\alpha = 0$)
FEATURE EXTRACTION	Geometrical normalisation of tracking signals	Centring using zero mean
	Head features	Normalised head positions
	Mouth features	Centred major and minor axis
	PCA reduction	No
	Dimensionality of feature space	10
MODEL ESTIMATION	GMM parameter estimation	Expectation-maximisation (EM)
	Gaussian components	4
	Initialisation	K-means
		Uniform weights
		Cluster means
		Uniform covariances
CLASSIFICATION	Number of background (cohort) models	2

Table 3: summary of the parameters for the default configuration of the recognition subsystem using head and mouth motion.

Then, in the default configuration of the subsystem using facial appearance, all images are firstly pre-processed with a *histogram equalisation*, colour component by colour component, to reduce the mismatches due to illumination variations. Next, the data set is represented by using the *NTSC colour space* (which consists of: luminance, hue and saturation), because it empirically provides more discriminative signals than the RGB does. Due to the problem of approximating high dimensional distributions with a limited amount of data (Section V.B.2), we are obliged to adopt serious restrictions on the dimension of the face space and the number of Gaussian components, in order to satisfy the minimum number of images recommended for a reliable GMM parameter estimation (Section IV.B.4). In fact, with 228 images per person in the enrolment subset, we should use an eigenspace of dimension 10 or less for being able to reliably train 2 components, and 8 or less for 3, which is excessively constraining because too much discriminative information is lost with such a reduced space. Hence, in the default configuration the client models are estimated using a single Gaussian component (which reverts on using a multivariate normal distribution), in a small face space of dimension 27, and the feature vectors are calculated by *whitening* the projection coefficients. Finally, the impostor models for verification are approximated by taking the average of the best 2 background (or cohort) models. A summary of the parameters for the default configuration of the recognition subsystem using facial appearance is proposed in Table 4.

		NORMALISED DATABASE
DATABASE	Database name	Image database of Italian TV speakers
	Normalisation	Accurate: in-plane rotated and aligned
PRE-PROCESSING	Image size	32 pixel rows (or height)
		32 pixel columns (or weigth)
	Image resizing interpolation method	Nearest neighbour
	Image pre-processing	Histogram equalisation
	Colour space	NTSC (luminance, hue & saturation)
	Vertical mirroring	No
FEATURE EXTRACTION	Image space reduction method	Centered PCA
	Subspace dimension	27
	Whitening of feature vectors	Yes
MODEL ESTIMATION	GMM parameter estimation	Direct mean and covariance calculation
	Gaussian components	1
CLASSIFICATION	Number of background (cohort) models	2

Table 4: summary of the parameters for the default configuration of the recognition subsystem using facial appearance.

Concerning the score integration step, the best results have been obtained by choosing an *equal weighting* of the previous two complementary recognition subsystems.

Finally, as we did in the experimental section of Chapter IV, in the following experiments we express the results relative to the default configuration of the multimodal recognition system with a blue colour line, in order to simplify the understanding and comparison between different graphs.

V.C.2. Recognition results in diverse experimental conditions

In this section, we assess the performance of our person recognition system by testing it in diverse experimental conditions, with various sources of biometric information and fusion strategies. We present and comment a selection of experiments, preferring those that better illustrate the properties of our approach, and help understanding the choices towards the best configuration.

Concerning the measures of performance, we express the identification results by reporting the *correct identification rates (CIRs)*, and by plotting the *cumulative (correct) match scores (CMSs)* as a function of the M best matches retained (Section II.E.2). For the verification scenario, we report the *equal error rates (EERs)* and we show the *receiver operating characteristic (ROC)* curves, which offer a global description of the system from low to high security applications (Section II.E.1).

We firstly compare the discriminatory power of the different sources of biometric information. We remind from Chapter IV that the recognition subsystem using only unconstrained head motion obtains good results, with a CIR of 90.4% and an EER of 3.0%; then, we notice that combining the head and the mouth motion information improves the scores to: 93.3% of CIR and 2.6% of EER. However, it is with the integration of facial appearance that the multimodal recognition system obtains the best performance overall; in fact, the spatial subsystem alone already shows good results, with a CIR of 93.3% and an EER of 2.2%, but it achieves the best performance when combined with the temporal information in the default configuration: still 93.3% of CIR, but 2.1% of EER and a small overall improvement in the verification scores.

Figure 35 provides a complete overview of the recognition results: the experiments show that facial appearance conveys the most discriminative information, followed by head and mouth motion. We explain this outcome by keeping in mind that the noise caused by our automatic tracking process reduces the recognition capabilities of natural head motion (Sections IV.C.2 and IV.C.5); likewise, the low quality of our video database makes the precise localisation of the outer lip contour really challenging for the lip segmentation algorithm, and in some cases the mouth features are excessively noisy. Though, the main reason behind the weakness of the mouth information compared to the head one is that the feature space of the temporal subsystem is mainly composed by head parameters (8 over 10), and only marginally by the mouth ones (only 2). In conclusion, by carefully examining the curves in Figure 35 we notice that each addition of biometric identifiers improves the recognition results, and that the fusion of the facial appearance information with both the head and mouth one provides the best results overall.

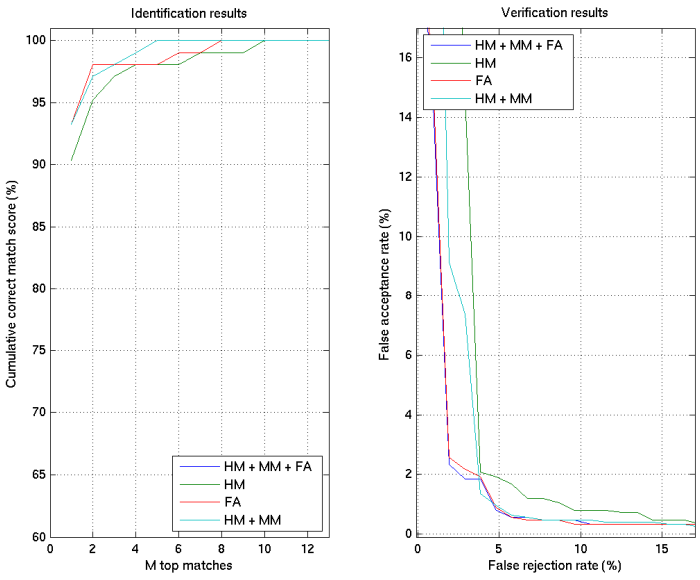

Figure 35: recognition results with different sources of biometric information.

Afterwards, we study the efficiency of the two different fusion strategies on the performance of the multimodal recognition system. The experiments on person identification show that the two weighting schemes obtain the same results, with a CIR of 93.3%. On the other hand, the equal weighting seems better suited in a verification scenario, where it presents a smaller EER than the adaptive fusion strategy: 2.1% rather than 2.9%.

The plots for the CMSs and the DETs are illustrated in Figure 36 and confirm the numerical results above: in general, the adaptive weighting should be preferred in an identification scenario, while the equal weighting is the best choice for a verification one. To explain this outcome we need to consider the *decision rules* of the two operational modes (Sections II.E.1 and II.E.2): in an identification problem, the system chooses the class with the best similarity score independently test by test; on the other hand, in a verification one each similarity score is compared with a threshold value, which remains the same for all the tests. For this reason, the adaptive weighting top performs in the identification scenario, because it is supposed to estimate the best weights for every video; at the same time, this test-dependent weighting strategy becomes unpredictable for the whole data set, so the similarity values are less commensurate and it is more difficult to find an optimal common threshold value. In our experimental situation, we preferred the equal weighting for our default configuration, due to the following reasons: it shows a significant improvement in verification results obtaining at the same time the best identification outcomes, it is computationally more efficient, and it provides a more balanced and stable fusion of modalities.

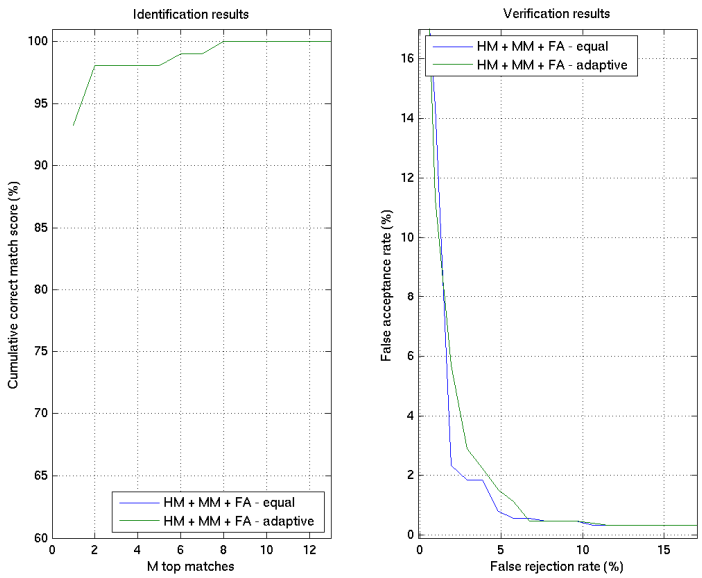

Figure 36: recognition results with different fusion strategies.

V.C.3. Comparison with the eigenface technique

We compare our multimodal person recognition system with the state of the art eigenface technique [87], which exploits only facial appearance. For our experiments, we keep the same eigenface implementation as in Section IV.C.4, whose parameters are reported in Table 5.

		NORMALISED DATABASE	RAW DATABASE
DATABASE	Database name	Image datab. of Italian TV speakers	Image datab. of Italian TV speakers
	Normalisation	Accurate: in-plane rotated and aligned	None
PRE-PROCESSING	Image size	32 pixel rows (or height)	48 pixel rows (or height)
		32 pixel columns (or weigth)	61 pixel columns (or weigth)
	Resizing interpolation method	Nearest neighbour	Nearest neighbour
	Image pre-processing	Histogram equalisation	Histogram equalisation
	Colour space	NTSC (luminance, hue & saturation)	NTSC (luminance, hue & saturation)
	Vertical mirroring	No	No
FEATURE EXTR.	Image space reduction method	Centered PCA	Centered PCA
	Subspace dimension	243	243
	Whitening of feature vectors	Yes	Yes
MODEL ESTIM.	Client model generation method	Centroid vector (average of features)	Centroid vector (average of features)
CLASSIFICATION	Similarity measure	Based on cosine distance	Based on cosine distance

Table 5: summary of the parameters for our eigenface implementation.

