UNIVERSIDADE ESTADUAL PAULISTA
"Júlio de Mesquita Filho"
Pós-Graduação em Ciência da Computação

Giovani Chiachia

Improving Face Recognition with Multispectral Fusion and Support Vector Machines

BAURU

2009

Giovani Chiachia

Improving Face Recognition with Multispectral Fusion and Support Vector Machines

Dissertação apresentada para obtenção do título de Mestre em Ciência da Computação, área de Processamento de Imagens e Visão Computacional, junto ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Biociências, Letras e Ciências Exatas, da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

Orientador: Prof. Dr. Aparecido Nilceu Marana

Co-Orientador: Dr. Christian Küblbeck

BAURU

2009

Giovani Chiachia

Improving Face Recognition with Multispectral Fusion and Support Vector Machines

Dissertação apresentada para obtenção do título de Mestre em Ciência da Computação, área de Processamento de Imagens e Visão Computacional, junto ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Biociências, Letras e Ciências Exatas, da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

BANCA EXAMINADORA

Prof. Dr. Aparecido Nilceu Marana
Professor Assistente Doutor
UNESP - Bauru
Orientador

Prof. Dr. Roberto Marcondes Cesar Junior
Professor Titular
USP - São Paulo

Prof. Dr. Ivan Rizzo Guilherme
Professor Assistente Doutor
UNESP - Rio Claro

Bauru, 19 de Junho de 2009

Aos meus pais, Valdir e Maria de Fátima.

Acknowledgment

More than two years were necessary to carry out this work and many people got involved in different steps of this, personally speaking, amazing experience.

First of all I would like to thank Prof. Dr. Aparecido Nilceu Marana for his belief on supervising me. His guidance, availability and enthusiasm were fundamental within this period.

Many thanks to Dr. Christian Küblbeck, Andreas Ernst, and Tobias Ruf from the Cognitive System team at Fraunhofer Institute. Their supervision during four months in Erlangen was decisive for this thesis. Along this period, I had the extraordinary chance to work with this high quality team and, without the tools and the C++ classes (with double meaning) provided, a considerable part of this work would not be reachable. I would also like to thank Dr. Patrick Flynn from the University of Notre Dame for providing the IR face database applied in the experiments.

Thanks to my colleagues Bruno, Christoph and Lukas for sharing with me the student's side of the force. The moments we spent together were great.

I do not want to forget my roommates and very best friends Bona, Saggioro, and Rani for their support and friendship along these academic years. We all grew up together in this journey.

A special gratitude to the Brazilian people for funding my studies by means of Unesp and Capes.

I would also like to thank the jury members of my thesis committee for the interesting advices that helped to improve this manuscript.

Most importantly, I am deeply grateful to my lovely family for their absolute support. Special regards to my uncle Antonio and my grandmother Aparecida. At last, my sincerely acknowledgment to my adorable Thais for her encouragement and her unconditional love.

“We shall not cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time.”
T.S. Eliot

Resumo

O reconhecimento facial é uma das principais formas de identificação humana. Apesar das pesquisas em reconhecimento facial automático terem crescido substancialmente ao longo dos últimos 35 anos, identificar pessoas a partir da face continua sendo um desafio para as áreas de Visão Computacional e Reconhecimento de Padrões. Em função dos cenários variarem desde a identificação a partir de fotografias até o reconhecimento baseado em vídeos sem nenhum tipo de controle ao serem gravados, os maiores desafios estão relacionados à independência contra diferentes tipos de iluminação, pose e expressão. O objetivo desta dissertação é propor técnicas que possam contribuir para a melhoria dos sistemas de reconhecimento facial. A primeira técnica endereça o problema da iluminação através da fusão dos espectros visível e infravermelho da face. Através desta abordagem, as taxas de reconhecimento foram melhoradas em 2.07% enquanto a taxa de erro igual (EER) foi reduzida em 45.47%. A segunda técnica trata do caso da extração e classificação de características faciais. Ela propõe um novo modelo para reconhecimento facial através do uso de características extraídas por Histogramas Census e de uma técnica de reconhecimento de padrões baseada em Máquinas de Vetores de Suporte (SVMs). Este outro grupo de experimentos nos possibilitou aumentar a precisão do reconhecimento no teste FERET fa/fb em 0.5%. Além destes resultados, algumas contribuições adicionais deste trabalho que merecem ser destacadas são a análise da dependência estatística entre classificadores de espectros diferentes e considerações sobre o comportamento de uma única C-SVC SVM para identificação de pessoas de forma eficaz.

Palavras-chave: *Reconhecimento Facial, Imagens Infravermelho, Multibiometria, Máquinas de Vetores de Suporte, Transformação Census.*

Abstract

Face recognition is one of the primary ways of human identification. Although researches on automated face recognition have broadly increased along the last 35 years, it remains a challenging task in the fields of Computer Vision and Pattern Recognition. As the scenarios varies from static and constrained photographs to uncontrolled video images, the challenging issues on automatic face recognition are usually related with variations in illumination, pose and expressions. The goal of this master thesis is to propose techniques for the improvement of face recognition systems. The first technique addresses the problem of illumination by fusing the visible and the infrared spectra of the face. With this approach the recognition rates were improved in 2.07% while the Equal Error Rate (EER) were reduced in 45.47%. The second technique addresses the issue of face features extraction and classification. It proposes a new framework for face recognition by using features extracted by Census Histograms and a pattern recognition technique based on Support Vector Machines (SVMs). This other group of experiments enabled us to increase the recognition accuracy in the FERET fa/fb test in 0.5%. Beyond these results, additional contributions of this work that deserve to be highlighted are the statistical dependency analysis between face recognition systems based on different spectra and a better comprehension about the behavior of a single C-SVC SVM to reliably predict faces identities.

Keywords: *Face Recognition, Infrared Images, Multibiometrics, Support Vector Machines, Census Transform.*

Contents

List of Figures

List of Tables

Acronyms

3-D	*Three-Dimensional*. 2, 16, 20–22, 36, 38
CH	*Census Histograms*. 3, 40, 60, 77–79
CMC	*Cumulative Match Characteristic*. 9, 10
COTS	*Commercial Off-The-Shelf*. 15
CT	*Census Transform*. 64, 65
DLA	*Dynamic Link Architecture*. 37
EBGM	*Elastic Bunch Graph Matching*. 35, 37
EER	*Equal Error Rate*. 12, 74, 76
FAR	*False Acceptance Rate*. 11, 12, 16
FAR	*False Rejection Rate*. 11, 12, 16
FERET	Facial Recognition Technology. 3, 13–15, 17, 67–69, 74, 78–82
FRGC	Face Recognition Grand Challenge. 13
FRVT	Face Recognition Vendor Tests. 10, 13–17, 69, 74
HMM	Hidden Markov Models. 36, 37
HVS	*Human Visual System*. 1
ICA	Independent Component Analysis. 36, 37, 60
ICP	*Iterative Closest Point*. 38
ICPR	International Conference on Pattern Recognition. 13

IR	*Infrared*. 2, 3, 20–24, 26, 27, 68, 69, 82
KKT	*Karush-Kuhn-Tucker*. 56, 57
kNN	*k-Nearest Neighbor*. 45
LBP	*Local Binary Pattern*. 35
LDA	*Linear Discriminant Analysis*. 34, 37
LFA	*Local Features Analysis*. 35
LWIR	*Long-Wave Infrared*. 4, 23
MWIR	*Mid-Wave Infrared*. 23
NETD	*Noise Equivalent Temperature Difference*. 26
NIR	*Near-Infrared*. 23
NN	Neural Network. 36, 37
PCA	*Principal Component Analysis*. 3, 4, 34, 36, 37, 40–44, 60, 74
RBF	*Radial Basis Function*. 59, 77–79, 81
ROC	*Receiver Operating Characteristics*. 11, 12
SOM	*Self-Organizing Map*. 37
SV	*Support Vectors*. 50, 57, 59
SVC	*Support Vector Classification*. 3, 60, 82
SVD	*Singular Value Decomposition*. 42, 43
SVM	*Support Vector Machines*. 3, 4, 35, 37, 39, 40, 45, 46, 49, 50, 56–58, 60–62, 66, 67, 71–73, 76–83
SWIR	*Short-Wave Infrared*. 23
VC	*Vapnik-Chervonenkis*. 52, 53

Chapter 1

Introduction

The earliest work on face recognition can be traced back at least to the 1950s in psychology and to the 1960s in the engineering literature. However, research on automatic machine recognition of faces really started in the 1970s [1]. Even with a significant increase on researches over the last 35 years, automatic face recognition remains a challenge for the areas of Computer Vision and Pattern Recognition [2]. Such an effort may be explained based on the fact that face recognition offers a non-intrusive, and perhaps the most natural, way of identification. Although several biometric methods based on other characteristics (e.g. fingerprints, retina, and iris patterns) can be used, they mostly rely on the cooperation of the participants. On the other hand, human recognition using faces is intuitive and can be done without the participants' cooperation or knowledge. The wide range of commercial possibilities, which raised in the early 1990s with real-time hardwares, has also played an important role on this research trend [1].

Application areas for face recognition technology are broad, including human-computer interaction, identification for law enforcement, border control, access control to secure places and financial transactions, and video surveillance. Despite many face recognition techniques have been proposed with significant promise, robust face recognition is still difficult. The major shortcomings on computer-based face recognition are variations in illumination, pose and facial expressions [1].

1.1 Challenges

Besides the visual information, the *Human Visual System* (HVS) carries out face recognition by using high-level knowledge like age, gender, race, identity and emotional state [2]. The HVS is also able to track facial motions that alter configuration of features, making it difficult to encode the face structure. Face movements include gestures while speaking, changes in the gaze or in the head pose, and expressions, such as smile or anger. These movements play an important role in human social interactions and can be employed to call the attention to an object or event of importance in the environment. Humans are capable of these impressive feats of visual information processing even after several years of separation and when viewing conditions are variable or less than optimal, for example, when the face is poorly illuminated, viewed from an unusual perspective or at a distance so the resolution of the image is limited. Therefore, face recognition in adverse situations is part of human lives and such a complexity is difficult to be mathematically modeled.

Applications of face recognition varies from static and constrained photographs to uncontrolled video images, posing a wide range of technical challenges and requiring an equally wide range of image processing and pattern recognition techniques [1].

By definition, a face is a *Three-Dimensional* (3-D) object illuminated from different light sources and surrounded by arbitrary background. Hence, the appearance of a face varies significantly when projected onto a 2-D image [2]. Different viewpoints also cause important changes in its appearance. Therefore, robust face recognition systems must embed the ability to recognize faces despite such variations. At the same time, they must be strong to typical image acquisition problems such as noise, distortion, and poor image resolution. Many methods have been proposed to handle these issues [3]. Earlier approaches treated face recognition as a typical pattern recognition problem. However, with the research advances, such methods also started to consider issues related to the representation aspects of faces [1]. The domain knowledge obtained by considering the facial components and its dynamics enables a better comprehension about the inherently uniqueness that faces represents.

1.2 Scope and Contributions

With the given scenario, the purpose of this master thesis is to address some of these deficiencies by taking as advantage methods for face recognition improvement based on three different aspects: new sensor modalities, *multibiometrics*, and *machine learning* techniques.

Regarding the new sensors issue, the thermal *Infrared* (IR) imaging is taken as a confident measurement of the heat pattern of faces. The IR images can be acquired by means of special cameras at the same time that visible recordings are taken [4]. This thesis presents some comparative evaluations of each face recognition modality (IR and visible).

In addition, the fusion of such multispectral classifiers is evaluated. A key contribution of this work is the estimation of different spectra classifiers' dependency and the relationship between this dependency and the recognition/error rates obtained through the score level fusion. The employment of IR images combined with visual images refers to techniques related to multibiometrics, an important tendency in automated human identification. In this work, normalization and fusion techniques were also addressed in the experiments.

With respect to the machine learning technique, the face recognition by *Support Vector Machines* (SVM) is assessed in this thesis. The approach firstly proposed by Phillips [5] based in a *Difference Space* is replicated. However, instead of using *Principal Component Analysis* (PCA) to extract representative face features, the trained SVMs have features based on local pixels structures as inputs. These features extracted by the *Census Transform* [6] are combined by a scanning window technique called *Census Histograms*. The fundamental contribution highlighted by Phillips [5] in the adoption of SVMs for face recognition purposes is also reproduced. This contribution is related to employing the distance that a test pattern has to the separating hyperplane as a similarity measurement, instead of using the default binary decision provided by C-SVC SVM. Additionally, the popular FERET *fa/fb* evaluation protocol was employed. The framework assessment regarding this protocol leads to a better comprehension by means of comparing SVMs to other face recognition techniques. Moreover, investigations linking data linearity, number of *Support Vectors*, and training expenditure are presented.

1.3 Organization of the Thesis

In short, this thesis is organized with the next three Chapters covering the necessary backgrounds. After, Chapters 5 and 6 present the material and the methodology used to assess the proposed improvements. Together with the methodology, Chapter 6 also presents the equivalent experimental results.

Regarding the backgrounds, to place face recognition as a *Biometric* modality, Chapter 2 covers face recognition in a broader sense by considering its characteristics and standards of evaluation. Section 2.1 is related to concepts of *Biometrics* common to different Biometric traits, including faces. In section 2.2, a summary of the main face recognition comparative evaluations is provided, while Section 2.3 presents the architecture of a typical operational scenario of face recognition systems.

Chapter 3 depicts the three already mentioned ways to overcome major challenges on face recognition. In Section 3.1, two sensing modalities are discussed. Special emphasis is given to *Long-Wave Infrared* (LWIR) images, which is part of the further experiments. Yet, in Section 3.2 multibiometrics is introduced with its peculiarities regarding modes of operation, levels of fusion, and score normalization. Closing this Chapter, Section 3.3 presents a brief review about learning approaches that have been successfully applied to recognize faces.

Aimed at deeper explaining the techniques used in the experiments, Chapter 4 covers PCA, Census Histograms, and SVMs in a detailed manner. Its content ranges from fundamental concepts to the application of such techniques to handle with face images. Section 4.1 accounts for the PCA used within the multispectral fusion. As the highlighted learning approach, Support Vector Machines is even more detailed. The Census Histograms feature extraction technique used to construct the SVM input space is presented in Section 4.2. At last, Section 4.3 deals with the SVM theory and its application to face images.

In this thesis, all investigations were done considering the face databases covered in Chapter 5. As already stated, the methodology and results of the experiments are presented in Chapter 6.

Finally, Chapter 7 discusses and presents some conclusions about the results obtained and provides a few remaining considerations about the topics covered by the thesis.

Chapter 2

Face Recognition

A general statement regarding the problem of machine recognition of faces can be formulated as follows: given still or video images of a scene, identify or verify one or more persons in the scene using a stored database of faces [1].

To address this given problem, a general procedure can be established by face detection, feature extraction, and recognition. Face detection is used to separate face-like objects from cluttered scenes. To be recognized, faces are usually represented in terms of vectors in a lower dimensional feature space extracted from the images [2]. Once the face features are available, the recognition step is based on the establishment of a similarity between feature vectors. Such a relation between two vectors of a known and an unknown identity is normally carried out by means of a *similarity score*.

In this chapter the main topics of Biometrics are presented in Section 2.1, giving a clear overview about how face recognition systems are situated on this field. Further, the protocols for evaluation are discussed in Section 2.2. These standards are quite important to analyze different face recognition systems, since there are many details to be considered when comparing performance ratings. As briefly mentioned above, such a recognition system has steps to be considered which are sharply distinct. Thus, the outline architecture of such systems is covered in Sections 2.3.

2.1 Within Biometrics

Humans use body characteristics such as face, voice, gait, etc, to recognize each other. In the mid 19^{th} century, the chief of the criminal identification division of the police department in Paris, Alphonse Bertillon, was the first to propose the idea of using some body measurements to identify criminals. Just as his idea was gaining popularity, it was obscured by a far more significant and practical discovery of the distinctiveness of the human fingerprints in the late of the same century [7].

In its beginning, biometrics efforts was focused on identifying criminals, but nowadays it is also being increasingly used to establish person recognition in a large number of civilian applications. When investigating biometrics, the main question that arises is what biological measurements qualify to be a biometric trait. In general, any human physiological and/or behavioral characteristic can be used as a biometric characteristic as long as it satisfies the following requirements [7]:

Universality: each person should have the characteristic;

Distinctiveness: any two persons should be sufficiently different in terms of the characteristic;

Permanence: the characteristic should be sufficiently invariant (with respect to the matching criterion) over a period of time;

Collectability: the characteristic can be measured quantitatively.

Nevertheless, in a practical biometric system there are some other issues that should be considered [7], including:

Performance: refers to the achievable recognition accuracy and speed, the resources required to achieve the desired recognition accuracy and speed, as well as the operational and environmental factors that affect the accuracy and speed;

Acceptability: indicates the extent to which people are willing to accept the use of a particular biometric identifier (characteristic) in their daily lives;

Circumvention: reflects how easily the system can be fooled using fraudulent methods.

On Table 2.1 a brief comparison carried out by Jain et al. [7] of 15 biometric modalities based on the seven factors mentioned above is provided. The applicability of a specific biometric technique depends heavily on the requirements of the environment. No single technique can outperform the others in all operational environments. In this way, each biometric technique is acceptable and there is no optimal biometric characteristic [7].

Table 2.1: Comparison of different biometric modalities provided by Jain et al. [7]. High, Medium, and Low are denoted by H, M, and L, respectively.