The recognition results for our multimodal approach are calculated using the default configuration and the database of Italian TV speakers: the subsystem working on unconstrained head and mouth motion exploits the video data set (Section VIII.A), while the other subsystem using facial appearance works with the normalised image version of the database (Section VIII.B). In the previous experiments we have already seen that the integration of biometric sources of information improves the discriminatory power of the system exploiting only head motion: the CIR increases from 90.4% to 93.3% and the EER decreases from 3.0% to 2.1%. Then, as we did in Section IV.C.4, we evaluate the performance of the eigenface technique on both the normalised and not normalised image databases. We remind that in the first favourable case, it achieves perfect recognition, with 100.0% of CIR and 0.0% of EER; though, in the second somewhat unfavourable case its results are poor: 69.2% of CIR and 10.8% of EER. This time, we also test the eigenface technique with a small face space of dimension 27, which is surely an adverse situation but it allows a fair comparison with our spatial recognition subsystem, and with our multimodal system; in this particular condition, its recognition results are noticeably worse, even if using the normalised data set: the CIR is 65.4% and the EER is 8.7%.

Figure 37 provides a practical and complete overview of these experimental results: the multimodal person recognition system still performs in between the favourable and unfavourable eigenface approach, but it also noticeably improves the results of the one exploiting only head motion. Nevertheless, we believe that the only reason why the eigenface technique still outperforms both its probabilistic extension and the multimodal recognition system, is the difficulty of estimating high dimensional distributions with a limited amount of training data; in particular, in our experimental conditions the spatial subsystem must adopt a small face space of dimension 27, which approximates only coarsely the facial appearance manifold, and so it possesses a reduced discriminatory power. In fact, the recognition results for the eigenface technique working in the same reduced face space (violet curves) corroborate our assertion: in this more appropriate comparison, it performs clearly worse than both our spatial subsystem and our multimodal one. For this reason, we are confident that in the future the use of larger databases will confirm the superior performances of our approach; in fact, the availability of a bigger enrolment subset would allow our system to use a bigger face space and a more complex GMM modelling (with more Gaussian components), and consequently to fully exploit the complementary and more discriminative nature of its multimodal identifiers. In conclusion, these empirical results let us believe that not only facial appearance but also head and mouth motion possess a potentially relevant discriminatory power, and that the integration of different sources of biometric information from video sequences is the key strategy to develop more accurate and reliable recognition systems.

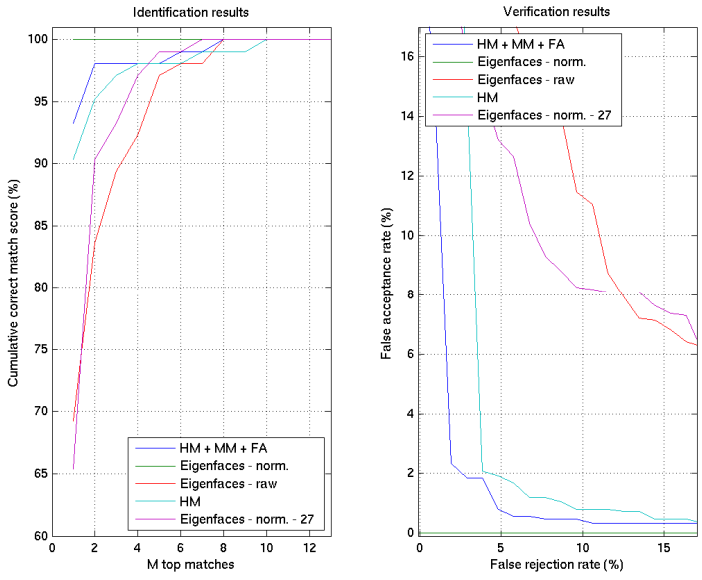

Figure 37: comparison ofperson recognition results between: the proposed method, eigenfaces, and the system using only head motion.

V.C.4. Gender recognition results

To conclude these experimental sections, we would like to evaluate the performance of our approach when applied to a different scenario: a gender recognition application; more precisely, we want to compare the gender discriminatory power of the multimodal system, integrating facial appearance and natural head and mouth motion, with that of the eigenface technique, exploiting facial appearance alone. Therefore, similarly to what we did in Section V.C.3 for identity, we consider an equivalent experimental set-up that is adapted to the gender recognition task, with the same eigenface implementation, databases of Italian TV speakers and dimensions of face spaces. It is worth noting that due to the particular nature of this recognition problem, which consists of only two classes, the CIRs and EERs are directly related: $\eta^{(CIR)} = 100\% - \xi^{(EER)}$.

Similarly to what we have seen for the person recognition scenario, we notice that the integration of multiple sources of biometric information clearly improves the performance of the individual modalities. In fact, reminding that the system exploiting only head motion obtains poor gender recognition scores with a CIR of 84.6% and an EER of 15.4% (Section IV.C.6), we observe that the addition of mouth motion and then of facial appearance in our multi-biometric system increases the CIR to 96.2% and 99.0%, and decreases the EER to 3.8% and 1.0% respectively. Moreover, this time our gender recognition approach performs in between the perfect recognition scores (100.0% of CIR and 0.0% of EER) of eigenfaces in its favourable condition, and the poor results in its unfavourable one, which are: a CIR of 89.4% and an EER of 10.6%.

By looking at Figure 38 we can visually evaluate the benefit of integrating multiple sources of biometric information; these graphs show that, even if head motion is not such an important discriminative identifier for gender recognition applications, it can still achieve excellent results if supported by the mouth motion information, and particularly by the facial appearance one[3]. In fact, these experiments are a clear demonstration of the advantage of multi-biometrics, in which complementary sources of biometric information can increase the accuracy and augment the reliability of the resulting multimodal system, by taking advantage of their redundant and richer information to compensate their individual weaknesses. We finally remember that due to the noisy tracking signals and the noisy mouth parameters (Section V.C.2) in the temporal subsystem, along with the reduced dimensionality of the face space and complexity of the GMM modelling in the spatial one (Section V.C.3), our recognition system is still far from its optimal working condition, so the potential discriminatory power of facial appearance and head and mouth motion for gender recognition is probably higher than the one established in these experiments.

[3] In other experiments (not reported here), we also divided our database into two random classes mixed in gender and consisting of multiple individuals, to evaluate the discriminatory power of the head and mouth motion information for gender recognition. The CIRs obtained for these random sets were between 60% and 70%, a bit better than the random choice (the system is still learning some patterns, and using some information for classification), but definitively worse than the corresponding CIR for gender recognition, 96.2%. The outcome of these experiments also supports the assertion that the head and mouth motion information contains some discriminative information related to identity and gender; information that can be used in person or gender recognition applications.

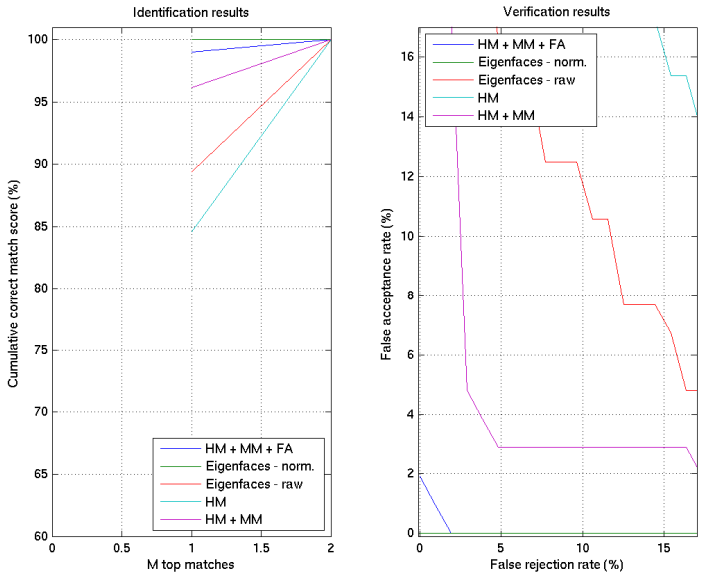

Figure 38: comparison ofgender recognition results between: the proposed method, eigenfaces, and the system using only head motion.

V.D. Concluding summary

In this chapter we proposed a multimodal extension of our person recognition system; in particular, we successfully integrated the head motion information with mouth motion and facial appearance, by taking advantage of a unified probabilistic framework. In fact, we developed a new temporal subsystem that had an extended feature space enriched by some additional mouth parameters; at the same time, we introduced a complementary spatial subsystem based on a probabilistic extension of the original *eigenface approach*. In the end, we implemented an integration step to combine the similarity scores of the two parallel subsystems, using a suitable *opinion fusion* (or *score fusion*) strategy.

In the experimental section we assessed the performance of our multimodal approach and we deduced the following considerations. First of all, the potential discriminatory power of the biometric identifiers integrated in our multimodal recognition system is probably higher than what is established by our experiments; in fact, our approach cannot be tested in its optimal condition due to: noisy tracking signals and noisy mouth parameters in the temporal subsystem, along with the reduced dimensionality of the face space and complexity of the GMM modelling in the spatial one. However, we observe that the integration of multiple sources of biometric information noticeably improves the performance of the separate unimodal systems, and that facial appearance conveys the most discriminative information, followed by head and mouth motion. After all, our multimodal approach obtains good person recognition results and very good gender recognition ones, and it performs closely to the eigenface technique in its most favourable testing condition. Finally, we believe that the integration of different sources of biometric information extracted from video sequences is the key strategy to develop more accurate and reliable recognition systems.