Biometric modalities	Universality	Distinctiveness	Permanence	Collectability	Performance	Acceptability	Circumvention
DNA	H	H	H	L	H	L	L
Ear	M	M	H	M	M	H	M
Face	H	L	M	H	L	H	H
Facial thermogram	H	H	L	H	M	H	L
Fingerprint	M	H	H	M	H	M	M
Gait	M	L	L	H	L	H	M
Hand geometry	M	M	M	H	M	M	M
Hand vein	M	M	M	M	M	M	L
Iris	H	H	H	M	H	L	L
Keystroke	L	L	L	M	L	M	M
Odor	H	H	H	L	L	M	L
Palmprint	M	H	H	M	H	M	M
Retina	H	H	M	L	H	L	L
Signature	L	L	L	H	L	H	H
Voice	M	L	L	M	L	H	H

Still on Table 2.1, one can observe that both face recognition modalities have the strongness of the high collectabilitty and acceptabillity and the weakness of permanence and performance. Once again, this is mainly due to changes in illumination, pose, facial expressions, aging, and physiological state (in thermograms). In a roundabout manner, these are the problems that this thesis addresses.

One peculiarity that must be noticed is that, beside the face itself, additional information extracted from it, such as race, age, gender, and facial expression can be used to enhance the recognition accuracy [2]. Recently, significant research efforts have been focused on video-based face modeling/tracking, recognition, and system integration. New datasets have been created and evaluations of recognition techniques using these databases have been carried out [8, 9]. Therefore, face recognition has been considered one of the most active applications of pattern recognition and image understanding [1].

2.1.1 Modes of Operation

Biometric systems can work both on verification and identification modes. In the verification mode, the system validates the identity of a person by comparing the captured biometric data with her own biometric template(s) stored in the system's database. In such a system, an individual who desires to be recognized claims an identity and the system conducts a one-to-one comparison to determine whether the person really is who he/she claim to be. Identity verification is typically used for *positive recognition*, where the purpose is to prevent multiple people from using the same identity [7].

In the identification mode, the system recognizes an individual by searching the templates of all the users in the database for a match. Hence, a one-to-many comparison is conducted by the system trying to determine an individual's identity without the subject having to claim for it. The ability to identify an individual is the key for *negative recognition* applications, where the system establishes whether the person is who he/she denies to be. The purpose of negative recognition is to prevent a single person from using multiple identities [7]. While traditional methods of personal recognition may work for positive recognition, negative recognition can only be established through biometrics.

The identification task can be further split in two categories. The *closed-set* identification is the classic performance measure used in the automatic face recognition community. With closed-set identification, the basic question asked is: Whose face is this? In such a closed-set scenario, this question is meaningful because a biometric sample to be identified is always that of someone in the gallery. The general case of closed-set identification is the *open-set* identification. With open-set identification, the person in the probe does not have to be somebody in the gallery and here the basic question asked is: Do we know this face? Thus, in this case, a system has first to decide if the probe contains an image of a person in the gallery. As can be noted, open-set identification is the general case task, with verification and closed-set identification being its special cases [3].

2.1.2 Evaluation Standards

The response of a biometric matching system is the matching score, $S(X_Q, X_I)$ (typically a single number), that quantifies the similarity between the input and the database template representations (X_Q and X_I, respectively). The higher the score, the more confident is the system that the two biometric measurements come from the same person.

Computing performance of a biometric system requires three sets of images. The first is the *gallery* G, which contains biometric samples of the people known to a system. The other two are the *probe sets*. A *probe* is a biometric sample that is presented to the system for recognition. The first probe set is P_G, which contains different biometric samples of people present in the gallery. The second probe set is P_N, which contains biometric samples of people who are not in the gallery [3].

Let us assume $\{g_1, ..., g_{|G|}\}$ as the gallery set G consisting of one or more biometric samples per person, and $g*$ being the unique match from all comparisons between a probe p_j and G. In this case, p_j has rank n if $S(p_j, g*)$ is the nth largest similarity score. This is denoted by $rank(p_j) = n$.

In the closed-set identification evaluation, where P_N is empty, the first step consists in sorting the similarity scores between a probe p_j and a gallery set G, and computing the $rank(p_j)$. The identification rate for rank n, $P_I(n)$, is the fraction of probes at rank n or lower. For rank n, let

$$C(n) = |\{p_j : \text{rank}(p_j) \leq n\}| \tag{2.1}$$

be the cumulative count of the number of probes of rank n or less. The identification rate at rank n is

$$P_I(n) = \frac{|C(n)|}{|P_G|} \tag{2.2}$$

where $|P_G|$ is the number of samples in the probe set.

The functions $C(n)$ and $P_I(n)$ are nondecreasing in n. The identification rate at rank 1, $P_I(1)$, is also called the correct identification rate, the top match rate, or just $TOP1$ [3], whereas the notion of *full retrieval* can be interpreted as n when $P_I(n) = 1$, in other words, the minimum $C(n)$ which comprises all probes.

Closed-set identification performance is normally reported on a *Cumulative Match Characteristic* (CMC) [3]. A CMC plots $P_I(n)$ as a function of rank n. Figure 2.1 shows an example of CMC, where the horizontal axis is the rank on a logarithmic scale. Closed-set identification performance is most often summarized with rank 1 performance, the other points such as rank 5, 10, or 20 are also commonly used. The strength and weakness of the CMC is its dependence on the gallery size $|G|$.

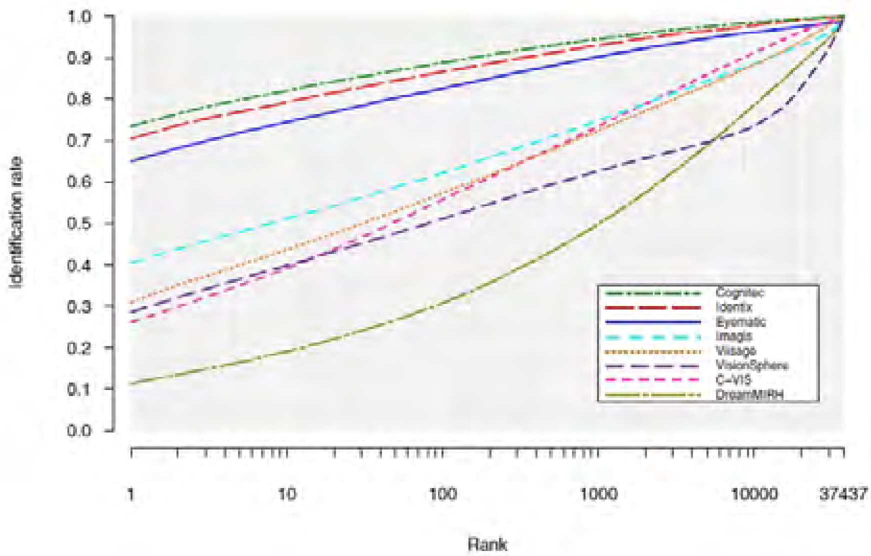


Figure 2.1: FRVT 2002 (Section 2.2) identification performance reported on a CMC [3].

2.1.3 Error Rates

Two samples of the same biometric characteristic from the same person (e.g., two face images) are not exactly the same due to imaging conditions, changes in the user's physiological or behavioral characteristics and environment conditions (e.g., temperature). Hence, there is often an overlap between the distribution of scores generated from pairs of samples from the same person and the distribution of scores from pairs of different persons. These distributions are called, respectively, *genuine distribution* and *impostor distribution* as showed in Figure 2.2 (a). To deal with this overlapping, the system decision has to be regulated by a threshold t: pairs of biometric samples generating scores higher than or equal to t are classified as *mate pairs* (i.e., belonging to the same person); pairs of biometric samples generating scores lower than t are classified as non-mate pairs (i.e., belonging to different persons).

A biometric system can make two types of errors: (i) the *false match*, when biometric measurements from two different persons are mistaken as belonging to the same person, and (ii) the *false non-match*, when two biometric measurements from the same person are mistaken as belonging to different persons. These two types of errors are often termed as *false acceptance* and *false rejection*, respectively [7].

There is a trade-off between *False Acceptance Rate* (FAR) and *False Rejection Rate* (FAR) in every biometric system. In fact, both FAR and FAR are functions of the system threshold t; if t is decreased to make the system more tolerant to input variations and noise, then FAR increases. On the other hand, if t is raised to make the system more secure, then FAR increases accordingly. The system performance at all the operating points t can be depicted in the form of a *Receiver Operating Characteristics* (ROC) curve [7], (see Figure 2.2 (b).

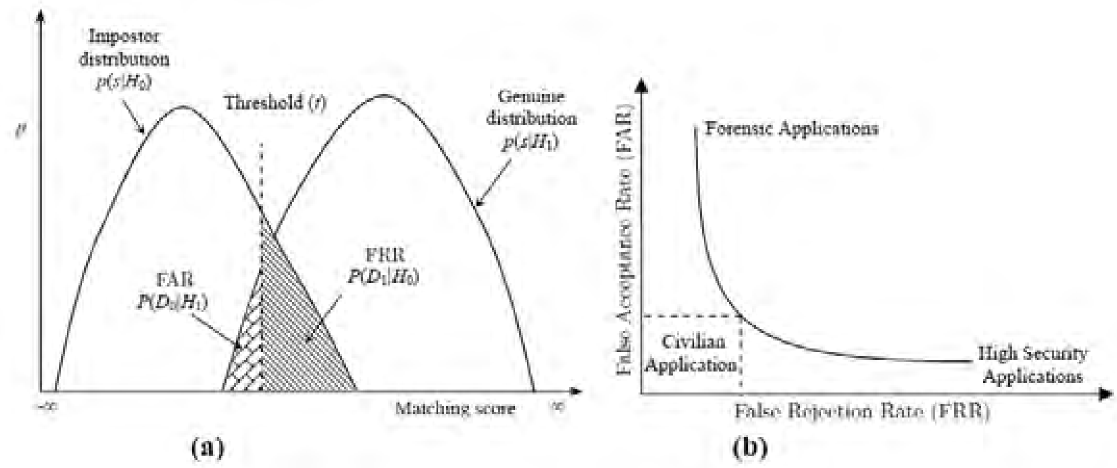


Figure 2.2: Biometric system error rates. (a) FAR and FAR for a given threshold t are displayed over the genuine and impostor score distributions; FAR is the percentage of non-mate pairs whose matching scores are greater than or equal to t, and FAR is the percentage of mate pairs whose matching scores are less than t. (b) Choosing different threshold values results in different FAR and FAR. The curve relating FAR to FAR at different thresholds is referred to as ROC. Typical operating points of different biometric applications are displayed on an ROC curve.

Let us assume g_i as the stored biometric template of the person i and p_j as the input template to be recognized. Let us also consider H_0 as the hypothesis that p_j does not come from the same person as the template g_i and H_1 as the hypothesis that p_j represents the same person as in g_i. Thus, the mathematical definitions of FAR and FAR are [7]:

$$FAR = P(D_1|H_0); \tag{2.3}$$

$$FRR = P(D_0|H_1) \tag{2.4}$$

where D_1 is the decision that the person is who he/she claims to be and D_0 the decision that the person is not who she claims to be. Both D_1 or D_0 are determined by comparing $S(g_i, p_j)$ with the threshold t. If less than t, then decide D_0, else decide D_1.

Finally, a one-number representation of the FAR and FAR can be obtained by the *Equal Error Rate* (EER), which is the negation of the system errors provided they are equalized by the threshold t. That is,

$$EER = 1 - FAR_t \quad \text{with} \quad \{t|FAR_t = FRR_t\} \tag{2.5}$$

2.2 Protocols

Automatic face recognition has a long history of independent evaluations with the three Facial Recognition Technology (FERET) evaluations (1994, 1995, and 1996), the International Conference on Pattern Recognition (ICPR) Face Recognition Contest 2000, the three Face Recognition Vendor Tests (FRVT) evaluations (2000, 2002, and 2006) and the Face Recognition Grand Challenge (FRGC) (2006) [1, 3, 10]. These evaluations have provided a basis for measuring progress in face recognition, determining the most promising approaches and identifying future research directions, with the FERET and FRVT being the "de facto" standards.

The general purpose to design and to conduct evaluations is to assess the state of the art, to advance the state of the art, and to point future research directions. The key of such evaluations is the experimental protocol. The protocol defines a set of data, how it should be used by a system to perform experiments and how the performance should be computed (format of the results). Its design might cope with the following precepts [3]:

1. Evaluations are designed and administered by groups that are independent from the algorithm developers and vendors being tested.

2. Test data are separated and not seen by the participants prior to an evaluation.

3. The evaluation test design, protocol, and methodology are published.

4. Performance results are spread in a manner that allows for meaningful differences among the participants.

Precepts 1 and 2 ensure fairness of an evaluation. Precept 1 provides assurance that the test is not designed to favor one participant over another. Independent evaluations enforces precepts 2 and 4. In addition, precept 2 ensures that systems are evaluated for their ability to generalize performance to new data sets, not the ability of the system to be tuned to a particular set of biometric samples. Publishing the evaluation protocol, as recommended in precept 3, lets the readers of published results understand how the results were computed.

The former approaches for face recognition evaluations were based on *similarity matrix*. In this approach, the input to an algorithm or system being evaluated is two sets of biometrics samples: a target set T and a query set Q. Galleries and probe sets were constructed from the target and query sets, respectively. The output from an algorithm was the similarity measure $S(t_i, q_j)$ between all pairs of images t_i from the target set and q_j from the query set, the referred similarity matrix. Performance statistics were then computed from this matrix. However, the most recent FRVT protocol (2006) was structured substantially different from its previous editions. On FRVT 2006, the format for submissions was binary executables that could be run independently on the test server. All submitted executables had to run using a specified set of command line arguments, including an experiment parameter file, files that contained the sets of biometric samples to be matched, and name of the output similarity file.

The main attributes and achievements of the mentioned FRVT and FERET evaluations were [1, 3, 10]:

FERET 1994: Until 1994, there did not exist a common evaluation protocol that included a large data set and a standard evaluation method. This evaluation established a baseline for face recognition algorithms and was designed to measure the performance of algorithms that could automatically locate, normalize, and identify faces.

FERET 1995: It consisted of a single test that measured the identification performance from a gallery of 817 individuals and included 463 duplicates[1] in the probe set (there were only 60 duplicates in the previous evaluation).

FERET 1996: The FERET protocol available nowadays comes from this edition. A number of lessons were learned from the FERET evaluations. The first was that performance depends on the probe category. A clear difference between the best and the average algorithms performance could also be noted. Another lesson was that the operational scenario considerably impacts the performance.

[1] A duplicate is a probe for which the corresponding gallery image was taken on a different day.

In 2000 the entire FERET database (Section 5) was released along with the FERET 1996 evaluation protocols, evaluation scoring code, and baseline algorithm, which provided a good benchmark for performance of new algorithms and had a significant impact on progress in the development of face recognition algorithms. In general, the FERET results revealed three major problem areas: recognizing people over time, under illumination variations, and under pose variations.

The FERET 1996 evaluation measured performance on prototype laboratory systems. After March 1997 there was rapid advances in the development of commercial face recognition systems. These advances represented both a maturing of face recognition technology and the development of the supporting system and infrastructure necessary to create *Commercial Off-The-Shelf* (COTS) systems.

By the beginning of 2000, COTS face recognition systems were readily available. To assess them, the FRVT 2000 was organized:

FRVT 2000: The protocol was the same used on FERET 1996, but the evaluation was significantly more demanding. Participation in FRVT 2000 was restricted to COTS systems. A greater variety of imagery was used in FRVT 2000 than in the FERET evaluations. FRVT 2000 showed that two common concerns, the effects of compression and recording media, do not affect performance. It also showed that future areas of interest continue to be duplicates, pose variations, and illumination variations generated when comparing indoor images with outdoor images.

FRVT 2002: It was a large-scale evaluation of automatic face recognition technology. The objective was to provide performance measures for assessing the ability of automatic face recognition systems to meet real-world requirements. It was also the first edition which made use of videos sequences. FRVT 2002 results showed that normal changes in indoor lighting do not significantly affect performance of the top systems. However, face recognition from outdoor imagery remained a challenging area of research. It also established that the use of morphable models can significantly improve nonfrontal face recognition. Another assertion was that identification performance decreases linearly in the logarithm of the size of the gallery and, in face recognition applications, accommodations should be made for demographic information, as characteristics such as age and sex can significantly affect performance.

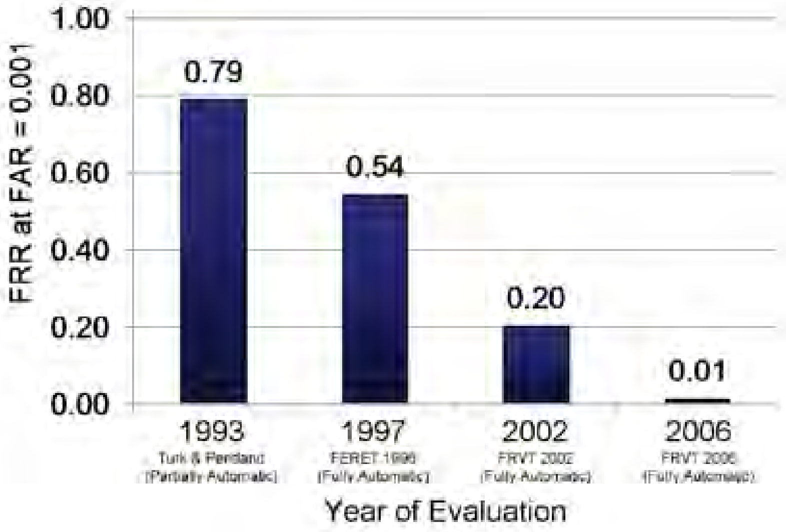


Figure 2.3: The reductions in error rate for state-of-the-art face recognition algorithms as documented through the FERET, the FRVT 2002, and FRVT 2006 evaluations. More details about the algorithms as well as about the protocols are available in [11].

FRVT 2006: It was the first technology evaluation that employed very-high resolution and 3-D range datasets. The human ability to recognize faces was also one of the targets. They concluded that the ability of algorithms to recognize faces across illumination changes has made significant progress. Another statement was that the human visual system contains a very robust face recognition capability that is excellent at recognizing familiar faces, but for unfamiliar faces this capability falls dramatically. It was found that algorithms are capable of human performance levels, and that at a FAR ranging 0.05, machines can out-perform humans.