Chapter VI. Tomofaces: spatio-temporal facial features for recognition

VI.A. Introduction

In Chapter III we have seen that the video data does not provide only abundant spatial information but also the temporal one, and that the face is now considered as a hybrid identifier. Then, in Chapter V we developed a multimodal recognition system, which exploited different sources of biometric information present in video sequences, and we noticed that, even if facial appearance still conveys the most discriminative information, its performance can be improved through the integration of some head and mouth motion features. All these elements motivate us to conclude that there is a real interest in investigating novel spatio-temporal strategies for video person and gender recognition.

In addition, we are aware that the performances of the majority of biometric systems using image and video data strongly depend on the accuracy of some complex pre-processing. In fact, it is well documented that recognition approaches exploiting facial appearance are extremely sensitive to inter-frame variations, like inconsistent facial alignments or changes in illumination, pose and head size; however, the precise and automatic normalisation of video frames is hard to achieve in practice, and it is computationally expensive. Concerning the exploitation of the behavioural information of the face, the situation is even more critical; in fact, the need for a temporal synchronisation of video chunks, which is necessary for a consistent matching between different gestures, has probably prevented an efficient use of the video temporal information for recognition, and put back the development of new strategies.

For these reasons, we investigated a practical method for extracting the spatio-temporal biometric information from video sequences, and we used it to discriminate identity and gender. Inspired by the research on *discrete video tomography (DVT)* [1][39] for camera work estimation, we developed a recognition system called *tomofaces*, which applies the temporal X-ray transformation of a video sequence to summarise the facial motion and appearance information of a person into a single X-ray image. The main advantages of this approach are that it does not require a complex pre-processing for accurate spatial normalisations, and that it avoids the temporal synchronisation problem; in particular, it is well suited to our unconstrained recognition scenario, where there is no prior knowledge on the explicit gesture that each user is doing in every video sequence. It is worth noting that in our experimental framework there is no camera motion, because the camera is fixed, there are no zooms or changes in scale, and that the depth variation of the head movement is insignificant, because the camera is far. This situation implies that all motion that can be extracted from the sequences is relative to the behaviour of the individual, and that in our system the discrete video tomography is no more applied for camera motion estimation, but for the extraction of spatio-temporal features for recognition.

The remainder of this chapter is organised in two main sections: one theoretical part that details the structure of our tomoface recognition system, and one experimental part that thorough fully evaluates the performances of our approach in various conditions.

VI.B. Proposed method

The architecture of the tomoface approach that exploits spatio-temporal facial features for recognition is illustrated in Figure 39, and closely resembles the one for the general biometric system, which has been introduced in Section II.D.

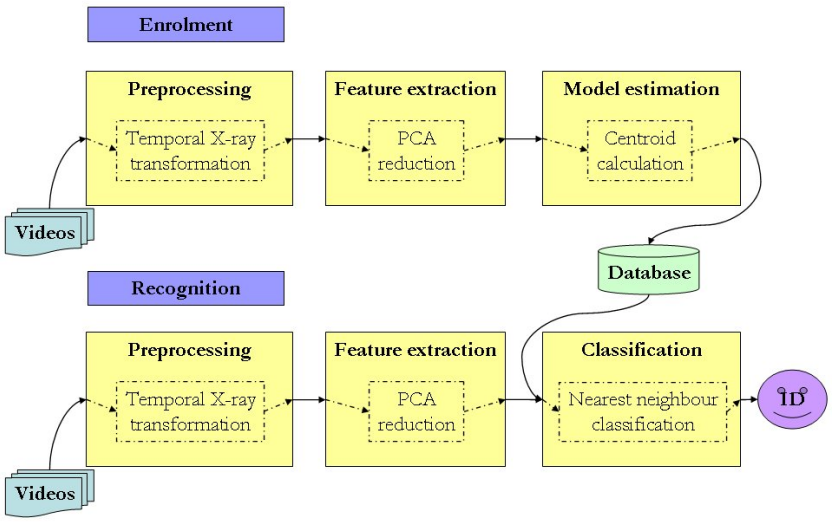

Figure 39: architecture ofthe tomoface approach, which exploits spatio-temporal facial features for recognition.

The pre-processing step firstly generates the X-ray images, by applying the *discrete video tomography (DVT)* introduced by Akutsu and Tonomura [1] to the input video sequences. Then, the X-ray image space is reduced into a low dimensional space, in which spatio-temporal features are extracted to provide a better discriminative representation. After that, the enrolment module estimates each client model by calculating its cluster centre, and in the end person (or gender) recognition is achieved through a nearest neighbour classifier. The four steps of our system are detailed in the following sections.

VI.B.1. Pre-processing: temporal video X-ray transformation

The pre-processing step computes the *temporal video X-ray transformation* of a sequence, in order to summarise the facial motion and appearance information of a person into a single image; more precisely, it generates a vector, $\mathbf{s} \in \mathrm{N}^{RC}$, representing the *video X-ray image* from each colour sequence of length T and frame size $R \times C$: $\Phi = \left\{ \Phi_t \in \mathrm{N}^{R \times C \times 3} \mid t = 1, \ldots, T \right\}$.

First of all, this step applies a contrast enhancement to all video frames, like a *histogram equalisation* or a *contrast stretching* (colour component by colour component) [28], that is useful to reduce the impact of inter-image illumination and colour variations. After that, it converts the sequence from the RGB colour space to the greyscale one, and then it generates the *edge map sequence*, $\Psi = \left\{ \Psi_t \in \mathrm{N}^{R \times C} \mid t = 1, \ldots, T \right\}$, by applying the *Canny edge finding method* [8] frame by frame. We remind that the Canny edge detection algorithm extracts the local maxima of the gradient by using two thresholds, in order to detect both strong and weak edges; in particular, the weak edges are included in the binary edge map only if they are connected to strong edges, which improves their detection accuracy.

Afterwards, the *temporal video X-ray transformation* consists of adding up the edge map frames, Ψ_t, along the temporal axis, in order to generate the *video X-ray image* of the sequence:

$$\Gamma = A \sum_{t=1}^{T} \Psi_t$$

where A is a scaling factor, which is constant for the whole database and is used to adjust the upper range values of the video X-ray images. In Figure 40, there is a visual example that summarises the major phases of the temporal video X-ray transformation. By looking at the lower left picture, corresponding to the video X-ray image, Γ, it is possible to notice that the static textured background generates very dark areas and very vivid contours; this information is not related to facial motion or appearance, and may negatively affect the discriminatory power of the X-ray image. For this reason, the pre-processing step allows filtering Γ in order to attenuate its brightest background contours, by putting to black all the pixels above a threshold value, θ: $\gamma_{r,c} = 0$ for $\{(r,c) | \gamma_{r,c} > \theta\}$. It is worth noting that the choice of the threshold value is quite important, because it represents a trade-off between the amount of background contours that are removed by filtering, and the amount of head and mouth motion that is preserved; in fact, the lower is the threshold, the stronger is the attenuation of the background, but also the higher is the loss of useful spatio-temporal information.

Figure 40: example of the temporal video X-ray transformation; from left to right, starting from the top: original frame, edge map frame, video X-ray image, attenuated video X-ray image.

Finally, the amount of data can be optionally incremented by mirroring each attenuated X-ray image along its vertical axis (a process called *vertical mirroring*), before that the ordinary image vectorisation takes place.

VI.B.2. Feature extraction: PCA reduction

The feature extraction step isolates the discriminative information that characterises the individual and discards the irrelevant one, by transforming the vectorised X-ray image $\mathbf{s} \in \mathrm{N}^{RC}$ into the corresponding feature vector: $\mathbf{x} \in \Re^{D}$. First of all, it applies a linear transformation from the high dimensional *X-ray image space*, N^{RC}, to a lower dimensional space, $\Re^{D}$, which is much smaller: $D << RC$. More precisely, each vectorised X-ray image $\mathbf{s} \in \mathrm{N}^{RC}$ is approximated with its projection in the reduced space $\mathbf{v} \in \Re^{D}$ by the following *linear transformation*:

$$\mathbf{v} = \mathbf{W}^{T}(\mathbf{s} - \boldsymbol{\mu})$$

where $\mathbf{W} \in \Re^{RC \times D}$ is a projection matrix with orthonormal columns, and $\boldsymbol{\mu} \in \Re^{D}$ is the mean X-ray image vector of the whole training set:

$$\boldsymbol{\mu} = \frac{1}{J}\sum_{j=1}^{J}\mathbf{s}_{j}$$

in which J is the total number of sequences in the training set, and $\mathbf{s}_{j} \in \mathrm{N}^{RC}$ is the vectorised X-ray image belonging to video Φ_{j}.

The *optimal projection matrix* $\mathbf{W}$ is computed using the *principal component analysis (PCA)* (also called the *Karhunen-Loeve transform (KLT)*) [22], which has the property of optimally representing the distribution of data in the root mean squares sense; the details on the calculation of $\mathbf{W}$ can be found in [87] and in Section III.B.1.

Once the vectorised X-ray image is projected into the reduced space, the corresponding feature vector $\mathbf{x} \in \Re^{D}$ is generated by choosing either:

1. The *projection in the reduced space*: in this case, $\mathbf{x} = \mathbf{v}$.
2. The *whitened projection in the reduced space*: the *whitening* process (Section III.B.1) rescales the projection coefficients v_{d} to counterbalance the overweighting of the low frequencies as following: $x_{d} = \frac{v_{d}}{\sqrt{\lambda_{d}}}$ for $d = 1, \ldots, D$, in which λ_{d} is the d-th largest eigenvalue.

VI.B.3. Model estimation and classification

The model estimation and classification steps of the tomoface recognition approach are similar to those of the original eigenface technique [87].

The personal models are characterised by representative points in feature space, which summarise the distribution of feature vectors belonging to each client; in particular, for a given individual k the model estimation step calculates his *cluster centre* $\mathbf{g}_{k} \in \Re^{D}$, by taking the average (the centroid) or the median feature vector of the user training data belonging to his enrolment subset.

Then, the recognition task is achieved by using a *nearest neighbour classifier*, which compares an unknown feature vector $\mathbf{x} \in \Re^D$ with all the client models stored in the system, $\{\mathbf{g}_k \in \Re^D \mid k = 1, \ldots, K\}$, and looks for the closest match. In this situation, the similarity measure is inversely proportional to one among these simple distance metrics in feature space:

- *City-block distance* (L_1): $d^{(L_1)}(\mathbf{x}, \mathbf{g}_k) \equiv \sum_{d=1}^{D} |x_d - g_{k,d}|$.
- *Euclidean distance* (L_2): $d^{(L_2)}(\mathbf{x}, \mathbf{g}_k) \equiv \sqrt{\sum_{d=1}^{D} (x_d - g_{k,d})^2}$.
- *Cosine distance*: $d^{(\cos)}(\mathbf{x}, \mathbf{g}_k) \equiv \frac{1}{2} - \frac{\mathbf{x}^T \mathbf{g}_k}{2\|\mathbf{x}\|\|\mathbf{g}_k\|} \equiv \frac{1}{2} - \frac{\sum_{d=1}^{D} x_d g_{k,d}}{2\sqrt{\left(\sum_{d=1}^{D} x_d^2\right)\left(\sum_{d=1}^{D} g_{k,d}^2\right)}}$.

VI.C. Experimental results

Due to the absence of standard video databases suited for our approach, we assess the performance of our person recognition system on our video database of Italian TV speakers: please refer to Section VIII.A for a discussion on existing data sets, a description of our database, and the structure of the enrolment and recognition subsets. It is worth noting that all experimental results and relative comments are related to our small video database of Italian TV speakers, so that they should not be considered as absolute general conclusions.

In the following sections, we firstly introduce the *default configuration*, which obtains the best recognition results overall. Then, we evaluate the performance of our system in a person recognition scenario, and we compare our results with the state of the art eigenface technique and the multimodal recognition system of Chapter V. Finally, we evaluate the discriminatory power of our method in a gender recognition application.

VI.C.1. Default configuration

We denote the parameter configuration that attains the best overall recognition performance as the *default configuration*, and we use it as a reference throughout the experiments; a summary of the parameters for the default configuration of the tomoface recognition system is presented in Table 6.

DATABASE	Database name	Video database of Italian TV speakers
	Normalisation	None
PREPROCESSING	Image pre-processing	Histogram equalisation
	Colour space	Greyscale
	Background attenuation threshold	66%
	Image size	64 pixel rows (or height)
		81 pixel columns (or weigth)
	Image resizing interpolation method	Nearest neighbour
	Vertical mirroring	No
FEATURE EXTRACTION	Image space reduction method	Centered PCA
	Subspace dimension	103
	Whitening of feature vectors	Yes
MODEL ESTIM.	Client model generation method	Centroid vector (average of features)
CLASSIFICATION	Similarity measure	Based on cosine distance

Table 6: summary of the parameters for the default configuration of the tomoface recognition system.

In the default configuration of the tomoface approach, all video frames are firstly pre-processed with a *histogram equalisation*, colour component by colour component, to reduce the mismatches due to illumination variations. Next, the temporal X-ray transformation is applied on the greyscale version of the sequence, and the threshold value for the attenuation of background contours in the video X-ray image is set to 66% of the grey level range. After that, the filtered X-ray image is resized to 64 pixel rows (height) and 81 pixel columns (width), by using the *nearest neighbour interpolation method*; in fact, we have empirically found that this image size best isolates the discriminative information of the original X-ray image, once it is approximated with its whitened projection in the reduced space of dimension 103 (the maximum possible with our training data). Then, the client models are registered into the system by using their *centroid vectors*, which are calculated by taking the average of the feature vectors in the enrolment subset; in the end, recognition is achieved by using a *nearest neighbour classifier* with *cosine distances* in (the whitened) reduced space.

Finally, as we did in the experimental sections of Chapter IV and Chapter V, in the following experiments we express the results relative to the default configuration of the tomoface recognition system with a blue colour line, in order to simplify the understanding and comparison between different graphs.

VI.C.2. Person recognition results

In this section we assess the performance of our person recognition system, and we compare it with the state of the art eigenface technique [87] and the multimodal recognition system of Chapter V. For our experiments, we keep the same eigenface implementation as in Section IV.C.4, and the same default configuration for the multimodal approach as in Section V.C.1.

Concerning the measures of performance, we express the identification results by reporting the *correct identification rates (CIRs)*, and by plotting the *cumulative (correct) match scores (CMSs)* as a function of the M best matches retained (Section II.E.2). For the verification scenario, we report the *equal error rates (EERs)* and we show the *receiver operating characteristic (ROC)* curves, which offer a global description of the system from low to high security applications (Section II.E.1).

The person recognition results of our tomoface approach using spatio-temporal facial features are pretty good: its CIR is 74.0% and its EER is 7.7%. Though, they are not as good as those discussed in Chapter V for the multimodal recognition system, which obtains a CIR of 93.3% and an EER of 2.1%. Eigenfaces is still the top performing technique when tested in its optimal condition using the normalised image database of Italian TV speakers: in this case it achieves the perfect recognition outcome with 100.0% of CIR and 0.0% of EER. On the other hand, the same technique shows poor discriminative properties when tested using the raw (not normalised) version of the data set; in this unfavourable condition, it has a CIR of 69.2% and an EER of 10.8%.

By looking at Figure 41 we can easily compare the identification and verification results of the various approaches: there is no doubt that the person recognition system using video tomography performs better than the eigenface one tested with raw images, but it is noticeably inferior to the multimodal recognition strategy, or to eigenfaces in its optimal experimental condition. In particular, in those common situations where the frame normalisation is not accurate, the tomoface approach performs better than the eigenface technique, which corroborates the assertion that spatio-temporal facial features are more discriminating than sole appearance ones. However, these experiments also reveal that the biometric information relative to facial appearance and motion is more distinctive if it is extracted separately through facial pictures and temporal signals, rather than when it is combined into video X-ray images; consequently, a post-mapping score integration like in the multimodal system of Chapter V should be preferred to a pre-mapping fusion at feature level, as in the tomoface approach. In conclusion, even if the comparison with the eigenface technique in its optimal condition is unfair, we are aware that our system is not as accurate as the best methods exploiting facial appearance [69], and we do not still regard the spatio-temporal facial features as a practical alternative to appearance ones.

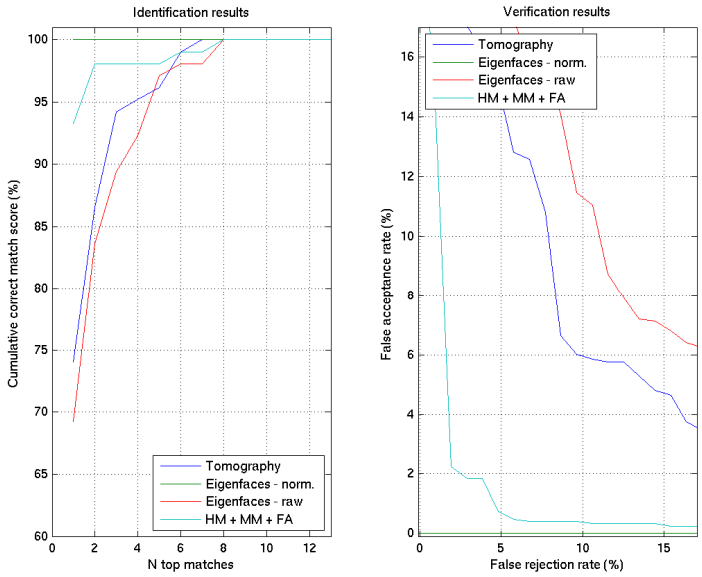

Figure 41: comparison of person recognition results between: the proposed method, eigenfaces, and the multimodal system of Chapter V.

VI.C.3. Gender recognition results

We also evaluate the performance of the tomoface system in a gender recognition application; hence, in this section we compare the gender discriminatory power of spatio-temporal facial features with that of facial appearance alone, and with that of facial appearance combined with head and mouth motion, similarly to what we did in Section VI.C.2 for identity. Therefore, we consider an equivalent experimental set-up that is adapted to the gender recognition task, with the same eigenface implementation and databases of Italian TV speakers. It is worth noting that due to the particular nature of this recognition problem, which consists of only two classes, the CIRs and EERs are directly related: $\eta^{(CIR)} = 100\% - \xi^{(EER)}$.

In a gender recognition scenario, the tomoface approach obtains very good results: it presents a CIR of 98.1% and so an EER of 1.9%. It also performs very close to its alternative strategies: we remind that the system exploiting facial appearance and head and mouth motion has a CIR of 99.0% and an EER of 1.0%, and the eigenface technique achives perfect recognition when in its favourable experimental condition, with a 100.0% of CIR and 0.0% of EER. On the contrary, the same eigenface strategy applied on a not normalised database presents poor results: a CIR of 89.4% and an EER of 10.6%.

By looking at Figure 42 we notice that, similarly to the person recognition case, the tomoface system performs better than the eigenface technique when it is applied on raw images, but this time its scores are really close to those belonging to the multimodal recognition approach, or to eigenfaces in its optimal experimental condition. All these elements confirm that spatio-temporal facial features are more discriminating than sole appearance ones, and let us believe that in a gender recognition scenario the biometric information relative to facial appearance and motion possesses almost the same discriminatory power, if it is extracted separately through facial pictures and temporal signals, or rather if it is combined into video X-ray images. In conclusion, we suppose that spatio-temporal features are potentially more discriminating than sole facial appearance, and we consider the tomoface technique promising, even if it is still too immature for real applications.