When research in automatic face recognition began, the primary goal was to develop recognition algorithms. With progress in face recognition, the goal of developing algorithms was extended to the goal of understanding the properties of such algorithms by means of computational experiments. Such a comprehension together with new algorithms proposals have enabled a great performance improvement on recent face recognition technologies.

To close this Section, Figure 2.3 illustrates the mentioned performance achievements through the years. For each milestone, the FAR at a FAR of 0.001 (1 in 1000) is given for the most state-of-the-art representative algorithms.

2.3 System Architecture

Over the last two decades, research has focused on how to make face recognition systems fully automatic by dealing with problems such as localization of a face in a given image or video sequence and extraction of features such as eyes, mouth, etc. Meanwhile, significant advances have been made in the design of classifiers [1].

According to Li and Jain [3], a face recognition system generally consists of four modules as depicted in Figure 2.4: detection, alignment, feature extraction, and matching, where localization and normalization (face detection and alignment) are processing steps before face recognition (facial feature extraction and matching) is performed.

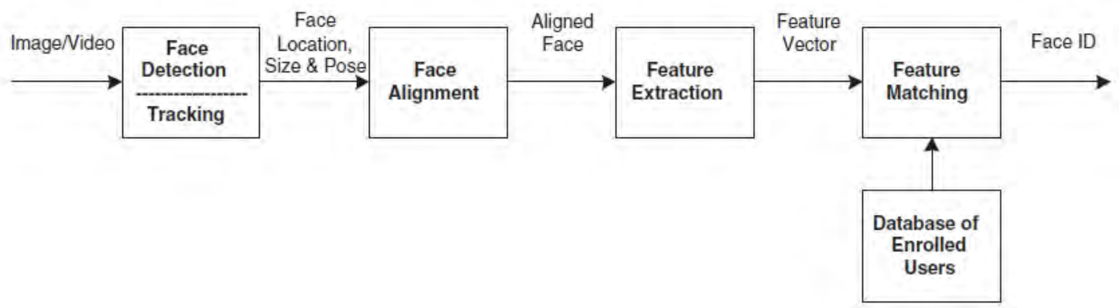


Figure 2.4: A typical face recognition system architecture [3].

Face detection segments the face areas from the background by coarse estimating its location and scale in a given scene. In the case of video, the detected faces may also need to be tracked using a face tracking component [2].

The purpose of the *face alignment* is to refine the location and to normalize the faces provided by the face detection. Facial components, such as eyes, nose, mouth and facial outline, are prior located. Based on their positions, the input face image is normalized with respect to geometrical properties, such as size and pose, using geometrical transforms or morphing. The face is usually further normalized with respect to photometrical properties such as illumination and gray scale [3].

Subsequent to the alignment step, the *feature extraction* is performed on this stable representation to provide effective information that is useful for distinguishing among faces of different persons [3].

Afterwards, on the *face matching* stage, the extracted feature vector of the input face is matched (similarity measurement) against those of enrolled faces in the database to determine the identity of the face when a match has sufficient confidence or to indicate that the face is unknown otherwise [3].

Face localization and normalization are the basis for extracting effective features to represent the face pattern. They play a crucial role in how efficient the classification methods will distinguish between faces.

Some other attributes that may classify face recognition systems consider its operational scenario. Within the possibilities significant differences exist. They are in terms of static images or video-based, image quality, amount of background clutter, variability of the images of a particular individual that must be recognized, matching criterion, and the nature, type, and amount of input from a user [1].

Chapter 3

Overcoming Challenges

Although automatic face recognition techniques have been successful in many practical applications, recognition based only on the visual spectrum has difficulties when applied in uncontrolled operating environments. Performance of visual face recognition is quite sensitive to changes in illumination. Moreover, variations between images of the same face due to changes in illumination and viewing directions are typically larger than image variations raised from changes in face identity. Other factors such as facial expressions and pose variations further complicate the face recognition task [2]. Visual face recognition techniques have difficulty in identifying individuals wearing makeup or disguises such as a fake nose or beard, which substantially change a person's visual appearance. In the same way, visual identification of identical twins or faces in which the appearance has been altered through plastic surgery is almost impossible.

To overcome such a challenges, three highlighted areas of research in face recognition are: (I) the application of new sensor modalities for face imaging, (II) the investigation of new learning techniques, and (III) the fusion of different kinds of machine knowledge to produce better results than its individual components, extracting and combining the powerfulness of each one. The next three sections detail some of these issues.

3.1 Sensor Modalities

Recognition of faces using different imaging modalities, such as IR and 3-D imaging sensors has become an area of growing interest [12, 13, 14].

Due to the focus of this work in the further implemented techniques and its experimental results on IR imagery, 3-D face range scanning is be briefly described in next section regarding its great potential on the improvement of unconstrained face recognition. Afterwards, in Section 3.1.2, the IR imagery issue is discussed in details, regarding its advantages, shortcomings, and why apply it for face recognition purposes.

3.1.1 Three-Dimensional

Because the human face is a 3-D object whose 2-D projection is sensitive to the already mentioned changes in pose and illumination, utilizing 3-D facial information is promising in the improvement of face recognition performance [15].

Over the past few years, a number of 3-D scanner and measurement techniques have been developed, with laser scanners being the first commercial product for 3-D face imaging. Thus, many researches were conducted based on this technology [13, 15]. These scanners use structured laser light to record the face image (e. g. Minolta VIVID 910 [1]), where each point in a scan has a texture color (r, b, g) as well as a location in 3-D space (x, y, z). Thus, the range images captured by the sensor contain both surface shape and texture information (Figure 3.1(a)). The 3-D shape of facial surface represents the facial structure, which is related to the internal anatomical structure instead of external appearance and environment. Compared with 2-D images, it is also more difficult to fake a 3-D image to fool the face recognition system. In real world scenarios, similar to the current 2-D camera capture systems, 3-D sensors provide only partial views of the face. However, during the training stage, 3-D face model can be constructed by taking several scans from different viewpoints (Figure 3.1(b)).

Although range scanners remain expensive, as it becomes popular, it is expected that its financial cost falls. Additionally, the probed performance and the potential of this sensing approach increasingly stimulates research efforts in this area.

On [16], several approaches for 3-D capture are listed. Another two common sensing techniques for 3-D face recognition that might be mentioned in time are: (i) triangulation for depth estimation based on structured light, where a single image of the scene illuminated by a known pattern allows 3-D and texture capture and (ii),

[1] Manufacturer names in this thesis are given only to specify details, and not to imply any endorsement of a particular equipment.

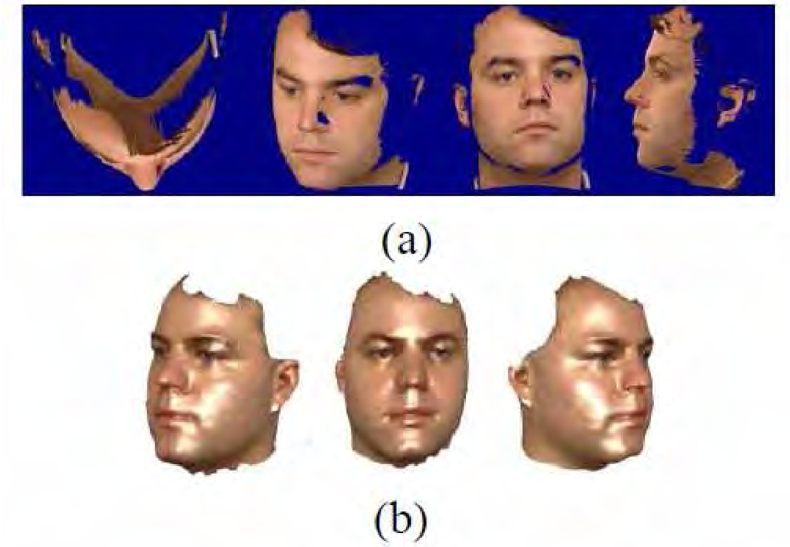


Figure 3.1: Partial scan and 3-D face model. (a) Partial face from a single capture; (b) Full constructed 3-D model [15].

triangulation applied by a stereo pair of images. However, the last one is much more an interesting measurement algorithmic technique rather than an alternative sensing modality.

A drawback of 3-D sensing technology is that all the elements of the system need to be well calibrated and synchronized to acquire accurate 3-D data. Moreover, most of the systems require the cooperation or collaboration of the subject, which restricts them to controlled scenarios. However, as the 3-D imaging technology is progressing quickly, non-intrusive 3-D data capture along with texture information will become readily available.

3.1.2 Infrared

As electromagnetic waves with length below the visible spectrum are harmful to the human body (e.g. X-rays and ultraviolet radiation), they cannot be employed for face recognition applications [2]. On the other side, thermal IR imagery has been suggested as an alternative source of information for detection and recognition of faces [2].

While visual cameras measures the electromagnetic energy in the visible spectrum range ($0.4 - 0.7\mu m$), sensors in IR cameras respond to thermal radiation in the infrared spectrum range at $0.7 - 14.0\mu m$. The infrared spectrum comprises the *reflected IR* and the *thermal IR* wavebands. The reflected IR band ($0.7 - 2.4\mu m$)

is associated with reflected solar radiation that contains no information about the thermal properties of materials. The *Near-Infrared* (NIR) ($0.7 - 0.9\mu m$) and the *Short-Wave Infrared* (SWIR) ($0.9 - 2.4\mu m$) spectra are reflective and differences in appearance between the visible and the reflective IR are due to the reflective properties of the materials. This radiation is for the most part invisible to the human eye [2].

The thermal IR band is associated with thermal radiation emitted by the objects. The amount of emitted radiation depends upon both the temperature and the emissivity of the material. The thermal IR spectrum is divided into two primary bands: the *Mid-Wave Infrared* (MWIR) of the spectral range $3.0 - 5.0\mu m$ and the already mentioned long-wave infrared LWIR, which ranges from $8.0 - 14.0\mu m$ [2].

Strong atmospheric absorption exists at $2.4 - 3.0\mu m$ between the SWIR and MWIR range and at $5.0 - 8.0\mu m$ between the MWIR and LWIR band, where imaging becomes extremely difficult. The human face and body emit thermal radiation in both bands of the thermal IR spectrum and thermal IR cameras can sense temperature variations in the face at a distance to produce thermograms in the form of 2-D images. Normally, LWIR is preferred for face recognition in the thermal IR due to much higher emissions in this band than in the MWIR [2].

One advantage of using thermal IR imaging over visible spectrum sensors arises from the fact that the light in the thermal IR range is emitted rather than reflected. Thermal emissions from skin are an inherent property, independent of illumination. Therefore, the face images captured using thermal IR sensors will be nearly invariant to changes in lighting. Such an energy can be measured in any light condition and is less subject to scattering and absorption by smoke or dust than visible light. The within-class variability is also significantly lower than that observed in visible imagery.

Therefore, the infrared spectrum has been found to have some advantages over the visible spectrum for face detection, detection of disguised faces, and face recognition under poor lighting conditions [2].

Facial expression recognition based on thermal imaging has also been investigated. An interesting and novel framework for face recognition is based on physiological information. The motivation behind this effort is to capitalize on the permanency of innate characteristics that are under the skin, using thermal IR band not only as a way to see in the dark or reduce the deleterious effect of light variability [17].

Thermal IR has also been used in some other applications such as the study of skin temperature distribution for breast cancer detection, heat source recognition for electronic parts inspection, and target detection in military applications. The detection of suspects engaged in illegal and potentially harmful activities using thermal images is another possibility, where symptoms of alertness and anxiety causes abrupt changes in local skin temperature [2].

Compared to visual face recognition, efforts in face recognition using infrared sensors have been relatively limited for a variety of reasons. Among those is the availability and low cost of ordinary cameras and the undeniable fact that visual face recognition is one of the primary activities of the human visual system. A comparative performance study of multiple face recognition methodologies using visual and thermal IR imagery was conducted in [18].

As already mentioned, visual cameras represent the reflectance information of a face object, while thermal IR sensors measure the subsurface anatomical information. Figure 3.2 shows visual and thermal image characteristics of faces with abrupt variation in illumination. A face image taken under very little ambient light is unrecognizable in Figure 3.2(c), while its counterpart thermal IR image in Figure 3.2(d) taken in the same environment reveals robust characteristics of the face regardless the illumination. In thermal IR images, background clutter is not registered and the task of face detection and segmentation are relatively easier and more reliable than in visual images [2].

(a) (b) (c) (d)

Figure 3.2: Comparison of visual and thermal IR images under abrupt changes in illumination. Visual (a) and thermal IR (b) images from the same person and their counterparts (c) and (d) taken under very low lighting. Extracted from [2].

The human face and body temperature is quite uniform, varying from 35.5 to 37.5°C and providing a consistent thermal pattern. Skin temperature in a 21°C ambient room has also a small range of variance between 26 and 28°C. The thermal signature of the faces are derived primarily from the pattern of superficial blood vessels under the skin, whose diameter is on the order of 0.1”. The vessels transport warm blood throughout the body and heat the skin directly above them on average 0.1°C more than the adjacent skin.

The complexity and vastness of about 5 kilometers of blood vessels in the head and face (Figure 3.3) assures that each person’s vascular arrangement is irreproducible and hence unique. It is also known that even identical twins have different thermal patterns [17].

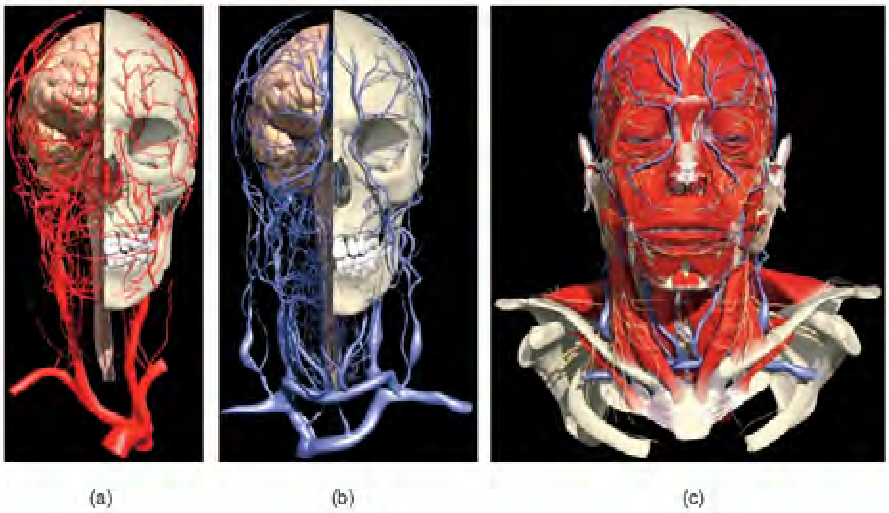


Figure 3.3: Generic map of superficial blood vessels on the face [17]. (a) Overview of an arterial network. (b) Overview of a venous network. (c) Arteries and veins together underneath the surface of the facial skin.

Roughly, the eyeglass region (if the person is wearing glasses) is usually associated with the lowest temperature of the face, followed by the nose. The eyes are relatively warmer areas, while the eyebrows, the mouth, and the cheeks often have similar temperature ranges [19]. Figure 3.4 shows the histograms of some of these facial components.

Fortunately, all anatomical and physiological characteristics presented are accomplished by the current thermal IR imaging technology. The sensitivity in such cameras are measured by the *Noise Equivalent Temperature Difference* (NETD) and cameras with NETDs about 0.01°C are broadly available (e. g. Flir Phoenix). The passive nature of such systems decreases their complexity and increases their reliability [2]. However, their prices are still an obstacle for commercial applications. Different technologies can be employed to sense temperature regarding its waveband and chemical principle. Jarvis et al. [4] provide a detailed report about them.

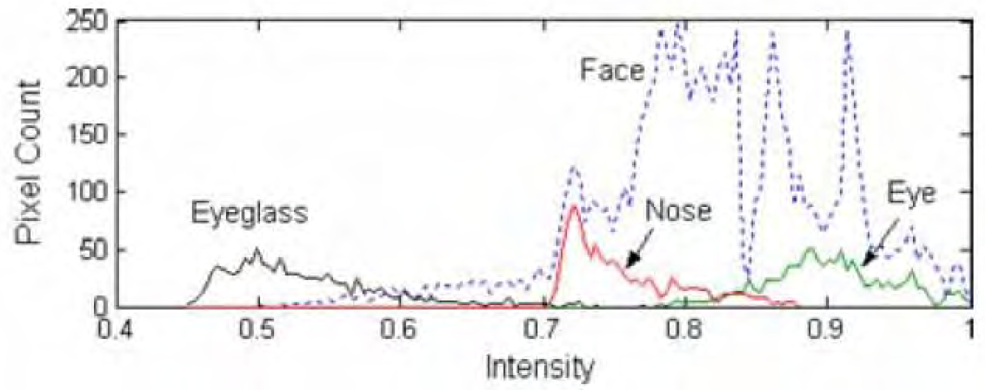


Figure 3.4: Histograms of facial components in terms of relative intensity. The sizes of components used in the histogram are 35x25 pixels for the eye and glasses, 24x35 pixels for the nose, and 150x120 for the face in images of size 240x320 [19].

Despite the benefits, thermal imaging has limitations in situations such as recognition of a person wearing glasses or seated in a moving vehicle. Glass blocks a large portion of thermal energy resulting in a loss of information near the eyes. Variations in ambient or body temperature also significantly change the thermal characteristics of the face, while the visual image features do not show a significant difference. To accomplish with such deficiencies, studies has been conducted on the fusion of both modalities [12, 13, 14], and the way we address thermal IR images in this thesis is also through multispesctral fusion.

In Section 5, a more detailed explanation is given about the collection of thermal IR images employed on this thesis experiments.