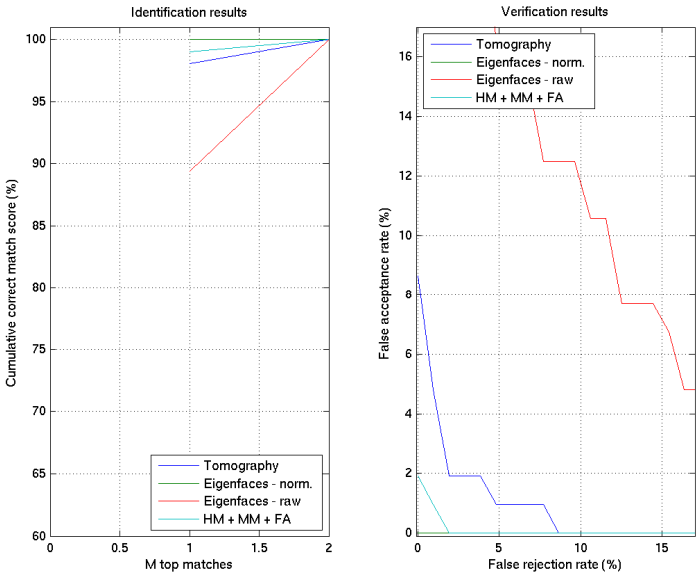

Figure 42: comparison of gender recognition results between: the proposed method, eigenfaces, and the multimodal system of Chapter V.

VI.D. Concluding summary

In this chapter we investigated a practical method for extracting novel spatio-temporal facial features from video sequences, which were used to discriminate identity and gender. For this purpose we developed a recognition system called *tomofaces*, which applied the *temporal X-ray transformation* of a video sequence to summarise the facial motion and appearance information of a person into a single X-ray image. Then, we detailed the linear projection from the X-ray image space to a low dimensional feature space, the estimation of the client models obtained by computing their cluster representatives, and the recognition of identity and gender through a *nearest neighbour classifier* using distances.

The experiments run to evaluate the performance of the tomoface approach showed that it achieves pretty good person recognition results and very good gender recognition ones, and supported us for the following conclusions. Firstly, when video frames are not accurately normalised, the tomoface recognition system performs better than the eigenface technique, so the spatio-temporal facial features seems more discriminating than the sole appearance ones. However, the biometric information extracted separately through facial pictures and temporal signals, as in the multimodal system of Chapter V, appears to be more distinctive than when it is combined into video X-ray images. Finally, the spatio-temporal facial features do not represent a practical alternative to appearance ones yet; in fact, even if they possess a relevant potential discriminatory power, their calculation is still immature and needs some more investigation.

Chapter VII. Conclusion and perspectives

VII.A. Concluding summary

After having introduced the discipline of biometrics and its evolution towards multi-biometrics in Chapter II, we reviewed the literature on those person recognition strategies exploiting facial video information in Chapter III. We observed that only recently the attention of the scientific community has been attracted towards the use of facial video information for person recognition. We also noticed that the research on this domain has been mostly focused on developing straightforward extensions of image-based approaches, which exploit only the spatial information in video sequences; furthermore, most of temporal strategies took only advantage of the evolution of facial appearance over time. We concluded that the use of the face as a hybrid identifier, for example by exploiting facial appearance and motion for recognition, was still a largely unexplored topic.

In Chapter IV we presented a novel person recognition system that exploited the unconstrained head motion information, obtained by tracking a few facial landmarks in the image plane. We saw how the tracking signals automatically extracted by our system were not very accurate; for this reason, their potential discriminatory power and the performance of our recognition approach were significantly reduced. Nevertheless, we remarked that our biometric system was able to achieve good person recognition results, but poor gender recognition ones. We deduced that natural head motion possessed enough discriminatory power to be used as a possible biometric in recognition applications, but it was not yet a practical alternative to facial appearance.

In Chapter V we proposed a multimodal extension of our person recognition system; in particular, we successfully integrated the head motion information with mouth motion and facial appearance, by taking advantage of a unified probabilistic framework. We remarked that the potential discriminatory power of the biometric identifiers integrated in our multimodal recognition system was probably higher than what has been established by our experiments; in fact, our approach could not be tested in its optimal condition due to: noisy tracking signals, noisy mouth parameters, the reduced dimensionality of the face space and complexity of the GMM modelling. However, we observed that the integration of multiple sources of biometric information noticeably improved the performance of the separate unimodal systems, and that facial appearance conveyed the most discriminative information, followed by head and mouth motion. After all, our multimodal approach obtained good person recognition results and very good gender recognition ones, and it performed closely to the eigenface technique in its most favourable testing condition.

In Chapter VI we investigated a practical method for extracting novel spatio-temporal facial features from video sequences, and we developed a recognition system called *tomofaces*, which applied the temporal X-ray transformation to summarise the facial motion and appearance information into a single X-ray image. This approach achieved pretty good person recognition results and very good gender recognition ones, and it performed better than the eigenface technique when video frames were not accurately normalised. We deduced then that spatio-temporal facial features seemed more discriminating than the sole appearance ones; however, the biometric information extracted separately through facial pictures and temporal signals appeared to be more distinctive than when it was combined into video X-ray images. In the end, we concluded that the calculation of these novel spatio-temporal facial features was still immature and needed some more investigation.

We completed this doctoral dissertation with the appendices in Chapter VIII, where we detailed our database of Italian TV speakers in its video and image formats.

VII.B. Future works

So far, the issue of recognising people by using head motion, mouth motion and facial appearance from video sequences has been intensively discussed. Despite its promise, which has been shown in this thesis, this work has some limitations; in this section we discuss these limitations and possible directions for future research.

First of all, the recognition results presented in this thesis should be validated using other databases. In particular, it would be necessary to assess the performance of our biometric systems with a larger dataset, containing far more individuals. After that, it could be interesting to evaluate the discriminatory power of our approaches in scenarios that are different from the one proposed here with professional TV speakers; for example, we would like to study the impact of the stress and various emotional states on the behavioural features of everyday users. Unfortunately, as far as we know there are no other databases available that could be suited to evaluate our techniques (Chapter VIII).

Then, to improve the performance of our recognition system using head motion, which has been presented in Chapter IV, we should increase the accuracy of its tracking signals by implementing a more robust tracker. In fact, in this dissertation we have verified that the discriminatory power of the tracking signals is considerably affected by their precision, and that the signals extracted using a template matching technique possess a significant noise; hence, we are convinced that implementing a more accurate technique will surely provide better recognition results. In addition, it would also be of interest to study and take advantage of the head motion information at a higher level, for example at a gesture level. A possible strategy could include a local analysis of the tracking signals using a sliding window, coupled with a modelling of the distinct head gestures through multiple GMMs or HMMs. In this case, however, it would not be possible to avoid the temporal synchronisation problem between similar gestures, and some more issues would also arise, like: the definition of those "fundamental" head gestures that could be relevant for recognising identities, the automatic segmentation of video sequences based on those core gestures, and the bigger amount of data necessary for training multiple GMMs or HMMs.

Afterwards, the multimodal biometric system presented in Chapter V can be enhanced in several ways. The temporal subsystem may be improved by increasing the accuracy of the head tracking signals, or the precision of the segmented lip contours; in the former case, we have already proposed to implement a more robust tracking algorithm. In the second case, the best solution would be to exploit a video database of higher quality, because we observed that the main cause of error in the localisation of the lip contour is probably the low quality of the video frames; otherwise, it may be possible to refine the lip segmentation process, by investigating some additional error correction strategies, or to explore other image segmentation techniques [15][51], like *active contours*. Furthermore, it could be interesting to increase the present feature space with some more behavioural parameters from facial motion. One possibility is to study the discriminative properties of the lip curvature, area or shape variation over time; otherwise, it may worth exploring the extraction of some features related to the movement of the pupils and the eyebrows, or the blinking of the eyes. However, we are afraid that the automatic location of these facial landmarks and the extraction of the corresponding behavioural parameters could be too challenging because of the low quality of our video database. Concerning the spatial subsystem of our multimodal biometric approach, the literature on person recognition using facial appearance has clearly demonstrated that PCA is not suited for computing the most discriminative features [12][67][96]. Hence, it would be necessary to examine more performing approaches, and investigate their compatibility with our probabilistic framework for model estimation and user classification; for this purpose, one possible candidate may be the probabilistic subspace strategy proposed by Moghaddam in [58]. Finally, the integration step can be improved by exploring more sophisticated fusion techniques, like for example the use of a post-classifier [22] (Section II.H.3) to integrate the different similarity scores; however, the amount of data required for estimating its parameters can be enormous compared to the one available.

A future way to improve the tomoface recognition system introduced in Chapter VI, could be to develop an alternative space reduction strategy that demonstrates better discriminative properties than PCA; for this purpose, we suggest looking into some more linear techniques, like LDA or *canonical correlation analysis (CCA)* [85], or eventually their non-linear extensions, like *kernel principal component analysis (KPCA)* [82] or *kernel linear discriminant analysis (KLDA)* [57]. Moreover, it could be important to evaluate the effect of uncontrolled body and camera motion on the resulting video X-ray images, which are supposed to summarise only the facial motion and appearance information. We expect that it would be necessary to develop some further pre-processing to compensate for these additional disturbing movements, before extending the tomoface approach to these novel situations as well.

Finally, considering the evolution of the research on speaker verification [27][76], all biometric systems proposed in this thesis could be improved by developing a *universal background model (UBM)* approach for impostor modelling, which could replace the background or cohort model estimations in our classification steps.

VII.C. Scientific publications derived from this research

Some parts of the work presented in this doctoral dissertation resulted in the following scientific publications:

1. Matta F., Saeed U., Mallauran C. and Dugelay J.-L.
Facial gender recognition using multiple sources of visual information.
IEEE Proceedings on Multimedia Signal Processing, pag. 785-790, October 2008.

2. Matta F. and Dugelay J.-L.
Tomofaces: eigenfaces extended to video speakers.
IEEE Proceedings on Acoustics, Speech and Signal Processing, pag. 1793-1796, April 2008.

3. Matta F. and Dugelay J.-L.
Introduction de paramètres dynamiques en reconnaissance faciale (french).
Proceedings on COmpression et REprésentation des Signaux Audiovisuels, November 2007.

4. Matta F. and Dugelay J.-L.
Video face recognition: a physiological and behavioural multimodal approach.
IEEE Proceedings on on Image Processing, vol. 4, pag. 497-500, September 2007.

5. Saeed U., Matta F. and Dugelay J.-L.
Person recognition based on head and mouth dynamics.
IEEE Proceedings on Multimedia Signal Processing, pag. 29-32, October 2006.

6. Matta F. and Dugelay J.-L.
Person recognition using human head motion information.
International Conference on Articulated Motion and Deformable Objects, LNCS, vol. 4069/2006, pag. 326-335, July 2006.

7. Matta F. and Dugelay J.-L.
A behavioural approach to person recognition.
IEEE Proceedings on Multimedia and Expo, pag. 1461-1464, July 2006.

8. Matta F. and Dugelay J.-L.
Towards person recognition using head dynamics.
IEEE Proceedings on Image and Signal Processing and Analysis, pag. 306-309, September 2005.

Chapter VIII. Appendices

VIII.A. Video database of Italian TV speakers

There are two fundamental requirements for developing and assessing the performance of our recognition systems that exploit the temporal information in videos:

1. A data set with natural head and facial motion in an unconstrained scenario.
2. Enough data per user to enable the enrolment and recognition using behavioural biometric identifiers; for our techniques we estimate that they require at least 3-4 minutes of video per person.

VIII.A.1. Analysis of standard video databases

Unfortunately, to the best of our knowledge there are really few standard video databases available, and none of them is suited for our research. The main reason is that few recognition approaches exploit the temporal information in videos, and that the use of face as a hybrid (both physiological and behavioural) identifier has been largely unexplored until now, as investigated in Chapter III. In fact, the existing databases have been conceived for multi-biometric recognition systems that integrate two of more of these well known biometric identifiers: facial appearance, voice, 3D scans, signatures or fingerprints; in contrast, the head and facial motion has not been considered.

In particular, we examined the following video databases:

- The *extended M2VTS database (XM2VTSDB)* [99]. It fulfils none of the requirements: the users are reading numbers in a constrained scenario with no head motion, and there is not enough data per user (less than 20 seconds of video).
- The *VALID database* [100]. Again, it satisfies none of the requirements: 5 recordings per person (less than 30 seconds) are not enough, and the users are static readers with no motion.
- The *My IDea database* [101]. It does not fulfil the requirements as well. First of all, even if it has more data than the previous ones (less than 120 seconds per user), it is still not enough. In addition, the videos have been recorded in too heterogeneous situations: in some shots the users repeat a phrase, in some others they count, sometimes they rotate the head horizontally, sometimes vertically, and some videos have full facial occlusions. Hence, the actual data that could be used to represent the natural motion of the users is minimal.

VIII.A.2. Description of the database

Due to the lack of standard video databases suited for our research, we were obliged to create our own data set. Hence, we have been recording a compressed version of the TV news from the Italian national channel *RAI 1* [102], over a period of 21 months. Next, we manually isolated a few short video clips, in which the TV speakers introduce the coverage; we selected those sequences where the movement of the announcer is natural and no capricious events are occurring, like a scene change, a discussion with a guest or a reporter, etc. In the end, we produced a small video database consisting of 208 video clips from 13 TV speakers (8 men and 5 women) of 13 seconds each.

Figure 43 illustrates our data set by showing the first 7 frames of each speaker. It is important to notice that: there is no camera motion, because the camera is fixed, there are no zooms or changes in scale, and that the depth variation due to the in-depth movement of the speaker is insignificant, because the camera is far. Hence, all motion that can be extracted from these sequences is relative to the behaviour of the announcers.

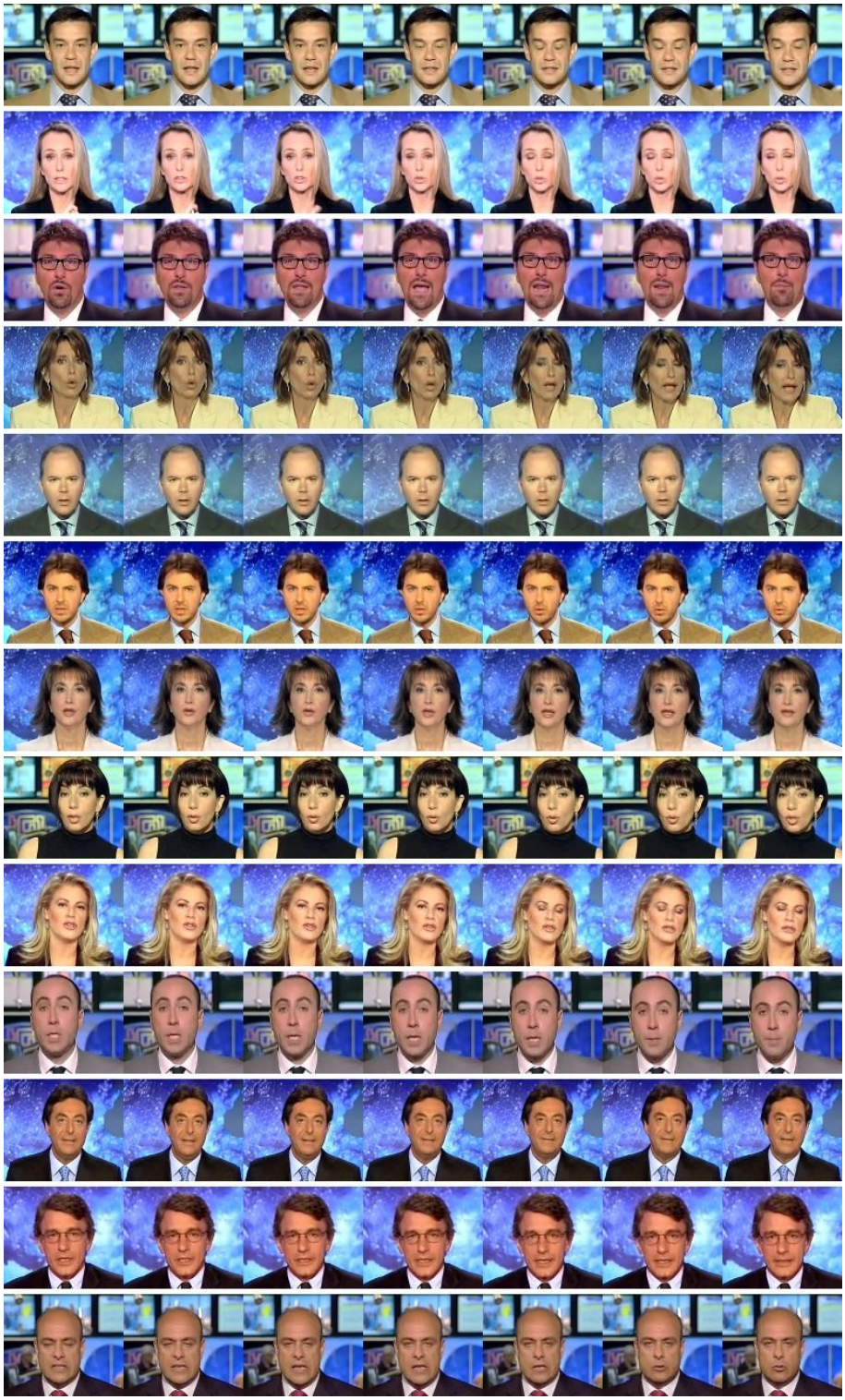

Figure 43: illustration ofour video database with the first 7 frames ofeach TV speaker.

In Figure 44 we have an example of variations present in our database; there are: assorted haircuts, clothing and backgrounds, lesser appearance variations due to aging and makeup, and various ornamentations (ear rings, necklaces).

Figure 44: example ofvariations in our video database.

In Figure 45 there is close up of a video key frame and of a predicted frame, to give an idea of the visual quality of the database; the compression artefacts are easy to remark by looking at the eyes, lips and textured background. From one side, the automatic location and segmentation of facial landmarks becomes challenging, which could actually affect their exploitation for recognition; on the other hand, this low quality database better simulates the real operational conditions of those systems that use surveillance video data.

Figure 45: close-ups of a video key frame (left) and of a predicted frame (right).

VIII.A.3. Enrolment and recognition subsets

For a *technology evaluation* of our recognition systems (Section II.E), we split the whole database into an *enrolment subset* and a *recognition subset*: 104 video sequences (8 for each of the 13 clients) are employed for the training of our systems (enrolment), and the remaining 104 (8 for each of the 13 clients) are used for their testing (recognition). We explicitly keep the two data subsets disjoint, in order to reduce the risk of systematic errors in the test procedure, due to an experimental situation over fitted to our video database (Section II.E.4).

All technical details of our video database are summarised in Table 7.

		WHOLE DATABASE	ENROLMENT SUBSET	RECOGNITION SUBSET
OVERALL	Number of individuals	13	13	13
	Number of men	8	8	8
	Number of women	5	5	5
	Number of videos	208	104	104
	Number of frames	68640	34320	34320
	Length of videos	45min. 49sec.	22 min. 54sec.	22 min. 54sec.
PER PERSON	Number of videos	16	8	8
	Number of frames	5280	2640	2640
	Length of videos	3min. 31sec.	1min. 46sec.	1min. 46sec.
PER VIDEO	Number of frames	330	330	330
	Length	13 sec.	13 sec.	13 sec.
RESOLUTION	Spatial	192 pixel rows or height	192 pixel rows or height	192 pixel rows or height
		224 pixel columns or width	224 pixel columns or width	224 pixel columns or width
	Temporal	24.97 frames/second	24.97 frames/second	24.97 frames/second
COMPRESSION	Compression rate	118 Kbits/second	118 Kbits/second	118 Kbits/second
	Format	Windows Media Video 9	Windows Media Video 9	Windows Media Video 9

Table 7: summary of the technical details of our video database.

VIII.B. Image database of Italian TV speakers

Person recognition systems based on facial appearance usually do not need the huge amount of data that a video can offer (in our case: 330 frames at 24.97 frames per second). In fact, the appearance variation in consecutive frames is minimal, and so it is the supplementary information that can be exploited from them; moreover, the additional computational burden largely surpasses its benefits. On the contrary, appearance based recognition algorithms are highly sensitive to inter-frame variations, like inconsistent facial alignments or changes in pose and head size.