3.2 Multibiometrics

The success of face recognition systems on some civilian applications does not imply that the problem is already solved. A brief check in Section 2.2 allows the reader to conclude that there exists a plenty of scope for improvement in this biometric modality. Researchers are not only looking for issues related to reducing error rates, but they are also addressing ways to enhance the usability of such systems. A face recognition system that operate using characteristics from a single image may have the following limitations [7]:

Noise in sensed image: the sensed data might be noisy or distorted. A person wearing glasses or the presence of beard are examples of noisy data. Noisy data could also be the result of defective or improperly maintained sensors (e.g., dirty camera's lens) or unfavorable ambient conditions (e.g., poor illumination of a user's face). Noisy data may be incorrectly matched with templates in the database resulting in a user being incorrectly rejected;

Intra-class variations: the face image acquired from an individual during authentication may be very different from the one that was used to generate the template during enrollment, thereby affecting the matching process (e. g. pose changes). The varying psychological state of an individual also might result in variations (e. g. thermal IR images of a fevered person);

Distinctiveness: while a biometric trait is expected to vary significantly across individuals, there may be large inter-class similarities on images used to represent these traits.

Some of the limitations imposed by unimodal face recognition systems can be overcome by using multiple biometrics modalities (such as face and fingerprint of a person). Such systems, known as multimodal biometric systems, are expected to be more reliable due to the presence of multiple and independent evidences [7, 20].

3.2.1 Modes of Operation

A multimodal biometric system can operate in one of three different modes: serial mode, parallel mode, or hierarchical mode [7].

The difference between the serial and the parallel mode is crucial. In the serial mode of operation, the output of one biometric trait is typically used to reduce the number of possible identities before the next trait is tested. This serves as an indexing scheme in an identification system.

The operation in the parallel mode, on the other hand, uses information from multiple traits simultaneously to perform recognition.

In the cascade operational scheme, the various biometric characteristics do not have to be acquired simultaneously. Thus, a decision can be established without the acquirement of all traits, which reduces the overall recognition time. In the hierarchical mode, individual classifiers are combined in a tree-like structure.

3.2.2 Levels of Fusion

As presented by Figure 2.4, a face recognition system can be divided in some modules. The flow of information through these modules decreases as it is processed and/or fused along the steps.

In the literature, several different designations can be used to classify the same levels of fusion. Detailed explanation about them can be found in [7, 21, 22]. Figure 3.5 shows the information fusion taxonomy proposed by Jain et al. [21].

The main distinction regarding the levels of fusion takes into account if they are carried out prior or post the classification. The pre-classification fusion refers to combining information before the application of any classifier or matching algorithm, while in post-classification fusion the information is combined only after the decisions of the classifiers are obtained.

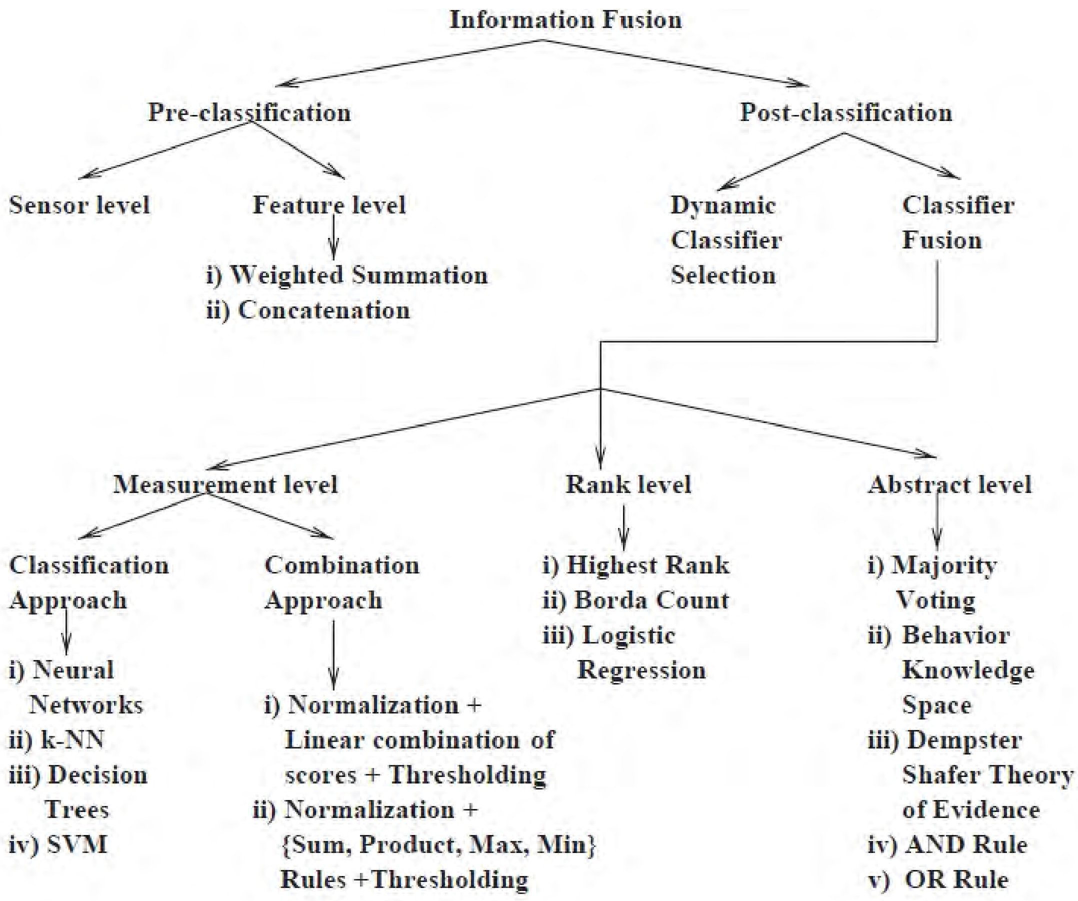


Figure 3.5: Taxonomy of information fusion [21].

According to a somewhat consensus, the levels of fusion can be depict in 4 categories [21]:

Sensor level: the raw data from sensors is combined. Depending on the application, there are two main methods to accomplish this: weighted summation and mosaic construction. In Figure 3.6, for example, the weighted summation can be employed to combine visual (3.6a) and infrared (3.6b) images into a single image (3.6c);

Feature level: the data obtained from each biometric modality is used to compute a single feature vector. If the features extracted from one biometric indicator are independent of those extracted from the other, it is reasonable to concatenate the two vectors into a single new vector, provided that the features from different biometric indicators are in the same measurement scale;

Score level: fusion at the matching score (measurement or rank) level consider the similarity score provided by each biometric matcher. These scores are then combined to assert a final confidence measurement;

Decision level: at this level, also known as abstract level, each biometric system makes its recognition decision based on its own input feature vector. Then, several techniques can be employed to fuse the individual decisions into a final unique decision.

(a) $V(x, y)$ (b) $T(x, y)$ (c) $F(x, y)$

Figure 3.6: Data fusion of visual and thermal images. (a) Visual image, (b) Thermal image, and (c) Average intensity image of (a) and (b) [19].

The integration prior the classification assumes a strong interaction among the input measurements and such schemes are referred to as *tightly coupled* integrations. The *loosely coupled* integration, on the other hand, assumes very little or no interaction among the inputs and integration occurs at the output of the classifiers, each independently estimating the inputs from its manner [7].

Despite the faith that a combination scheme applied as early as possible in the recognition system can be more effective, performing a combination at the data or feature level is more difficult because the relationship between the domains of different biometric systems may not be known and the representations may not be compatible. Additionally, a multimodal system may not have access to the input data of individual classifiers because of their proprietary nature. In such cases, fusions at the matching score or decision levels are the only options. Because of these difficulties, a few papers have been published reporting results on the fusion at the sensor or the feature level [21].

3.2.3 Score Normalization

Fusion at the matching score level can be approached in two distinct ways. In the first approach the fusion is viewed as a classification problem, while in the second approach it is viewed as a combination problem [21]. In the classification approach, a feature vector is constructed using the matching scores output from the individual matchers; this feature vector is then classified into one of two classes: Accepted (genuine user) or Rejected (impostor). In the combination approach, the individual matching scores are combined to generate a single scalar score which is then used to make the final decision. Despite the benefits of the combination approach showed in some studies [23], no single fusion scheme is good under all circumstances [21].

To combine scores from different matchers into a single one, it is necessary to consider some issues. The matching scores at the output of the individual classifiers may not be homogeneous. For example, one classifier may output a distance (dissimilarity) measure while another may output a proximity (similarity) measure. The outputs of the individual classifiers may also not be on the same numerical scale (range). Furthermore, the matching scores of the classifiers may follow different statistical distributions. Concerning these issues, score normalization is necessary to transform the scores of the individual classifiers into a common domain before combining them.

A good normalization scheme must be *robust* and *efficient* with respect to the location and the scale of the matching score distribution. Robustness refers to insensitivity to the presence of outliers, while efficiency refers to the proximity the normalized scores present with respect to the optimal estimate derived from the data distribution [21].

The simplest normalization technique is the *Min-Max*. It is best fitted in cases where the bounds of the scores are known. However, it can be employed in new and not bounded cases based on a prior matching distribution. Given a set of matching scores $s_k, k = 1, 2, ..., n$, the normalized scores are given by

$$s'_k = \frac{s_k - \min}{\max - \min} \tag{3.1}$$

Min-max normalization preserves the original distribution of scores except for a scaling factor and transforms all the scores into a common range $[0, 1]$. Therefore, it is not considered a robust technique due to its sensitiveness to the presence of outliers.

Among a number of normalization techniques, two approaches that are considered robust and efficient are: *double sigmoid function* and *tanh-estimators* [21]. With double sigmoid function the score is normalized by

$$s'_k = \begin{cases} \frac{1}{1+\exp(-2((s_k-t)/r1))} & \text{if} \quad s_k < t \\ \frac{1}{1+\exp(-2((s_k-t)/r2))} & \text{otherwise} \end{cases} \tag{3.2}$$

where t is the reference operating point and $r1$ and $r2$ denote the left and right edges of the region in which the normalization is linear. It also transforms the scores into the $[0, 1]$ interval, but it requires careful tuning of the parameters to obtain good efficiency. Normally, t is chosen to be some value in the region of overlap between the genuine and impostor score distributions, and $r1$ and $r2$ are made equal to the extent of overlap between the two distributions.

By means of the tanh-estimators the score is normalized with

$$s'_k = \frac{1}{2}\left\{\tanh\left(0.01\left(\frac{s_k - \mu_{GH}}{\sigma_{GH}}\right)\right) + 1\right\} \tag{3.3}$$

where μ_{GH} and σ_{GH} are the mean and standard deviation estimates, respectively, of the genuine score distribution as given by the Hampel estimators [21]. These estimators are based on the influence function

$$\psi(u) = \begin{cases} u & 0 \leq |u| < a, \\ a \ \mathrm{sgn}(u) & a \leq |u| < b, \\ a \ \mathrm{sgn}(u) \left(\frac{c-|u|}{c-b}\right) & b \leq |u| < c, \\ 0 & |u| \geq c. \end{cases} \tag{3.4}$$

This function reduces the influence of the points at the tails of the distribution which are identified by the variables a, b, and c. Hence, this method is not sensitive to outliers. If the influence of a large number of tail-points (outliers) is reduced, the estimate is more robust, but not efficient. On the other hand, if many tail-points influence the estimate, it becomes less robust, but the efficiency increases. Therefore, the parameters a, b, and c must be carefully chosen as a trade-off between these two characteristics. The nature of this normalization is such that the genuine score distribution in the transformed domain has a mean of 0.5 and a standard deviation of approximately 0.01. The constant 0.01 in the expression determines the spread of the normalized genuine scores [21].

3.3 Machine Learning Techniques

There are two main topics to concern when dealing with face recognition algorithms. The first one is to construct a "good" face representation space in which the descriptive data become as linearly separable and convex as possible. This includes two levels of processing: (i) normalize face images geometrically and photometrically, for instance, using morphing and histogram equalization; and (ii) extract features from these images which are stable with respect to the natural variations of faces [3].

The second topic is to come up with classifiers that are able to solve difficult nonlinear classification and regression problems in the feature space with good generalization. Although a good normalization and feature extraction reduces the nonlinearity and nonconvexity, they do not solve the problems completely. Hence, classification engines that are able to deal with such difficulties are still necessary to achieve high performance. A successful framework usually combines both strategies [3].

3.3.1 Face Representation

The former techniques to represent faces, which use geometric attributes of face features (e.g., the distances between important points such as eyes, nose, mouth), have the advantage of economy and efficiency when achieving data reduction and insensitivity to variations in illumination and viewpoint. Nevertheless, facial feature detection and geometric measurement techniques developed to date can not cope with reliable geometric feature-based recognition. The reason why early techniques are not effective is that rich information contained in the facial texture or appearance is discarded when only geometric attributes are considered [3].

Appearance-based approaches, such as Principal Component Analysis and *Linear Discriminant Analysis* (LDA) [24, 25], had significantly advanced face recognition techniques. Such approaches generally operates directly on an image-based representation (i.e., array of pixels intensities) from where they extract features to a new derived subspace.

Despite the fact that these linear and holistic appearance-based methods avoid the instability of the early geometric feature-based methods, they are not accurate enough to describe subtleties of original manifolds[2] in the original image space. This is due to their difficulty in handling with all nonlinearity and nonconvexity represented by unrestricted faces instances. Such linear methods can be extended by using nonlinear kernel techniques, such as *kernel PCA* and *kernel LDA*, to deal with nonlinearity in face recognition [3, 26]. In these techniques, a nonlinear projection (dimension expansion) from the image space to a feature space is performed. The manifolds in the resulting feature space becomes more simple, yet with its details preserved. Although kernel methods may achieve good performance on the training data, it may not be so for unknown data due to the flexibility and the overfitting that would be achieved thereof [3]. This kernel issue is briefly described in Section 4.3.4.

Another approach to handle with nonlinearities is to construct a local appearance-based feature space using appropriate image filters so the distributions of faces are less affected by changes. *Local Features Analysis* (LFA), *Elastic Bunch Graph Matching* (EBGM), and *Local Binary Pattern* (LBP) are some of the approaches that have been used to address this assumption [1, 2, 3, 27]. These algorithms usually combine geometric feature detection with local appearance feature extraction with

[2]A manifold is a mathematical space in which every point has a neighborhood which resembles Euclidean space, but in which the global structure may be more complicated.

the intention to increase stability of recognition performance under adverse scenarios. Due to they rely both on feature- and appearance-based characteristics, they are are called hybrids.

Figure 3.7 shows a taxonomy presented by Li and Jain [3] for face recognition algorithms with respect to pose dependency and the features being used to represent the faces.

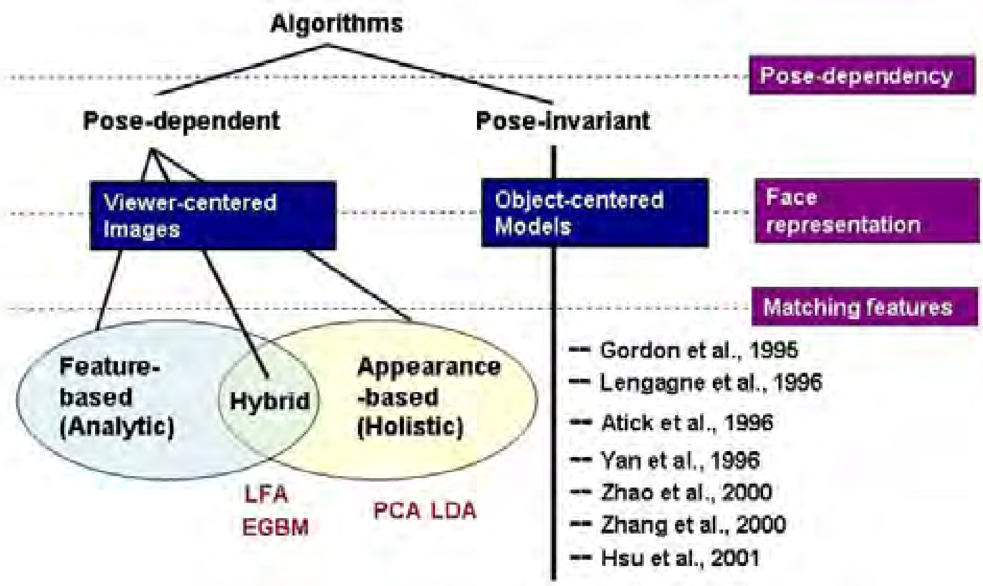


Figure 3.7: Taxonomy of face recognition algorithms based on pose-dependency, face representation, and features used in matching [3].

3.3.2 The Face Recognition Problem

The relationship between face representation and classification methods could lead to a somehow misunderstanding. When a discriminative feature space (representation) is reached, the classification engine can be as simples as the Euclidian or the L_1 distances. However, as the nonlinearity on the feature space grows, more sophisticated classifiers are demanded. Examples of approaches to deal with more intricate problems are the *Bayesian*, the *Dynamic Link Architecture* (DLA) used in the EBGMs, the HMMs, the Neural Networks, and the SVMs. Usually, such sophisticated techniques are based on training samples from where the knowledge to distinguish between faces is taken from.

While classical pattern recognition problems, such as character recognition, have a limited number of classes, typically less than 50, with a large number of training samples available for each category, the face recognition problem have a relatively small number of face images for training and a large number of possible classes of faces. Therefore, the lack of training samples representing the huge amount of possible faces and scenarios setups is the main reason why face recognition training-based algorithms still can not achieve such a good generalization under unconstrained environments.

However, despite training-based classifiers can not take into account training sets which provides all possible face variations, the prior knowledge learned from available face databases has been enabling many successful algorithms for face detection and matching [3].

3.3.3 Well-Known Techniques

The actual number of face recognition approaches is so huge that it would be impossible to enumerate them all. However, the emerging amount of possibilities is more related with different features being extracted than with the feature extraction or the classification procedures. In this Section, a description about few commonly used learning techniques is provided. The purpose is not to cover them in details, but to give conceptual ideas about renowned approaches being used to recognize faces. As the specificities of PCA and SVM are reviewed on Chapter 4, they are not listed below.