For these reasons, we derived a special version of the video database of Italian TV speakers by sub sampling and manually normalising some video frames. More precisely, for the *enrolment subset* we extracted 28 frames from each sequence, at a frame rate of 2 frames per second, whereas for the *recognition subset* we retrieved only the first key frame. After that, to normalise the video frames we firstly (in-plane) rotated the heads to horizontal eye position, then we cropped the face regions, and finally we aligned the images using the locations of the pupils. Finally, all technical details of our *normalised image database* are summarised in Table 8, and an illustration of this image database can be seen in Figure 46.

		WHOLE DATABASE	ENROLMENT SUBSET	RECOGNITION SUBSET
OVERALL	Number of individuals	13	13	13
	Number of men	8	8	8
	Number of women	5	5	5
	Number of images	3016	2912	104
PER PERSON	Number of images	232	224	8
RESOLUTION	Spatial	64 pixel rows or height	64 pixel rows or height	64 pixel rows or height
		64 pixel columns or width	64 pixel columns or width	64 pixel columns or width

Table 8: summary of the technical details of our normalised image database.

Figure 46: illustration ofour normalised image database with the first 9 images ofeach TV speaker.

Bibliographical references

In this chapter all bibliographical references are listed in alphabetical order, sorted using the surnames of the first author of each document.

[1] Akutsu A. and Tonomura Y.
Video tomography: an efficient method for camerawork extraction and motion analysis.
ACM Proceedings on Multimedia, pag. 349-356, October 1994.

[2] Barber C.B., Dobkin D.P. and Huhdanpaa H.T.
The quickhill algorithm for convex hulls.
ACM Transactions on Mathematical Software, vol. 22, iss. 4, pag. 469-483, December 1996.

[3] Belhumeur P.N., Hespanha J.P. and Kriegman D.J.
Eigenfaces vs. fisherfaces: recognition using class specific linear projection.
IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, iss. 7, pag 711-720, July 1997.

[4] Bezdec J.C.
Pattern recognition with fuzzy objective function algorithms.
Plenium Press, 1981.

[5] Bilmes J.A.
A gente tutorial on EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models.
International Computer Science Institute, April 1998.

[6] Bishop C.
Neural networks for pattern recognition.
Oxford University Press, 1995.

[7] Bourel F., Chibelushi C.C. and Low A.A.
Robust facial feature tracking.
Proceedings on British Machine Vision, vol. 1, pag. 232-241, September 2000.

[8] Canny J.
A computational approach to edge detection.
IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 8, iss. 6, pag 679-698, 1986.

[9] Canzler U. and Dziurzyk T.
Extraction of non manual features for video sign language recognition.
Proceedings on Machine Vision Applications, pag. 318-321, December 2002.

[10] Cappelli R., Maio D., Maltoni D., Wayman J.L. and Jain A.K.
Performance evaluation of fingerprint verification systems.
IEEE Transactions on Pattern Analysis an Machine Intelligence, vol. 28, iss. 1, pag. 3-18, January 2006.

[11] Chang K.I, Bowyer K.W. and Flynn P.J.
An evaluation of multimodal 2D + 3D face biometrics.
IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, iss. 4, pag. 619-624, April 2005.

[12] Chellappa R., Wilson C.L. and Sirhoey S.
Human and machine recognition of faces: a survey.
Proceedings ofthe IEEE , vol. 83, iss. 5, pag. 705-741, May 1995.

[13] Chen L.-F., Liao H.-Y.M. and Lin J.-C.
Wavelet-based optical flow estimation.
IEEE Transactions on Circuits and Systems for Video Technology, vol. 12, iss. 1, pag. 1-12, January 2002.

[14] Chen L.-F., Liao H.-Y.M. and Lin J.-C.
Person identification using facial motion.
Proceedings on Image Processing, vol. 2, pag. 677-680, October 2001.

[15] Cheng H.D., Jiang X.H., Sun Y. and Wang J.
Color image segmentation: advances and prospects.
Pattern Recognition, vol. 34, iss. 12, pag. 2259-2281, December 2001.

[16] Cootes T.F., Taylor C.J., Cooper D.H. and Graham J.
Active shape models – their training and application.
Computer Vision and Image Understanding, vol. 61, iss. 1, pag. 38-59, January 1995.

[17] Cootes T.F., Edwards G.J. and Taylor C.J.
Active appearance models.
Computer Vision, vol. 2, pag 484-498, 1998.

[18] Daugman J.
Biometric decision landscapes.
Technical report TR482, University of Cambridge, Computer Laboratory, 2000.

[19] Daugman J.
How iris recognition works.
IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, iss. 1, pag. 21-30, January 2004.

[20] Delac K. and Grgic M.
A survey of biometric recognition methods.
Proceedings ofInternational Symposium on Electronics in Marine , pag. 184-193, June 2004.

[21] Doucet A., Godsill S. and Andrieu C.
On sequential Monte Carlo sampling methods for Bayesian filtering.
Statistics and Computing, vol. 10, iss. 3, pag. 197-208, July 2000.

[22] Duda R.O., Hart P.E. and Stork D.G.
Pattern classification.
John Wiley & Sons Interscience, 2nd edition, 2000.

[23] Edwards G.J., Taylor C.J. and Cootes T.F.
Interpreting face images using active appearance models.
IEEE Proceedings on Automatic Face and Gesture Recognition, pag. 300-305, April 1998.

[24] Edwards G.J., Taylor C.J. and Cootes T.F.
Learning to identify and track faces in image sequences.
IEEE Proceedings on Automatic Face and Gesture Recognition, pag. 260-265, April 1998.

[25] Edwards G.J., Taylor C.J. and Cootes T.F.
Improving identification performance by integrating evidence from sequences.
IEEE Proceedings on Computer Vision and Pattern Recognition, vol. 1, pag. 486-491, 1999.

[26] Figueiredo M.A.F. and Jain A.K.
Unsupervised learning of finite mixture models.
IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, iss. 3, pag. 381-396, March 2002.

[27] Furui S.
Digital speech processing: synthesis and recognition.
CRC Press, 2nd edition, November 2000.

[28] Gonzalez R.C. and Woods R.E.
Digital image processing.
Prentice Hall, 3rd edition, August 2007.

[29] Gross R. and Shi J.
The CMU Motion of Body (MoBo) database.
Technical report of Robotics Institute at Carnegie Mellon University, June 2001.

[30] Hjelmas E. and Low B.K.
Face detection: a survey.
Computer Vision and Image Understanding, vol. 83, iss. 3. pag. 236-274, September 2001.

[31] Hock Koh L., Raganatah S. and Venkatesh Y.V.
An integrated automatic face detection and recognition system.
Pattern Recognition, vol. 35, iss. 6, pag. 1529-1273, June 2002.

[32] Howell A.J. and Buxton H.
Learning identity with radial basis function networks.
Neurocomputing, vol. 20, iss. 1-3, pag. 15-34, August 1998.

[33] Huang S.K. and Trivedi M.M.
Streaming face recognition using multicamera video arrays.
Proceedings of Pattern Recognition, vol. 4, pag. 213-216, 2002.

[34] Hwang W.-S. and Weng J.
Hierarchical discriminant regression.
IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, iss. 11, pag. 1277-1293, November 2000.

[35] Kittler J., Hatef M., Duin R.P.W. and Matas J.
On combining classifiers.
IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, iss. 3, pag. 226-239, March 1998.

[36] Jain A.K. and Pankanti S.
Advances in fingerprint technology.
Elsevier Science, 2001.

[37] Jain A.K., Ross A. and Prabhakar S.
An introduction to biometric recognition.
IEEE Transactions on Circuits and Systems for Video, vol. 14, iss. 1, pag. 4-20, January 2004.

[38] Jain A.K., Pankanti S., Prabhkar S., Hong L. and Ross A.
Biometrics: a grand challenge.
IEEE Proceedings on Pattern Recognition, vol. 2, pag. 935-942, August 2004.

[39] Joly P. and Hae-Kwang K.
Efficient automatic analysis of camera work and microsegmentation of video using spatio-temporal images.
Signal Processing: Image Communication, vol. 8, iss. 4, pag. 295-307, May 1996.

[40] Kirby M. and Sirovich L.
Application of the Karhunen-Loeve procedure for the characterisation of human faces.
IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, iss. 1, pag. 103-108, January 1990.

[41] Knight B. and Johnston A.
The role of movement in face recognition.
Visual Cognition, vol. 4, iss. 3, pag. 265-273, September 1997.

[42] Kong G.S., Heo J., Abidi B.R., Paik J. and Abidi M.A.
Recent advances in visual and infrared face recognition – a review.
Computer Vision and Image Understanding, vol. 97, iss. 1, pag. 103-135, January 2005.

[43] Lades M., Vorbruggen J.C, Buhmann J., Lange J., Malsburg C.V., Wurtz R.P. and Konen W.
Distortion invariant object recognition in the dynamic link architecture.
IEEE Transactions on Computers, vol. 42, iss. 3, pag. 300-311, March 1993.

[44] Lee K.-C., Ho J., Yang M.-H. and Kriegman D.
Visual tracking and recognition using probabilistic appearance manifolds.
Computer Vision and Image Understanding, vol. 99, iss. 3, pag. 303-331, September 2005.

[45] Li B. and Chellappa R.
A generic approach to simultaneous tracking and verification in video.
IEEE Transactions on Image Processing, vol. 11, iss. 5, pag. 530-544, May 2002.

[46] Li Y., Gong S. and Liddell H.
Modelling faces dynamically across views and over time.
IEEE Proceedings on Computer Vision, vol. 1, pag. 554-559, July 2001.

[47] Li Y., Gong S. and Liddell H.
Recognising trajectories of facial identities using kernel discriminant analysis.
Image and Video Computing, vol. 21, iss. 3-4, pag. 1077-1086, December 2003.

[48] Liu C. and Wechsler H.
Evolution of optimal projection axes (OPA) for face recognition.
IEEE Proceedings on Automatic Face and Gesture Recognition, pag. 282-287, April 1998.