Linear Discriminant Analysis: Using PCA, a face subspace is constructed to represent optimally the face, while using LDA, a discriminant subspace is constructed to distinguish optimally faces of different persons. Although one might think that LDA should always outperform PCA, since it deals directly with classes discrimination, empirical evidence suggests that when the number of samples per class is small or when the training data nonuniformly samples the test data distribution, PCA might outperforms LDA [25].

Independent Component Analysis: While PCA finds a set of orthogonal basis for face images of which the transformed features are uncorrelated, ICA seeks non-orthogonal basis that are statistically independent or as independent as possible, generalizing the concept of PCA to higher order image statistics relationships. ICA basis vectors are more spatially local than PCA basis vectors, and local features give better face representations. This local representation reduces the sensitivity of face variations such as small occlusions or pose changes [2].

Hidden Markov Models: Also known just as HMMs, this technique has been successful in modeling temporal information in many speech, image, and video applications. An HMM models a sequence of observations by assuming that there is an underlying sequence of states drawn from a finite state set. When applied for face recognition, it belongs to the class of the hybrids methods. Without finding the exact locations of facial features, HMM based methods use stripes of pixels that cover the forehead, eye, nose, mouth, and chin to formulate the states. These stripes can supply observations from its raw pixels data or from coefficients derived through transformation such as PCA [1, 2].

AdaBoosting: A finding to tackle the feature selection for nonlinear problems is the *AdaBoost* methods. The *boosting* approach has been used successfully in many applications and leads to a framework for learning both effective features and effective classifiers. In boosting, a number of weak classifiers are combined to form a final strong classifier in an iterative procedure [3, 27, 28].

Neural Networks: Neural Network (NN) approaches have been widely explored for face recognition. One of the formers techniques used a single-layer adaptive network containing a separate network for each stored individual. Hybrid NNs have also been proposed as a combination of local image sampling, a *Self-Organizing Map* (SOM) NN, and a convolutional NN. The *probabilistic decision-based* NN has also effectively been applied with this purpose [1]. More recent researches suggest the use of a *non-convergent chaotic* NN to recognize human faces [29]. The fusion of NNs whose outputs are combined to make a final decision on classifying a face, or the integration of NNs with statistical approaches (e.g. PCA space, non-negative matrix factorization) are two other suggested possibilities [29]. As with the other learning approaches, there are a profusion of NNs methods which learns from faces. Once again, they are much more related with different features representations than with the learning theory behind them. One shortcoming of NNs is related with the increasing of the number of people to be recognized. In general, NNs encounter problems when the number of classes increases. Despite the promising performance, the training difficulty has to be further investigated.

Elastic Bunch Graph Matching: Aimed at solving the conceptual problems of conventional NNs, the EBGM approach is based on the neural information processing concept, the DLA. In EBGM, comparing two faces corresponds to matching and adapting grids taken from two distinct images. In such a procedure, rotation in depth is compensated by elastic deformation of the graphs [2, 30].

3-D Approaches: The main advantage of using 3-D data is that depth information does not depend on pose and illumination and therefore the representation of the object does not change with these parameters. The first contributions to the field were mainly based on the extraction of a *curvature* representation of the face. Another family of approaches rely on a transformation which maps the *geodesic distances* between a certain numbers of sample points on the facial surface to less complex distances, such as Euclidean. In the 3-D *model-based* approaches the main idea is to fit distinct models, or a synthesized image to a model. This fitting leads to a measurement which is used to recognize the person. In this way, recognition can be based on the model's coefficients or some other feature- or appearance-based approach. A famous technique for surface matching is called *Iterative Closest Point* (ICP), which aims at finding the best correspondence between points from two surfaces. Despite their advantages, a drawback of 3-D model-based approaches is the high computational burden on the synthesizing and fitting procedures [13, 15, 30, 31].

As previously mentioned, there are many studies emerging every year proposing different methods for face recognition. Chapter 6 presents details about the two new methods covered in the experiments. Actually, Multispectral fusion has been investigated before, but not with focus on providing results about classifiers dependency and score normalization techniques. The SVM approach, however, is believed to be new due to its feature extraction and face representation procedures. The explored training setups and additional observations about dealing with SVMs have also the intention to provide new information about issues related to the SVMs experimentation.

Chapter 4

Adopted Recognition Techniques

Faces are usually represented by digital images whose dimension mn related to the number of pixels is a high number even for small images. Methods which operate in this pure representation has a number of potential shortcomings, most of them related to the well-known curse of dimensionality. However, much of the surface of a face is smooth and has regular texture, which means that the value of a pixel is typically highly correlated with the values of its neighbors. Moreover, faces constraints such as symmetry can be also considered somehow a kind of redundancy in its representation [3].

The sensing of faces represented as high-dimensional arrays of pixels often belong to a manifold of lower dimension. As a consequence, face recognition and computer vision research in general has shown increasingly interest in techniques that take advantage of this observation and apply algebraic and statistical tools to extract and analyze such manifolds. The features in these manifolds (i.e., subspaces) provide more prominent and richer information for recognition than the raw image. The use of subspace modeling techniques has significantly advanced face recognition technology [3].

To deal with the dimensionality problem, two different approaches are presented: PCA and Census Histograms (CH). To classify the features determined by the PCA, distance measurements are employed. On the other hand, to establish similarities from the CH feature space, SVMs are used.

4.1 Principal Component Analysis

One of the most versatile approaches for face recognition is derived from the statistical technique called Principal Component Analysis [32]. Popularized by Turk and Pentland [24], the use of PCA to represent faces is based on the idea that face recognition can be accomplished with a small set of features that best approximates the set of known facial images. Application of PCA for face recognition proceeds by first performing PCA on a set of training images of known human faces. From this analysis, a set of *principal components* is obtained, and the projection of the test faces on these components is used in order to compare them with each previously projected faces (i.e., the database). This leads not only to computational efficiency, but also makes the recognition more general and robust.

PCA is a dimensionality reduction technique based on extracting the desired number of principal components of the multidimensional data. The first principal component is the linear combination of the original dimensions that has the maximum *variance*. Hence, the n^{th} principal component is the linear combination with the highest variance subjected to being *orthogonal* to the $n-1$ first principal components.

These principal components, or basis vectors in the transformed space, can be calculated in the following way [32]. Let X be the $N \times M$ data matrix whose columns $x_1, ..., x_M$ are observations of a signal embedded in $\mathbb{R}^N$. The PCA basis Φ is obtained by solving the eigenvalue problem

$$\Lambda = \Phi^T \Sigma \Phi \tag{4.1}$$

where Σ is the covariance matrix of the data,

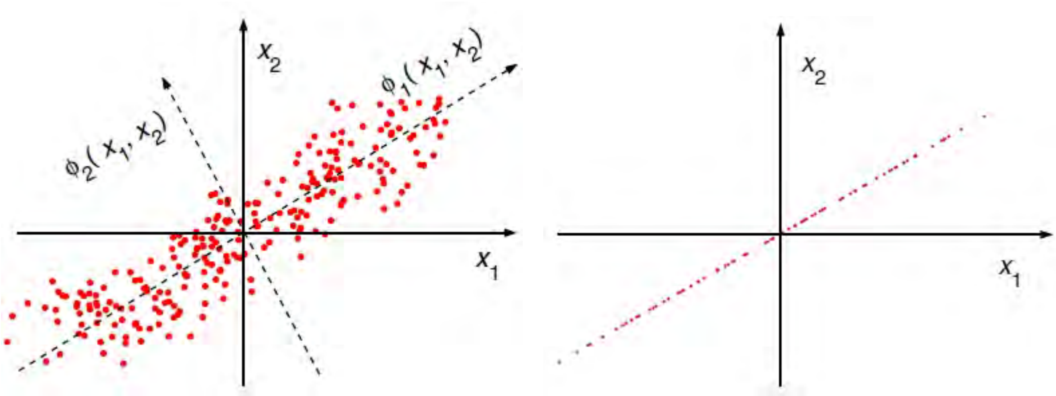
$$\Sigma = \frac{1}{M} \sum_{i=1}^{M} x_i x_i^T \tag{4.2}$$

$\Phi = [\phi_1 ..., \phi_m]^T$ is the eigenvector matrix of Σ, and Λ is the diagonal matrix with eigenvalues $\lambda_1 \geq ... \geq \lambda_N$ of Σ on its main diagonal. In this manner, ϕ_j is the eigenvector corresponding to the j^{th} largest eigenvalue, being λ_j also the variance of the data projected on it.

Hence, to extract k principal components of the data, one must project the data onto Φ_k - the first k columns of the PCA basis Φ which correspond to the k highest eigenvalues of Σ. This can be seen as a linear projection $\mathbb{R}^N \rightarrow \mathbb{R}^k$, which retains the maximum energy (i.e., variance) of the signal. Another important property of

PCA, is that it *decorrelates* the data, with the covariance matrix of $\Phi_k^T X$ always being diagonal. This comes from the orthogonality of the principal components Φ previously mentioned [32].

The idea of PCA is shown in Figure 4.1(a), where the axis labeled ϕ_1 corresponds to the direction of maximum variance and is chosen as the first principal component thereof. In a two-dimensional case, the second principal component is the remaining perpendicular axis ϕ_2. However, in a higher-dimensional space, the selection process would continue dictated by the variances of the projections. Figure 4.1(b) show how the original data is expressed only with the first principal component. Even being the most discriminant way for a one-dimensional projection of the dataset, it is possible to note some loss of information.



(a) PCA Basis (b) PCA reduction to 1D

Figure 4.1: The outline of PCA. (a) The solid axis represent the original basis, while the dashed axis represents the PCA basis. (b) The projection of the data using only the first principal component [32].

When the matrix X is small, PCA is rather non expensive to calculate. However, as X grows, the computation of Σ (eq.4.2) becomes quite expensive. Fortunately, PCA may be implemented via an iterative method called *Singular Value Decomposition* (SVD) [32]. The SVD of an $N \times M$ matrix $X(N \geq M)$ is given by

$$X = UDV^T \tag{4.3}$$

where the $N \times M$ matrix U and the $M \times M$ matrix V have orthonormal columns, and the $M \times M$ matrix D has the square root of the eigenvalues of XX^T on its diagonal entries, the singular values of X.

It can be shown that $U = \Phi$, so SVD allows efficient and robust computation of PCA without the need to estimate the data covariance matrix. When the number of examples M is much smaller than the dimension N, this is a definitive advantage.

4.1.1 From Faces to Feature Space

In the context of face recognition, the variables from the previous section M and $N = mn$ are the *number of available training images* of faces and the *number of pixels* of these images, respectively.

It is important to observe that each face has to be normalized in terms of rotation and scaling. This is normally carried out based on the location of the eyes. Despite the evidences about the principal components, it is recommended that the images be also illumination normalized due to the holistic property of such method. The following steps illustrates the outline of PCA for face recognition [24]:

1. Express each of the M training faces as one-dimensional vectors Γ of length N such that the rows of pixels in the image are concatenated one after the other;

2. Subtract the average face $\Psi = \frac{1}{M}\sum_{n=1}^{M}\Gamma_n$ of each Γ such that $X_i = \Gamma_i - \Psi$. This leads the mean on X to become zero;

3. Calculate the data covariance matrix Σ from X (equation 4.2);

4. Determine the eigenvalues Λ and eigenvectors Φ according to equation (4.1);

5. Select the Φ_k principal components concerning the highest k eigenvalues of Σ or adopting other empirical-based approach [25, 33].

Because the vectors Φ_k are the eigenvectors of the covariance matrix corresponding to the original face images, and because they are face-like in appearance, they are called "eigenfaces". Figure 4.2 shows an example of the average face and seven top eigenfaces.

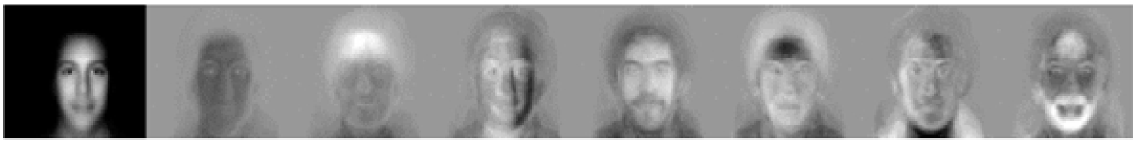


Figure 4.2: Eigenfaces: the average face on the left, followed by seven top eigenfaces. Extracted from [32].

Once the eigenfaces are created, the identification becomes a pattern recognition task. A new face image Γ is transformed into its eigenface (projected into the face space) by the operation $\Omega = \Phi_k(\Gamma - \Psi)$, which is a set of point-by-point image multiplications and summations, with the elements of Ω describing the contribution of each eigenface in representing the input face image.

To reach the dimensionality reduction, the basis vectors can be truncated not only from the front, but also from the back or any other position of the eigenvector matrix. Former researches suggests that front eigenvectors corresponds to not useful information, such as lighting variations, and the back eigenvectors corresponds to noise in the images [25, 33]. Thus, not only highly correlated face data (e.g., texture and symmetry), but also undesirable information can be eliminated.

4.1.2 Similarity Measurement

Due to the fact that PCA eliminates all of the statistical covariance in the transformed feature space, distance measures can be exploited as a confident way to predict similarity.

The simplest method for determining which face class provides the best description of an input face image p is to find the face class g that minimizes the distance $D = ||\Omega_g - \Omega_p||$, where Ω_g and Ω_p are arbitrary eigenfaces from the galery and the probe sets respectively. A face is classified as belonging to class g when the minimum D is bellow some chosen threshold t. Otherwise, the face is classified as "unknown".

As already mentioned, the Euclidean distance, also known as *2-Norm* or *L2 Distance*, is the most common similarity measurement employed on PCA. It derives from the family of the $L_{\mathcal{P}}$ *distances* with $\mathcal{P}$ particularly set to 2. Given a probe Ω_p and a gallery Ω_g, the $L_{\mathcal{P}}$ *distance* can be define as

$$L_{\mathcal{P}}(p, g) = \sqrt[\mathcal{P}]{\sum_{i=1}^{N} |\Omega_{pi} - \Omega_{gi}|^{\mathcal{P}}} \tag{4.4}$$

Among other metrics, the *L1 distance* (where $\mathcal{P} = 1$) has also been proposed to assess the similarity of faces. Such a distance is also called *1-Norm*, *City Block*, or *Manhattan distance*.

Another well-known and effective group of distances is related to measurements taken from the *Mahalanobis space* [34], which is a space where the variance along each dimension is one. Therefore, to transform an eigenface into the Mahalanobis feature space it is necessary to divide each coefficient in the vector by its standard deviation, which is the square root of the corresponding eigenvector. This transformation yields a feature space with unit variance in each dimension.

A probe and a gallery eigenvector can be compared with the distances mentioned above, which leads to the *Mahalanobis L1* and the *Mahalanobis L2* distances. Nevertheless, a very common approach in the Mahalanobis space is carried out by computing the *Cosine* or the *angle* between the already transformed Ω_p and Ω_g. These operation is made based on the *dot product* properties (Section 4.3.1), and can be defined as

$$M_C(p, g) = \frac{\Omega_p.\Omega_g}{|\Omega_p|.|\Omega_g|} \tag{4.5}$$

and

$$M_A(p, g) = arccos\left(\frac{\Omega_p.\Omega_g}{|\Omega_p|.|\Omega_g|}\right) \tag{4.6}$$

which leads to the called *Mahalanobis Cosine* and *Mahalanobis Angle* distances.

To close this section, the concept of *k-Nearest Neighbor* (kNN) [35] is briefly described. In a scenario where the gallery G is formed by more than one sample per subject, it may be interesting to turn on a majority voting scheme considering the kNN from the probe Ω_p. Thus, the identity of p can be assigned to the most common subjects in G amongst its k nearest neighbors. This technique is well used in many applications and can be employed with any of the reviewed distance metrics.

4.2 Census Histograms as Face Features

The representation of faces for the SVM learning process in the literature varies from using the raw image data [36, 37] to combining extracted facial components into a single vector [38]. Another approach, which seems to be more popular, is to apply PCA in the training set before submitting the data to the SVM [5, 39]. ICA in the place of PCA has also been assessed [40].

In this thesis, the Census Histograms (CH) is applied as the proposed model to describe faces to the SVM. The CH is a technique based on the Census Transform that leads to a structural description of the images and has been successfully used in many practical applications [8, 41].

The features used as the basis of the SVMs inputs can be defined as structure kernels of size 3×3 which summarize the local spatial image structure. Therefore, they are called *Local Structure Features*. Within such kernels, the structure is expressed as binary information $\{0, 1\}$ and the resulting binary patterns can represent oriented edges, line segments, junctions, ridges, saddle points, etc [8]. On the referred 3×3 lattice, there exist $2^9 = 512$ possible kernels[1], and the matching kernel can thus be used to describe each position of a given image. Figure 4.3 shows some examples of structure kernels.

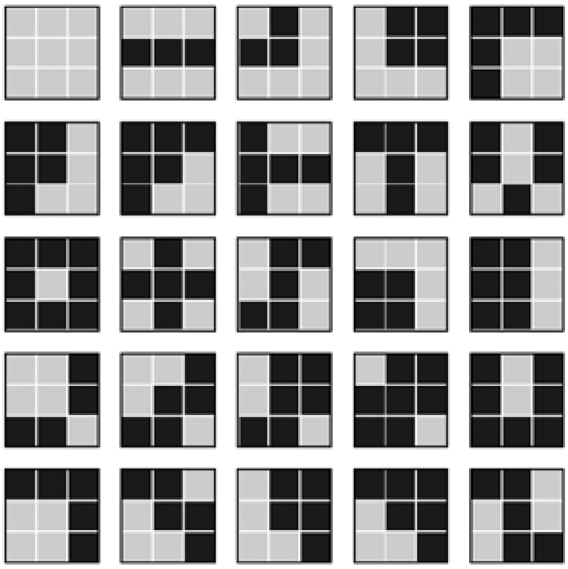


Figure 4.3: A randomly chosen subset of 25 out of $2^9 - 1$ possible Local Structure Kernels in a 3×3 neighborhood [8].

The *Census Transform (CT)* is a non-parametric local transform which was first proposed in [6] and follows the principle of structure kernels. It can be thought as an ordered set of intensities comparisons between a center pixel and its local neighborhood. Despite the size of the local neighborhood is not restricted, it is usually assumed as a 3×3 region. Let $\mathcal{N}(\mathbf{x})$ denote a local spatial neighborhood of the pixel $\mathbf{x}$ so that $\mathbf{x} \notin \mathcal{N}(\mathbf{x})$. The goal of CT is to generate a bit string representing which pixels in $\mathcal{N}(x)$ have an intensity lower than $\mathbf{I}(\mathbf{x})$. Concerning that pixel intensities are always zero or positive, the formal definition can be stated as follows: Let a comparison function $\zeta(\mathbf{I}(\mathbf{x}), \mathbf{I}(\mathbf{x}'))$ be 1 if $\mathbf{I}(\mathbf{x}) < \mathbf{I}(\mathbf{x}')$ and let $\otimes$ denote the concatenation operation, the Census Transform at $\mathbf{x}$ is defined as

$$\mathcal{C}(\mathbf{x}) = \otimes_{\mathbf{y} \in \mathcal{N}} \zeta(\mathbf{I}(\mathbf{x}), \mathbf{I}(\mathbf{y})) \tag{4.7}$$

[1]Actually, there are only $2^9 - 1$ reasonable kernels out of the 2^9 possible ones. This is because the kernel with all elements 0 and that with all 1 convey the same information (all pixels are equal) and so one is excluded.

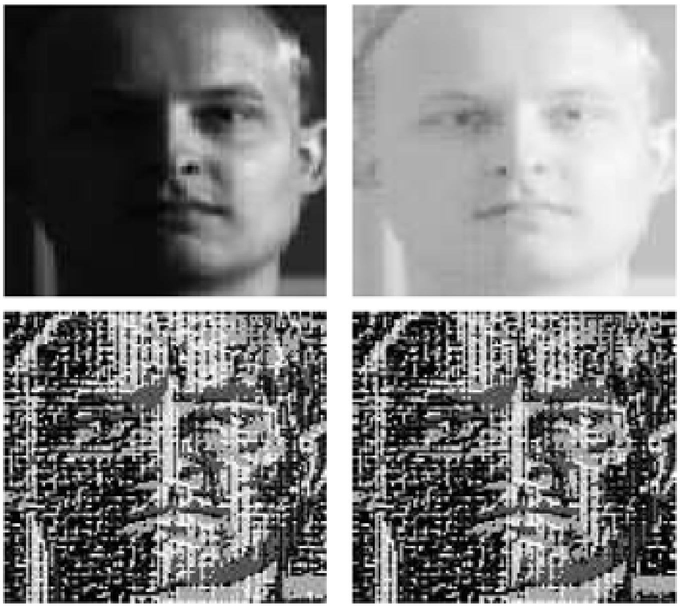


Figure 4.4: Example of the illumination invariance of the Census Transform. Despite the illumination in the second synthetically changed image varies considerably, the transformation results almost the same. Here the CT image is visualized taking kernel indexes as gray values [8].

It is important to observe that all bits of $\mathcal{C}(\mathbf{x})$ have the same significance level. Furthermore, different from linear transforms, the CT is not related to intensity or similarity. In fact, $\mathcal{C}(\mathbf{x})$ may be interpreted as the index of a structure kernel defined on $\mathcal{N}(\mathbf{x})$ with the center set to zero and the other kernel components (i.e., pixels) representing the outcome of the census comparisons $\zeta(.)$. This is the way Census Transform is connected to local structure kernels.

An image transformation which preserves the order of its arguments is stable with respect to the application of a linear and monotonic changing on the reflectance $R(\mathbf{x})$ due to illumination. This means that the Census Transform, which relies on the local intensity order, is unaffected by linear lighting variance [8]. In Figure 4.4 an illustrative example is provided, where the Census Transform is visualized as an index image with the kernel index determining the pixel intensity.

A Census Histogram is simply a histogram built from *Census Features* and expresses the structure kernels distribution in a given image region. Based on [27], the proposed method leads to the Difference Space (Section 4.3.5) by computing and summing the L_1 distance (as in Section 4.4 with $\mathcal{P} = 1$) between each bin of two Census Histograms from the same region of different face images.

The procedure is based on a scanning window which shifts and changes its scale over pairs of images, extracts the local Census Histograms and compute the dissimilarity between two corresponding local histograms. If both images are from the same identity, the dissimilarity measurements are labelled as positive features, otherwise as negative features. This scanning window leads to a sequence of regions $R_0, R_1, ..., R_m$. Assuming $c_l(x, y)$ as a census labelled image, a histogram of a region m can be defined as

$$H_{m,i} = \sum_{x,y} I\{c_l(x, y) = i\} \quad \text{subject to} \quad (x, y) \in R_m \tag{4.8}$$

where i represents the bins of the histogram, i.e., the kernel indexes.

After all, the resulting $L1$ dissimilarity between pairs of Census Histograms from corresponding regions of two different images comprises a SVM input vector.

4.3 Support Vector Machines

Commonly considered as the first practical derivation of statistical learning theory, Support Vector Machines (SVM) represents at the present time a field of research that has a large choice of topics to work on, and many of the issues are conceptual rather than merely technical [26, 42]. Over the last years, its scope has widened significantly, both in terms of new algorithms, such as kernel methods, and in terms of a deeper theoretical understanding [26]. It has become clear that kernel methods provide a framework to deal with profound issues in machine learning theory. Meanwhile, successful applications have demonstrated that SVMs not only have a better foundation than artificial neural networks, but are able to serve as a replacement for neural networks that perform as well or better in a number of fields [26].

In this thesis, the covered theory is introductory and the discussed concepts is narrowly related to SVMs for pattern recognition problems. A considerable part of this Section was condensed from the extensive study of SVMs provided by Schölkopf and Smola [26].

4.3.1 Basic Concepts

One of the fundamental problems of learning theory is stated as: given two classes of known objects, assign one of them to a new unknown object. Considering a given empirical data set, this problem can be formalized as follows [26]:

$$(x_1, y_1), ..., (x_m, y_m) \in \mathcal{X} \times \{\pm 1\} \tag{4.9}$$

Here $\mathcal{X}$ is some nonempty set usually referred to as the *domain* from where the *patterns* x_i (also known as *cases, inputs*, or *instances*) are taken, and y_i are called *labels, targets*, or *outputs*. In such a problem, there are only two classes of patterns, which for mathematical convenience are labelled by $+1$ and -1. This is a particularly simple situation, referred to as *(binary) pattern recognition* or *(binary) classification*.

In learning, the purpose is to be able to *generalize* the categorization of patterns to unknown data points. In the case of pattern recognition, this means that given some new pattern $x \in \mathcal{X}$, the corresponding $y \in \{\pm 1\}$ has to be predicted. In other words, it indicates that a y has to be chosen such that (x, y) is in some sense similar to the training examples, which leads to the notions of *similarity* in $\mathcal{X}$ and in $\{\pm 1\}$. The choice of this similarity measure for the inputs lies at the foundations of the machine learning area.

A simple type of similarity measure that is of particular mathematical appeal is the dot product. Given two vectors $x, x' \in \mathbb{R}^N$, the *canonical dot product* is

$$\langle \mathbf{x}, \mathbf{x}' \rangle = \sum_{i=1}^{N} [\mathbf{x}]_i [\mathbf{x}']_i \tag{4.10}$$

where $[x]_i$ denotes the ith element of x.

Also referred to as *inner product* or *scalar product*, the geometric interpretation of the canonical dot product is that it computes the cosine of the angle between the vectors x and x', provided they are normalized to length 1. In addition, it allows computation of the *length* (or *norm*) of a vector x as

$$||\mathbf{x}|| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \tag{4.11}$$

Similarly, the distance between two vectors is computed as the length of the difference vector. Therefore, being able to compute dot products amounts to being able to carry out all geometric constructions that can be formulated in terms of angles, lengths and distances.

In order to be able to use a dot product as a similarity measure, the patterns need to be represented as vectors in some dot product space $\mathcal{H}$. To this end, the map

$$\begin{aligned} \Phi :& \mathcal{X} \rightarrow \mathcal{H} \\ & x \rightarrow \mathbf{x} = \Phi(x) \end{aligned} \tag{4.12}$$

is used [26]. Even if the original patterns exist in a dot product space, it may still be desirable to considered more general similarity measures by applying a nonlinear Φ, which can lead to a more suitable representation of the given problem. Henceforth, the space $\mathcal{H}$ is called *feature space*. To summarize, embedding the data into $\mathcal{H}$ via Φ has three benefits [26]:

1. It lets define a similarity measure from the dot product in $\mathcal{H}$,

$$k(x, x') = \langle x, x' \rangle = \langle \Phi(x), \Phi(x') \rangle \tag{4.13}$$

2. It allows to deal with the patterns geometrically, and thus lets study learning algorithms using linear algebra and analytic geometry.
3. The independent choice of the mapping Φ enables the design of a large variety of similarity measures and learning algorithms.

With the representation and similarities above in mind, a simple pattern recognition learning algorithm can be described. Assuming that the data are embedded into a dot product space $\mathcal{H}$, the basic idea of the algorithm is to assign a previously unseen pattern to the class with closer mean. In this way, it starts by computing the means of the two classes in feature space

$$\mathbf{c}_+ = \frac{1}{m_+} \sum_{\{i|y_i=+1\}} \mathbf{x}_i \tag{4.14}$$

$$\mathbf{c}_- = \frac{1}{m_-} \sum_{\{i|y_i=-1\}} \mathbf{x}_i \tag{4.15}$$

where m_+ and m_- are the number of existent positive and negative labelled examples, respectively. Half way between $\mathbf{c}_+$ and $\mathbf{c}_-$ lies the point $\mathbf{c} = (\mathbf{c}_+ + \mathbf{c}_-)/2$, and thus, the class of $\mathbf{x}$ is computed by checking whether the vector $\mathbf{x} - \mathbf{c}$ connecting $\mathbf{c}$ to $\mathbf{x}$

encloses an angle smaller than $\pi/2$ with the vector $\mathbf{w} = \mathbf{c}_+ - \mathbf{c}_-$ connecting the class means. This lead to

$$\begin{aligned} y &= sgn\langle(\mathbf{x} - \mathbf{c}), \mathbf{w}\rangle \\ &= sgn\langle(\mathbf{x} - (\mathbf{c}_+ + \mathbf{c}_-)/2), (\mathbf{c}_+ - \mathbf{c}_-)\rangle \\ &= sgn(\langle\mathbf{x}, \mathbf{c}_+\rangle - \langle\mathbf{x}, \mathbf{c}_-\rangle + b) \end{aligned} \tag{4.16}$$

where b is an offset defined by

$$b = \frac{1}{2}(||\mathbf{c}_-||^2 - ||\mathbf{c}_+||^2) \tag{4.17}$$

with the norm $||\mathbf{x}|| = \sqrt{\langle\mathbf{x}, \mathbf{x}\rangle}$. If the class means have the same distance to the origin, then $b = 0$.

It is important to note that (4.16) produces a decision boundary which has the form of a *hyperplane* (Figure 4.5). That is, a set of points that satisfy a constraint expressible as a linear equation.

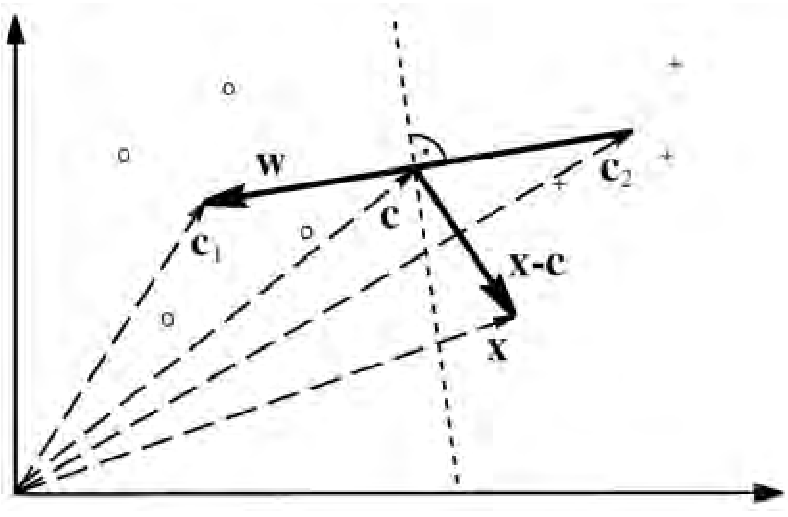


Figure 4.5: A simple geometric classification algorithm: given two classes of points (depicted by 'o' and '+'), compute their means $\mathbf{c}_1, \mathbf{c}_2$ and assign a test pattern $\mathbf{x}$ to the one whose mean is closer. This can be done by looking at the dot product between $\mathbf{x} - \mathbf{c}$ (where $\mathbf{c} = (\mathbf{c}_1 + \mathbf{c}_2)/2$) and $\mathbf{w} = \mathbf{c}_1 - \mathbf{c}_2$, which changes sign as the enclosed angle passes through $\pi/2$. The corresponding decision boundary is a hyperplane (the dotted line) orthogonal to $\mathbf{w}$. From [26].

It is worth to rewrite (4.16) in terms of the input patterns $\mathbf{x}_i$, using the kernel k to compute the dot products. As (4.13) only computes dot products between vectorial representations $\mathbf{x}_i$ of inputs x_i, it is necessary to express the vectors $\mathbf{c}_i$ and $\mathbf{w}$ in terms of $\mathbf{x}_1, ..., \mathbf{x}_m$. To this purpose, the substitution of (4.14) and (4.15) into (4.16) leads to the *decision function* [26]

$$\begin{aligned} y &= sgn\left(\frac{1}{m_+}\sum_{\{i|y_i=+1\}}\langle\mathbf{x},\mathbf{x}_i\rangle - \frac{1}{m_-}\sum_{\{i|y_i=-1\}}\langle\mathbf{x},\mathbf{x}_i\rangle + b\right) \\ &= sgn\left(\frac{1}{m_+}\sum_{\{i|y_i=+1\}}k(x,x_i) - \frac{1}{m_-}\sum_{\{i|y_i=-1\}}k(x,x_i) + b\right) \end{aligned} \tag{4.18}$$

and likewise, the offset becomes

$$b = \frac{1}{2}\left(\frac{1}{m_-^2}\sum_{\{(i,j)|y_i=y_j=-1\}}k(x_i,x_j) - \frac{1}{m_+^2}\sum_{\{(i,j)|y_i=y_j=+1\}}k(x_i,x_j)\right) \tag{4.19}$$

The classifier given by equation (4.18) is quite close to a SVM classifier. Both take the form of kernel expansions on the input domain,

$$y = sgn\left(\sum_{i=1}^{m}\alpha_i k(x,x_i) + b\right) \tag{4.20}$$

In both cases, the expansions correspond to a separating hyperplane in a feature space. In this manner, α_i can be considered a *dual representation* of the hyperplane's normal vector. Both classifiers are example-based in the sense that the kernels are centered on the training patterns, which means that one of the two arguments of the kernel is always a training pattern. A test point is then classified by comparing it to all the training points that appear in (4.20) with a nonzero weight [26].

The SVM algorithm differs from (4.16) mainly in the selection of the patterns on which the kernels are centered and in the choice of weights α_i that are placed on the individual kernels in the decision function. However, not all training patterns will appear in the kernel expansion in SVMs (there will be $\alpha_i = 0$), and the weights of the kernels in the expansion will no longer be uniform within the classes – concerning that according to 4.16 and 4.20, α_i has a dual and uniform representation being either $\alpha_i = (1/m_+)$ or $\alpha_i = (1/m_-)$ depending on the class to which the pattern belongs [26].

In the feature space representation, this statement corresponds to saying that the normal vector $\mathbf{w}$ of the decision hyperplanes can be represented as general linear combinations (i.e., with non-uniform coefficients) of the training patterns. With this assumption, it is possible to remove the influence of patterns that are very far away from the decision boundary since it is expected that they will not improve the generalization error of the decision function, or since is desirable to reduce the computational cost of evaluating the decision function (4.20). The hyperplane will then only depend on a subset of training patterns called *Support Vectors*s (SVs) [26].

4.3.2 Excerpts from Statistical Learning Theory

As previously stated, the objective in a two-class pattern recognition is to infer a function

$$f : \mathcal{X} \to \{\pm 1\} \tag{4.21}$$

based on the input-output training data.

Figure 4.6 shows a simple example of different models applied to classify the same problem. The task is to separate the solid dots from the circles by finding a function which takes the value 1 on the dots and -1 on the circles. In the rightmost plot, the classification function correctly separates all training points. However, it is unclear if the same would happen for test points which stem from the same underlying regularity. For example, new test points lying close to one of the two outliers located in the middle of the opposite class. Perhaps it is better not to allow the decision function to come up with own custom-made regions for them, which leads to a simpler model. The leftmost picture shows an almost linear separation of the classes. This separation, however, not only misclassifies the above two outliers, but also a number of easy points which are so close to the decision boundary that the classifier should be able to get them right. At last, the central picture represents a trade-off by using a model with an intermediate complexity, which gets most points right, without trusting too much in any individual point [26].

Statistical learning theory is aimed at placing these intuitive arguments in a mathematical framework. To this end, it studies the mathematical properties of the function class that a learning machine can implement.