[49] Liu J.S. and Chen R.
Sequential Monte Carlo methods for dynamic systems.
Journal of the American Statistical Association , vol. 93, iss. 443, pag. 1032-1044, September 1998.

[50] Liu X. and Cheng T.
Video-based face recognition using adaptive hidden Markov models.
IEEE Proceedings on Computer Vision and Pattern Recognition , vol. 1, pag. 340-345, June 2003.

[51] Lucchese L. and Mitra S.K.
Color image segmentation: a state-of-the-art survey.
Proceedings on Image Processing, Vision and Pattern Recognition , vol. 67A, iss. 2, pag. 207-221, March 2001.

[52] Luis-Garcia R., Alberola-Lopez C., Aghzout O. and Ruiz-Alzola J.
Biometric identification systems.
Signal Processing, vol. 83, iss. 12, pag. 2539-2557, December 2003.

[53] Maeda E. and Murase H.
Multi-category classification by kernel based nonlinear subspace method.
IEEE Proceedings on Acoustics, Speech and Signal Processing, vol. 2, pag. 1025-1028, March 1999.

[54] Maio D., Maltoni D., Cappelli R., Wayman J.L. and Jain A.K.
FCV2004: third fingerprint verification competition.
Lecture notes in Computer Science, pag. 1-7, July 2004.

[55] Maltoni D., Maio D., Jain A.K. and Prabhakar S.
Handbook of fingerprint recognition.
Springer Verlag, 2003.

[56] Mansfield A.J. and Wayman J.L.
Best practices in testing and reporting performance of biometric devices.
Technical report CMSC 14/02, Centre for Mathematics and Scientific Computing, National Physical Laboratory, Middlesex, U.K.

[57] Mika S., Ratcsh G., Weston J., Scholkopf B. and Muller K.-R.
Fisher discriminant analysis with kernels.
IEEE Proceedings on Neural Networks for Signal Processing, pag. 41-48, August 1999.

[58] Moghaddam B.
Principal manifolds and probabilistic subspaces for visual recognition.
IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, iss. 6, pag. 780-788, June 2002.

[59] Monrose F. and Rubin A.D.
Keystroke dynamics as a biometric for authentication.
Future Generation Computer Systems, vol. 16, iss. 4, pag. 351-359, February 2000.

[60] Murase H. and Nayar S.K.
Visual learning and recognition of 3-d objects from appearance.
International Journal of Computer Vision, vol. 14, iss. 1, pag. 5-24, January 1995.

[61] Nishiyama M., Yamaguchi O. and Fukui K.
Face recognition with multiple constrained mutual subspace method.
Proceedings of Audio- and Video-Based Biometric Person Authentication, vol. 3546/2005, pag. 71-80, June 2005.

[62] Nixon M.S., Carter J.N., Shutler J.D. and Grant M.G.
New advances in automatic gait recognition.
Information Security Technical Report, vol. 7, iss. 4, pag. 23-35, December 2002.

[63] Nixon M.S. and Carter J.N.
Advances in automatic gait recognition.
IEEE Proceedings on Automatic Face and Gesture Recognition, pag. 139-144, May 2004.

[64] Oja E.
Subspace methods of pattern recognition.
Research Study Press, December 1983.

[65] Otsu N.
A threshold selection method from gray-level histograms.
IEEE Transactions on Systems, Man and Cybernetics, vol. 9, iss. 1, pag. 62-66, 1979.

[66] Paalanen P., Kamarainen J.-K., Ilonen J. and Kalviainen H.
Feature representation and discrimination based on Gaussian model probability densities. – Practices and algorithms.
Research report of the Lappeenranta University of Technology, num. 95, 2005.

[67] Perronnin F.
A probabilistic model of face mapping applied to person recognition.
Doctoral dissertation at Ecole Politecnique Fédérale de Lausanne (EPFL) and Eurécom Institute, 2004.

[68] Perronnin F., Junqua J.-C. and Dugelay J.-L.
Biometric person authentication: from theory to practice.
EURASIP Newsletter, vol. 16, iss. 1, March 2005.

[69] Phillips P.J., Grother P., Micheals R.J., Blackburn D.M., Tabassi E. and Bone M.
Facial recognition vendor test 2002: evaluation report.
http://www.frvt.org/FRVT2002/, March 2003.

[70] Plamondon R. and Srihari S.N.
Online and off-line handwriting recognition: a comprehensive survey.
IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, iss. 1, pag. 63-84, January 2000.

[71] Przybocki M. and Martin A.
NIST speaker recognition evaluation chronicles.
http://www.nist.gov/speech/, 2004.

[72] Pun K.H. and Moon Y.S.
Recent advances in ear biometrics.
IEEE Proceedings on Automatic Face and Gesture Recognition, pag. 164-169, May 2004.

[73] Rabiner L.R.
A tutorial on hidden Markov models and selected applications in speech recognition.
IEEE Proceedings, vol. 77, iss. 2, pag. 257-286, February 1989.

[74] Raytchev B. and Murase H.
Unsupervised recognition by associative chaining.
Pattern Recognition, vol. 36, iss. 1, pag. 245-257, January 2003.

[75] Raytchev B. and Murase H.
Unsupervised face recognition on multi-view face sequences based on pairwise clustering with attraction and repulsion.
Computer Vision and Image Understanding, vol. 91, iss. 1-2, pag. 22-52, July-August 2003.

[76] Reynolds D.A., Quatieri T.F. and Dunn R.B.
Speaker verification using adapted Gaussian mixture models.
Digital Signal Processing, vol. 10, iss. 1-3, pag. 19-41, January 2000.

[77] Rosenberg A.E., DeLong J., Lee C.-H., Juang B.-H. and Soong F.K.
The use of cohort normalized scores for speaker verification.
Proceedings of Spoken Language Processing, pag. 599-602, October 1992.

[78] Roy Chowdhury A.K. and Chellappa R.
Face reconstruction from monocular video using uncertainty analysis and a generic model.
Computer Vision and Image Understanding, vol. 91, iss. 1-2, pag. 188-219, July-August 2003.

[79] Sanchez-Reillo R., Sanchez-Avila C. and Gonzalez-Marcos A.
Biometric identification through hand geometry measurements.
IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, iss. 10, pag. 1168-1171, October 2000.

[80] Sanderson C. and Paliwal K.K.
Identity verification using speech and face information.
Digital Signal Processing, vol. 14, iss. 5, pag. 449-480, September 2004.

[81] Satoh S.
Comparative evaluation of face sequence matching for content-based video access.
IEEE Proceedings on Automatic Face and Gesture Recognition, pag. 163-168, March 2000.

[82] Scholkopf B., Smola A.J. and Muller K.-R.
Kernel principal component analysis.
Advances in kernel methods: support vector learning, pag. 327-352, 1999.

[83] Steffens J., Elagin E. and Neven H.
PersonSpotter – fast and robust system for human detection, tracking and recognition.
IEEE Proceedings on Automatic Face and Gesture Recognition, pag. 516-521, April 1998.

[84] Sturim D.E., Reynolds D.A., Singer E. and Campbell J.P.
Speaker indexing in large audio databases using anchor models.
IEEE Proceedings on Acoustics, Speech and Signal Processing, vol. 1, pag. 429-432, May 2001.

[85] Thompson B.
Canonical correlation analysis: uses and interpretation.
John Wiley and Sons, 1984.

[86] Torres L. and Vilà J.
Automatic face recognition for video indexing applications.
Pattern Recognition, vol. 35, iss. 3, pag. 615-625, March 2002.

[87] Turk M.A. and Pentland A.P.
Face recognition using eigenfaces.
IEEE Proceedings on Computer Vision and Pattern Recognition, pag. 586-591, June 1991.

[88] Uludag U., Pankanti S., Prabhakar S. and Jain A.K.
Biometric cryptosystems: issues and challenges.
Proceedings ofthe IEEE, vol. 92, iss. 6, pag. 948-960, June 2004.

[89] Van Ouwerkerk J.D.
Image super-resolution survey.
Image and Video Computing, vol. 24, iss. 10, pag. 1039-1052, October 2006.

[90] Verbeek J.J., Vlassis N. and Krose B.
Efficient greedy learning of Gaussian mixture models.
Neural Computation, vol. 15, iss. 2, pag. 469-485, February 2003.

[91] Weng J., Evans C.H. and Hwang W.-S.
An incremental learning method for face recognition under continuous video stream.
IEEE Proceedings on Automatic Face and Gesture Recognition, pag. 251-256, March 2000.

[92] Wiskott L, Fellous J.-M., Kruger N. and Malsburg C.V.
Face recognition by elastic bunch graph matching.
IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, iss. 7, pag. 775-779, July 1997.

[93] Yamaguchi O., Fukui K. and Maeda K.I.
Face recognition using temporal image sequence.
IEEE Proceedings on Automatic Face and Gesture Recognition, pag. 318-323, April 1998.

[94] Yan P. and Bowyer K.W.
Ear biometrics using 2D and 3D images.
IEEE Proceedings on Computer Vision and Pattern Recognition, iss. 3, pag. 121-128, June 2005.

[95] Zhang D.D.
Palmprint authentication.
International Series on Biometrics, Springer, vol. 3, 2004.

[96] Zhao W., Chellappa R., Phillips P.J. and Rosenfeld A.
Face recognition: a literature survey.
ACM Computing Surveys, vol. 35, iss. 4, pag. 399-458, December 2003.

[97] Zhou S., Krueger V. and Chellappa R.
Probabilistic recognition of human faces from video.
Computer Vision and Image Understanding, vol. 91, iss. 1-2, pag. 214-245, July-August 2003.

[98] Zhou S., Chellappa R. and Moghaddam B.
Visual tracking and recognition using appearance-adaptive models in particle filters.
IEEE Transactions on Image Processing, vol. 13, iss. 11, pag. 1491-1506, November 2004.

[99] The extended M2VTS database (XM2VTSDB).
http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/

[100] The VALID database.
http://ee.ucd.ie/validdb/

[101] The My IDea database.
http://diuf.unifr.ch/diva/biometrics/MyIdea/

[102] The Italian national channel RAI 1.
http://www.raiuno.rai.it/