Assuming that the data are generated independently from some unknown (but fixed) probability distribution $P(x, y)$ (Independent and Identically Distributed), the

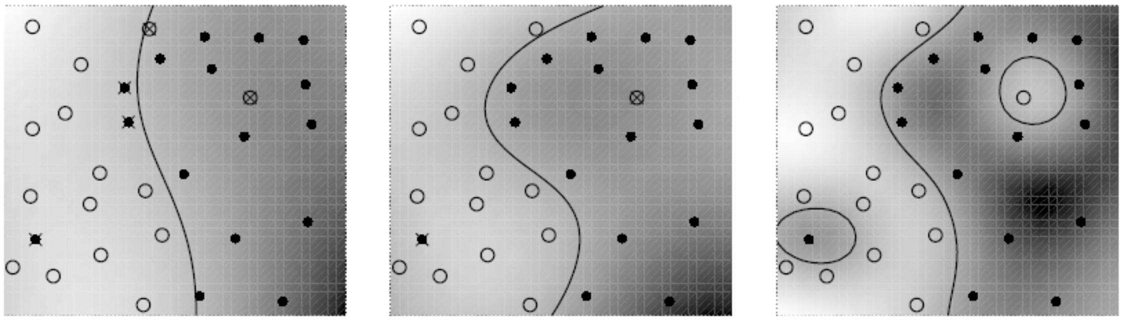


Figure 4.6: A simple example of different models. The models vary in complexity, ranging from a simple one (left), which misclassifies a large number of points, to a complex one (right), which believes in each point and comes up with solution that is consistent with all training points, but may not work well on new points [26].

goal is to find a function f that will correctly classify unseen examples (x, y) also generated from $P(x, y)$, such that $f(x) = y$. The accuracy of the classification can be measured by the *zero-one loss function* $c(x, y, f(x)) = \frac{1}{2}|f(x) - y|$, which is 0 if $f(x, y)$ is correctly classified, and 1 otherwise.

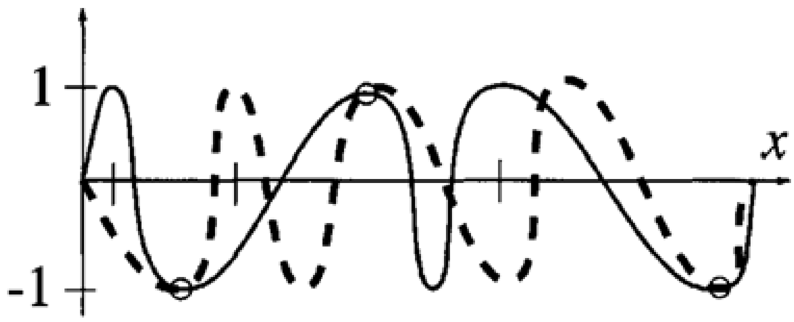


Figure 4.7: Set of functions applied to the same 1D problem, with a training set of three points (circles) and three test inputs (marked on x-axis). Classification is performed by thresholding the values of the functions. Both functions (dotted line, and solid line) perfectly explain the training data, but they give opposite predictions on the test inputs. With no additional information, it is not possible to decide which of the two functions is to be desirable [26].

If no restriction is made on the set of functions from which the chosen f arises, then even a function that does very well on the training data may not generalize well to test examples. Figure 4.7 illustrates an example of how two different functions can diverge on the test data despite they had perfectly agreed about the training data. As only the training data is available, there is no means to select which of the two functions is preferable. Through this statement, it is possible to conclude that only minimizing the *empirical risk* (or *average training error*),

$$R_{emp}[f] = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}|f(x_i) - y_i| \tag{4.22}$$

does not lead to the smallest *test error* (called *risk*),

$$R[f] = \int \frac{1}{2}|f(x) - y|dP(x,y) \tag{4.23}$$

averaged over test examples drawn from the same distribution $P(x, y)$ [26].

Statistical learning theory, or *Vapnik-Chervonenkis* (VC) theory [26, 42], shows that it is obligatory to restrict the set of functions from which f is chosen to one that has an appropriate *capacity* for the amount of available training data, providing *bounds* for the test error. The minimization of these bounds, which depend on both the empirical risk and the capacity of the function class, leads to the principle of *structural risk minimization* [42].

The capacity concept known as *VC dimension* is defined as follows: each function of the class separates the patterns in a certain way and thus induces a certain labelling of the patterns. Assuming the labels are in $\{\pm 1\}$, there are at most 2^m different labellings for m patterns. A very rich function class might be able to realize all 2^m separations, in which case it is said to *shatter* the m points. Nevertheless, a given class of functions might not be sufficiently rich to shatter the m points. The VC dimension is thus defined as the largest m such that there exists a set of m points which the class can shatter, or ∞ if no m exists. It can be thought of as a one-number indicator of a learning machine's capacity [26].

Assuming that $h < m$ is the VC dimension of the class of functions that the learning machine cope with, then for all functions of that class with a probability of at least $1 - \delta$ over the drawing of the training sample, the bound

$$R[f] \leq R_{emp}[f] + \phi(h, m, \delta) \tag{4.24}$$

holds, where the *confidence term* (or *capacity term*) ϕ is defined as

$$\phi(h, m, \delta) = \sqrt{\frac{1}{m}\left(h\left(ln\frac{2_m}{h} + 1\right) + ln\frac{4}{\delta}\right)} \tag{4.25}$$

with δ specifying the probability with which the bound is held [26].

Given a training set of finite size, and provided that there are no examples contradicting each other, it is always possible to come up with a learning machine which achieves zero training error. However, to correctly separate all training examples, this machine will necessarily require a large VC dimension h. This imply that the confidence term (4.25), which increases monotonically with h, will be large, and the bound (4.24) will show that the small training error does not guarantee a small test error. In order to get significant predictions from (4.24), the function class must be restricted such that its VC dimension is small enough, while it must be enlarged in order to provide functions that are able to model the dependencies hidden in $P(x, y)$ [26]. The choice of this set of functions is the core of learning from data.

4.3.3 Hyperplane Classifiers

According to the highlighted components for controlling the statistical effectiveness in the design of learning algorithms, it is mandatory that the capacity of the class of functions can be computed. In his former time, Vapnik [42] considered the class of hyperplanes in some dot product space $\mathcal{H}$,

$$\langle \mathbf{w}, \mathbf{x} \rangle + b = 0 \tag{4.26}$$

where $\mathbf{w}, \mathbf{x} \in \mathcal{H}, b \in \mathbb{R}$, corresponding to decision functions

$$f(x) = \operatorname{sgn}(\langle \mathbf{w}, \mathbf{x} \rangle + b) \tag{4.27}$$

and, based on two arguments, he proposed the *Generalized Portrait* learning algorithm for problems which are separable by hyperplanes:

1. Among all hyperplanes separating the data, there exists a unique *optimal hyperplane*, distinguished by the maximum margin of separation between any training point and the hyperplane. This is the solution of

$$\underset{\mathbf{w} \in \mathcal{H}, b \in \mathbb{R}}{\text{maximize}} \quad \min\{||\mathbf{x} - \mathbf{x}_i|| \,|\, \mathbf{x} \in \mathcal{H}, \langle \mathbf{w}, \mathbf{x} \rangle + b = 0, i = 1, ..., m\} \tag{4.28}$$

2. The capacity of the class of separating hyperplanes decreases with increasing margin.

Despite the similarity of the decision function (4.27) with the earlier example (4.16), the way in which they are established is quite different. Instead of computing the normal vector $\mathbf{w}$ from the mean of the classes, the present case will need some additional work to find the normal vector that leads to the largest margin. To construct the optimal hyperplane, it is necessary to solve

$$\underset{\mathbf{w}\in\mathcal{H},b\in\mathbb{R}}{\text{minimize}} \quad \tau(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2 \tag{4.29}$$

subject to

$$y_i(\langle\mathbf{w},\mathbf{x}_i\rangle + b) \geq 1 \quad \text{for all} \quad i = 1, ..., m \tag{4.30}$$

with the constraint (4.30) ensuring that $f(x_i)$ will be $+1$ for $y_i = +1$ and -1 for $y_i = -1$, and also fixing the scale of $\mathbf{w}$.

The reason behind the minimization of $\mathbf{w}$ (4.29) can be stated as follows: if $||\mathbf{w}||$ were 1, then the left hand side of (4.30) would equal the distance from x_i to the hyperplane. In general, it is necessary to divide $y_i(\langle\mathbf{w},\mathbf{x}_i\rangle + b)$ by $||\mathbf{w}||$ to transform it into this distance. Hence, if (4.30) is satisfied for all $i = 1, ..., m$ with a $\mathbf{w}$ of minimal length, the overall margin will be maximized. A summary of these arguments is given in Figure 4.8.

The function τ in (4.29) is called the *objective function*, while (4.30) are the *inequality constraints*. Together, they form a so-called *constrained optimization problem*, which is treated by the introduction of the *Lagrangian*

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{m} \alpha_i(y_i(\langle\mathbf{x}_i,\mathbf{w}\rangle + b) - 1) \tag{4.31}$$

where $\alpha_i \geq 0$ are the *Lagrange multipliers* and L has to be minimized with respect to the *primal variables* $\mathbf{w}$ and b and maximized with respect to the *dual variables* α_i. In other words, a saddle point has to be found. As one might note, the constraints were incorporated into the second term of the Lagrangian.

Looking better to what the constrained optimization problems just stated, it is possible to verify that if the constraint (4.30) on the second term of the Lagrangian is violated with $\sum_{i=1}^{m} \alpha_i(y_i(\langle\mathbf{x}_i,\mathbf{w}\rangle + b) - 1 < 0$, L can be increased by increasing the corresponding α_i. Meanwhile, $\mathbf{w}$ and b will have to change (reducing the margin) such that L decreases. Provided that the problem is separable, at a given step, the constraint will be finally satisfied. Likewise, one can understand that for all constraints which are not precisely met as equalities, the corresponding α_i must

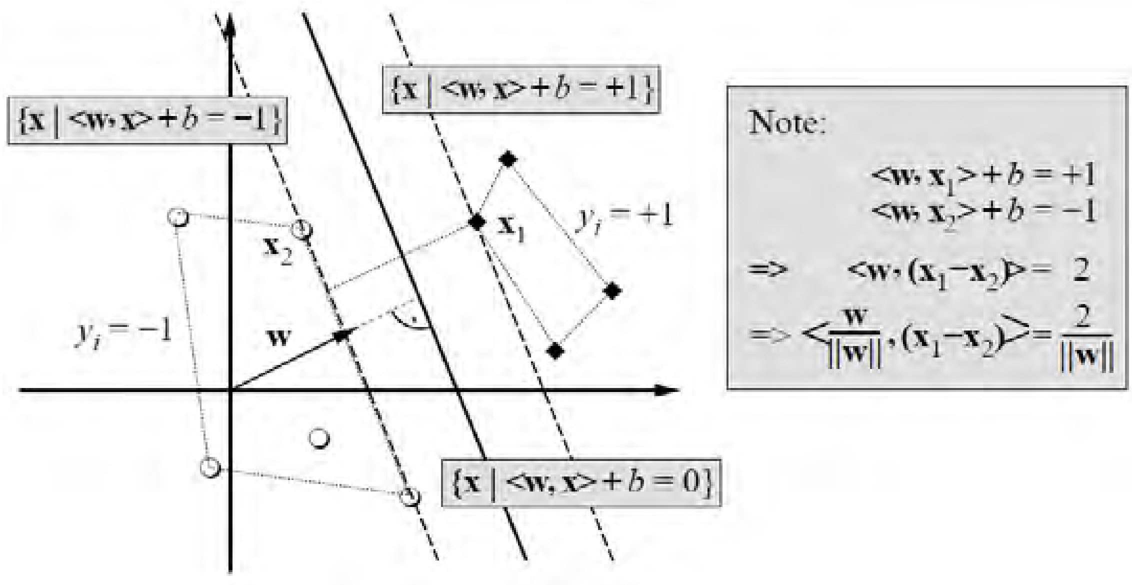


Figure 4.8: A binary classification problem whose purpose is to separate balls from diamonds. The optimal hyperplane (4.29) is shown as a solid line. The problem being separable: there exists a weight vector $\mathbf{w}$ and a threshold b such that $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) > 0$ $(i = 1, ..., m)$. Rescaling $\mathbf{w}$ and b in order that the point(s) closest to the hyperplane satisfy $|\langle \mathbf{w}, \mathbf{x}_i \rangle + b| = 1$, it is obtained the *canonical* form $(\mathbf{w}, b)$ of the hyperplane [26].

be 0: this is the value of α_i that maximizes L. The latter is the statement of the *Karush-Kuhn-Tucker* (KKT) complementarity conditions of optimization theory [26].

Considering that at the saddle point, the derivatives of L with respect to the primal variables must disappear,

$$\frac{\partial}{\partial b} L(\mathbf{w}, b, \alpha) = 0 \qquad \text{and} \qquad \frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, b, \alpha) = 0 \tag{4.32}$$

leads to

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \tag{4.33}$$

and

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \tag{4.34}$$

So, the solution vector has an expansion (4.34) in terms of a subset of the training patterns, namely those with non-zero α_i, which, in analogy with the idea in Section 4.3.1, are the Support Vectors.

By substituting (4.33) and (4.34) into the Lagrangian (4.31), one eliminates the primal variables **w** and **b**, arriving at the so-called *dual optimization problem*, which is the problem that SVMs usually solves in practice:

$$\underset{\alpha \in \mathcal{R}^m}{\text{maximize}} \quad W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \tag{4.35}$$

$$\text{subject to} \quad \alpha_i \geq 0 \quad \text{for all} \quad i = 1, ..., m \quad \text{and} \quad \sum_{i=1}^{m} \alpha_i y_i = 0 \tag{4.36}$$

and with the same (4.34), the hyperplane decision function (4.27) can be written as

$$f(x) = sgn\left(\sum_{i=1}^{m} y_i \alpha_i \langle \mathbf{x}, \mathbf{x}_i \rangle + b\right) \tag{4.37}$$

To complete this section, it is necessary to state how the remaining b can be calculated. According to the KKT conditions, all points with

$$\alpha_i [y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b) - 1] = 0 \quad \text{for all} \quad i = 1, ..., m \tag{4.38}$$

are SVs which lies on the margin and, therefore, $\alpha_i > 0$. In such cases, it can be shown that

$$\langle \mathbf{x}_i, \mathbf{w} \rangle + b = y_i \tag{4.39}$$

Thus, the threshold can for instance be obtained by averaging

$$b = y_i - \langle \mathbf{x}_i, \mathbf{w} \rangle \tag{4.40}$$

Alternatively, b can also be computed from the values of the corresponding dual variables α_i and α_j.

4.3.4 Support Vector Classification

Everything in the last section was formulated in a dot product space, as the feature space $\mathcal{H}$ stated in Section 4.3.1. To express the equations in terms of the input patterns in $\mathcal{X}$, it is necessary to employ (4.12), which denotes the dot product of the vectors $\mathbf{x}, \mathbf{x}'$ in terms of the kernel k evaluated on the inputs elements x, x',

$$k(x, x') = \langle \mathbf{x}, \mathbf{x}' \rangle \tag{4.41}$$

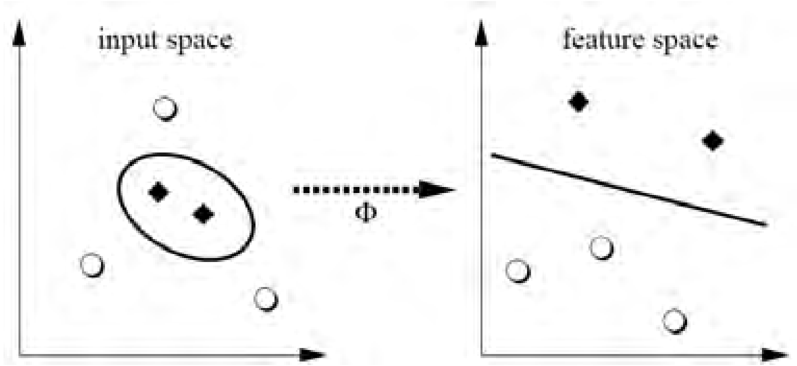


Figure 4.9: The idea of applying the kernel trick into SVMs: map the training data into a higher-dimensional feature space via Φ, and construct a separating hyperplane with maximum margin there. Despite the decision boundary being linear in the feature space, it is nonlinear in the original input space.

This replacement, also referred to as the *kernel trick* [26], is used to extend the concept of hyperplane classifiers to nonlinear support vector machines. It can be applied since all feature vectors only occurred in dot products, which can be noted in the equations (4.35) and (4.37). The weight vector (4.34) then becomes an expansion in feature space, and therefore will typically no more correspond to a vector of the input space. From (4.37), the decision function will then become

$$f(x) = sgn\left(\sum_{i=1}^{m} y_i\alpha_i\langle\Phi(x),\Phi(x_i)\rangle + b\right) = sgn\left(\sum_{i=1}^{m} y_i\alpha_i k(x,x_i) + b\right) \tag{4.42}$$

and the optimization problem (4.35) will take the form

$$\underset{\alpha\in\mathcal{R}^m}{\text{maximize}} \quad W(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j k(x_i,x_j) \tag{4.43}$$

subject to $\alpha_i \geq 0$ for all $i = 1,...,m$, and $\sum_{i=1}^{m}\alpha_i y_i = 0$. Figure 4.9 captures the idea of mapping data from the input space to the feature space.

Even with the advantage of "kernelizing" the problem, in practice, the separating hyperplane may still not exist, for example, if a high noise level causes a large overlap of the classes. To allow that some examples may violate equation 4.30, the slack variables are introduced [26, 43]

$$\xi \geq 0 \quad \text{for all} \quad i = 1,...,m \tag{4.44}$$

leading the constraints to

$$y_i(\langle \mathbf{w}, \mathbf{x_i} \rangle + b) \geq 1 - \xi_i \quad \text{for all} \quad i = 1, ..., m \tag{4.45}$$

A classifier that generalizes well is then found by controlling both the capacity (through $||\mathbf{w}||$) and the sum of the slacks variables $\sum_i \xi_i$. It can be shown that this summation provides an upper bound on the number of training errors.

In this context, a possible accomplishment of such a *soft margin* classifier is obtained by minimizing the objective function

$$\tau(\mathbf{w}, \xi) = \frac{1}{2}||\mathbf{w}||^2 + C \sum_{i=1}^{m} \xi_i \tag{4.46}$$

subject to the constraints (4.44) and (4.45), where the constant $C > 0$ determines the balance between margin maximization and training error minimization. The greater the C, the shorter the margin, and thus, smaller the training error and the capacity of the machine to generalize well.

Incorporating the kernel, and rewriting the problem in terms of the Lagrange multipliers, again leads to maximizing (4.43) subjected to the constraints

$$0 \leq \alpha_i \leq C \quad \text{for all} \quad i = 1, ..., m \quad \text{and} \quad \sum_{i=1}^{m} \alpha_i y_i = 0 \tag{4.47}$$

The only difference from the separable case is the upper bound C on the Lagrange multipliers α_i. Hence, the influence of the single patterns (which could be outliers) gets limited. The solution in (4.42) does not change and the threshold b can be computed by exploiting the fact that for all SVs x_i with $\alpha_i < C$, the slack variable ξ is zero. So,

$$\sum_{j=1}^{m} \alpha_j y_j k(x_i, x_j) + b = y_i \tag{4.48}$$

From the geometrical point of view, choosing b amounts to shifting the hyperplane. Concerning (4.48), such a shift is mandatory so that the SVs with zero slack variables lie on the ± 1 lines as shown in Figure 4.8.

The referred dot product kernel (4.41) is commonly called *Linear Kernel* and, given certain constraints [26], can be replaced by other functions.

Another well-known mapping kernel is the so-called *Gaussian Radial Basis Function* (RBF), which takes the form

$$k(x, x') = \exp\left(-\frac{||x - x'||^2}{2\sigma^2}\right) \tag{4.49}$$

where $\sigma > 0$ is the kernel parameter. Such a class of functions were also employed in the experiments.

At this point, all concepts about the type of SVMs employed to recognize faces in this thesis are covered. Due to the tunning variable C, this kind of SVM is normally referred to as C-SVC and represents SVM classification on its original form [44].

4.3.5 SVMs Applied to Face Recognition

Many researches with SVMs being used in some step of the face recognition problem were conducted in the last few years. Li and Jain [3] and Delac et al. [29] present a plenty of applications in face detection, pose estimation, expression recognition, and face identification. Two former studies that can be highlighted by applying SVMs for face recognition are provided by Phillips [5] and Guo et al. [39]. Since then, and following the remark about the growing importance of the area, many related works have been published [36, 37, 38, 40].

A fundamental point to be considered when formulating the framework of faces being recognized by SVMs is how to deal with the natural multi-class property of faces (where each individual represents a class) with a essentially binary classifier. To this problem, there also exists a range of possibilities [5, 26, 39, 36], which are presented below.

Identification as a Two Classes Problem

A typical face recognition algorithm treat each individual as a distinct class and, therefore, the distribution of its face is estimated or approximated. In such methods, for a gallery of k individuals, the identification problem is a k class problem.

Considering the traditional view-based approaches, two strategies for solving a k class problem with SVMs have been proposed [39, 36, 38]:

One Versus the Rest: In the one-versus-the-rest approach, k SVMs are trained. Each of the them separates a single class from all remaining classes.

Pairwise Classification: In the pairwise approach, $k(k-1)/2$ machines are trained. Each SVM separates a pair of classes. The pairwise classifiers are arranged in trees, where each tree node represents an SVM. Both top-down and bottom-up searching can be explored, with the later being similar to the elimination tree used in sports tournaments.

Regarding the training effort, the one-versus-the-rest has as a benefit the fact that only k SVMs have to be trained compared to $k(k-1)/2$ SVMs in the pairwise approach. The prediction complexity is similar: The one-versus-the-rest requires the evaluation of k SVMs, while the pairwise requires $k-1$. Since the number of SVMs to be trained is related with the number of classes to be recognized, both approaches would not cope with real scenario applications, where it is not desirable to train a SVM every time a new person is enrolled in the system.

To reduce face recognition to a single instance of a two class problem, Phillips [5] proposed a new face representation space, called Difference Space. By modeling the dissimilarities between faces, a k class problem can be transformed into a *within-class differences set* and a *between-class differences set* problem. This is a departure from traditional face space or view-based approaches, which encodes each facial image as a separate view of a face. Formally, let $T = t_1, ..., t_m$ be a training set of faces of k individuals, with multiple images of each of the k individuals. The within-class differences set C_1, which are the dissimilarities in facial images of the same person is

$$C_1 = \{t_i - t_j | t_i \sim t_j\} \tag{4.50}$$

and contains within-class differences for all k individuals in T. The between-class differences set C_2, which contains the dissimilarities among images of different individuals in the training set, is defined as

$$C_2 = \{t_i - t_j | t_i \nsim t_j\} \tag{4.51}$$

and classes C_1 and C_2 are then used as input to the SVM algorithm.

As showed in Section 4.3.4, in its pure paradigm SVM classifiers returns one of the two possible classes of an unknown test object. Thus, given the difference between facial images p_1 and p_2, the classifier estimates if they belong to the same person or not. This binary decision would considerably penalize the accuracy of the recognition, once the distance from a test sample to the separating hyperplane is reduced to a binary value.

In order to get a more sensitive similarity measurement, it is necessary to take of the signal function from (4.42), leading to

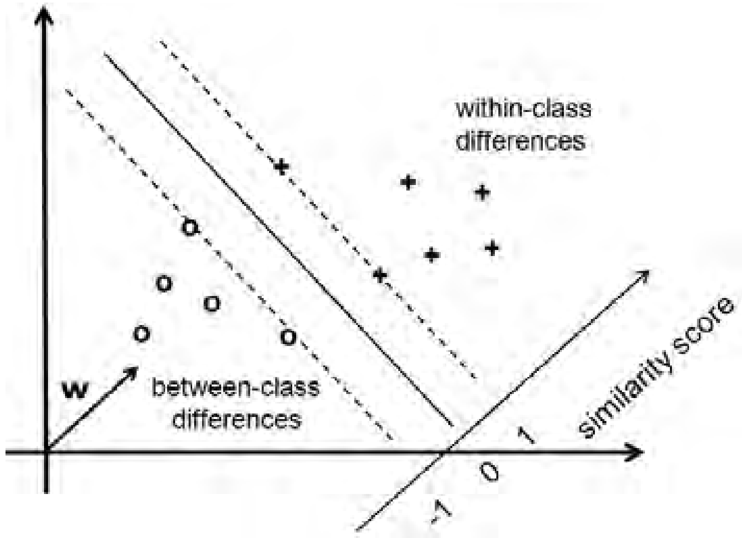


Figure 4.10: The notion of Support Vector Classification in a Face Differences Space. The similarity is reached through the distance a test pattern has to the separating hyperplane. The more inside a test pattern is with respect to the within-class set, the more reliable is the difference between such two faces (gallery and probe) from being of the same person.

$$f(x) = \sum_{i=1}^{m} y_i \alpha_i k(x, x_i) + b \tag{4.52}$$

This is a fundamental contribution introduced by Phillips in the adoption of SVMs for face recognition purposes. It gives the intuition of similarity as the distance a test pattern has to the separating hyperplane, leading to more reliable predictions. The higher the distance a test pattern belonging to the within-class set has to the hyperplane, the better. Figure 4.10 shows the idea of such face representation in a separating hyperplane problem.

In the adopted approach, just one SVM is necessary to handle the face recognition problem. From the moment a new person is enrolled in the system, its face-like gallery sample(s) will be used in pair with a probe to create *difference samples*, which in turn, will be used for the similarity assessment by the SVM.

Chapter 5

Face Databases

A face database plays a crucial role in the effective evaluation of a recognition algorithm. It must reflects the purpose from which the suggested technique arose from. In its previous time, when automatic face recognition had just become a reality, these databases expressed their challenges, with mostly collections representing constrained scenarios. However, as robustness against pose, illumination, age, occlusion, or expression started to be considered, new and carefully controlled variations were embedded to them [45].

The collection of a high quality database is a resource-intensive task. Furthermore, the size of the database has to be sufficiently large to cope with real scenario applications. However, despite the amount of work, the availability of public face databases is essential for the advancement of the field. Not only the specific needs of the algorithm under development, but also the increasingly interest in this area of research accounted for the large number of face databases available nowadays [45].

In this section four face databases are described. The first of them is the Notre Dame HumanID database [14]. It supplied the algorithms not only with good testings samples, but also with coherent images for their trainings. The other three databases are used in the experiments with SVMs. For the sake of comparisons and due to its broadly established protocol, the FERET database [46], with its respective gallery and probes sets, is presented. Additionally, the BANCA [47] and the VALID [9] are also introduced.

5.1 Notre Dame HumanID

The *University of Notre Dame* [14] has an important position in the field of infrared face recognition, providing collections of their biometrics databases to the research community. These collections are broken into components discriminated by the period and the sensing modality they were recorded.

Most of the data from the provided $X1$ collection was acquired at University of Notre Dame during 2002, with images from 241 distinct subjects. Each acquisition session consists of four views with different lighting and facial expressions in each sensor modality. For 82 subjects, image acquisitions were held weekly and most of them participated multiple times. The time-lapse records is a key feature of this database.

According to [14], the infrared images were acquired with a Merlin Uncooled long-wavelength IR camera, which provides a real-time, 60Hz, 12 bit digital data stream, has a resolution of 320×240 pixels, and is sensitive in the $7.0 - 14.0\mu m$. Visible-light images were taken by a Canon Powershot G2 digital camera with a resolution of 1200×1600 and 8 bit output.

Three Smith-Victor A120 lights with Sylvania Photo-ECA bulbs provided studio lighting. The lights were located approximately eight feet in front of the subject. One was approximately four feet to the left, one was centrally located, and one was located four feet to the right. All three lights were trained on the subject's face. The side lights and central light are about 6 feet and 7 feet high, respectively. One lighting configuration had the central light turned off and the others on. This is referred to as "FERET style lighting" or "LF". The other configuration has all three lights on, and is called "mugshot lighting" or "LM".

For each subject and illumination condition, two images were taken: one with neutral expression, which is called "FA", and the other image with a smiling expression, which is called "FB". For all of these images the subject was surrounded by a standard gray background. Since glass and plastic lenses are opaque in IR, subjects were asked to remove eyeglasses.

According to the lighting and expression, the collection provides four categories: (a) FA expression under LM lighting (FA|LM), (b) FB expression under LM lighting (FB|LM), (c) FA expression under LF lighting (FA|LF), and (d) FB expression under LF lighting (FB|LF). Figure 5.1 shows one subject in one session under these four conditions.

In order to create a larger training set for the experiments, 81 infrared and visible-light images of 81 distinct subjects acquired by Equinox Corporation [48] were also incorporated to the collection.

(a) FA|LF

(b) FB|LF

(c) FA|LM

(d) FB|LM

Figure 5.1: Visible and infrared images from the UND face database under the four different lighting and facial expression conditions. In the IR images, the gray level ranges from black (cold) to white (warm).

5.2 FERET

The Facial Recognition Technology (FERET) database [46] was collected at George Mason University and the US Army Research Laboratory facilities as part of the FERET program. The FERET, the Face Recognition Vendor Tests (FRVT) 2000, and independent evaluations used the database extensively [45], so detailed performance evaluations are available for many research algorithms as well as commercial face recognition systems. The lists of images used in training, gallery, and probe sets are distributed along with the database, so direct comparisons of recognizer performance with previously published results are possible.

The images were recorded in 15 sessions between August 1993 and July 1996. Each session lasted 1 or 2 days, and the location and setup did not change during the session. Because the recording equipment had to be reassembled for each session, slight variations between recording sessions are present. Images were recorded with a 35 mm camera, subsequently digitized, and then converted to 8-bit gray-scale images. The resulting images are 256×384 pixels in size.

Figure 5.2 shows frontal images from the five different sets of the database. The *fa* and *fb* image sets were obtained in close succession. The subjects were asked to display a different facial expression for the *fb* image. The resulting changes in facial expression are typically subtle, often switching between "neutral" and "smiling". The images in the *fc* category were recorded with a different camera and under different lighting. A number of subjects returned at a later date to be imaged again. For the images in the duplicate I set, 0 to 1031 days passed between recording sessions (median 72 days, mean 251 days). A subset of these images forms the duplicate II set, where at least 18 months separated the sessions (median 569 days, mean 627 days) [45].

(a) fa (b) fb (c) fc (d) DupI (e) DupII

Figure 5.2: The *fa* and *fb* were taken with the same lighting condition with different expressions, the *fc* image has a different lighting condition, and the *DupI, DupII* images were recorded in later sessions.

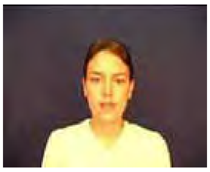
The database consists a total of 14,126 images and in late 2000 was entirely released to the research community. In its protocol, the gallery consists of one frontal image (fa) per person for 1196 individuals and the probe sets consists of (i) *fb probes*, (ii) *fc probes*, (iii) *DupI probes*, and (iv) *DupII probes* from 1195, 194, 722, and 234 individuals, correspondingly.

Ground-truth information, including the date of the recording and if the subject is wearing glasses, is provided for each image in the data set. In addition, the manually determined locations of left and right eye and the mouth center is available for a plenty of images.

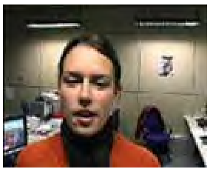
5.3 VALID

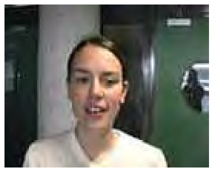
With the goal to facilitate the development of robust audio, face, and multi-modal person recognition systems, the large and realistic multi-modal (audio-visual) VALID database [9] was acquired in a noisy "real world" office scenario with no control on illumination or acoustic noise.

The database consists of five video recording sessions of 106 subjects over a period of one month. One session is recorded in a studio with controlled lighting and no background noise, the other 4 sessions are recorded in office type scenarios. The images have a resolution of 576×720 pixels and were extracted from the 1^{th}, 10^{th}, and the 50^{th} frames of each of the 5 sessions. Thus, the face database has $106 \times 5 \times 3 = 1590$ images. Figure 5.3 gives an example from each of the 5 sessions.


(a) 1


(b) 2


(c) 3


(d) 4

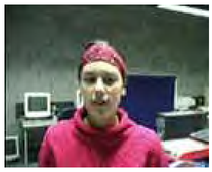

(e) 5

Figure 5.3: A VALID subject with still images taken from each of the five sessions.

5.4 BANCA

The BANCA multi-modal database [47] was collected as part of the European BANCA project, which aimed at developing and implementing a secure system with enhanced identification, authentication, and access control schemes for applications over the Internet.

The database was designed to test multimodal identity verification with various acquisition devices (high and low quality cameras and microphones) and under several scenarios (*controlled, degraded*, and *adverse*). Data were collected in four languages (English, French, Italian, Spanish) for 52 subjects each (26 men and 26 women).

Each subject was recorded during 12 different sessions over a period of 3 months. Recordings for a true client access (*Matched*) and an informed impostor attack (*Unmatched*) were taken during each session. For each recording the subject was instructed to speak a random 12-digit number along with name, address, and date of birth (client or impostor data). Recordings took an average of 20 seconds.

Figure 5.4 shows example images for all three recording conditions. The BANCA evaluation protocol specifies training and testing sets for a number of experimental configurations, thus, accurate comparisons between algorithms are possible.

(a) Controlled (b) Degraded (c) Adverse

Figure 5.4: Examples of the BANCA database images.

Chapter 6

Experiments and Results

Along Chapter 4, techniques for the multispectral fusion (PCA, Section 4.1) and techniques applied to the SVM experiments (CH in Section 4.2 and SVM in 4.3) were covered. This Chapter presents the results obtained with experiments carried out on the datasets described in Chapter 5 in order to assess the improvement that the proposed techniques can provide to the face recognition problem.

Regarding the typical system architecture presented in Section 2.3, the experiments are focused on extracting and matching face features, i.e., the third and the fourth modules of Figure 2.4, respectively.

6.1 Multispectral Fusion

The goal of the multispectral experiments was to overcome the recognition rates reached by the individual classifiers. To this end, the parallel fusion was adopted (Section 3.2.1) so that the visible and the thermal face images were concurrently used to converge onto a single classification response.

A total of 2.023 images in each spectrum from the UND face database were considered, with 187 subjects employed in the training phase and other 54 subjects selected for the gallery and the probe sets. This 54 subjects were the ones whose sessions attendance was bigger, with at least 7 and at most 10 weekly acquisitions. The first session of each subject was used in the gallery set, and the remaining 6 to 9 sessions constitutes the probe set. Hence, this experiment has two main characteristics to be highlighted:

1. The multisample gallery, composed of four images per subject in each spectrum. In both visible and infrared images, there were two samples with neutral and other two samples with smiling expressions;

2. The recognition over time, with probe images taken from 1 to 9 weeks after the gallery images. This addresses the major problem of duplicate recognition pointed out by the FERET and FRVT evaluations (Section 2.2).

Two different variations of the PCA face recognition method were assessed, namely: PCA with Euclidean distance and PCA with Mahalanobis Angle (Section 4.1), whose implementations are described in [33]. Each method was then applied individually in both spectra: Visible and IR.

The images were first converted to gray scale, followed by a geometrical normalization and an elliptical masking using the eyes as landmarks. The elliptical masking is aimed at cropping images such that information only from the forehead to the chin and from one cheek to the other remains. Figure 6.1 shows an example of (a) visible and (b) infrared images after the preprocessing.

(a) (b)

Figure 6.1: Visible and infrared preprocessed images from the UND face database.

Table 6.1 shows the obtained results of the individual modalities by means of TOP1 (Section 2.1.2) and EER (Section 2.1.3).

Despite the illumination invariance of the IR sensing approach, it is possible to observe that the method based on visible images has better performance in this controlled face database. However, this may not be the case in scenarios with adverse imaging conditions.

Table 6.1: The two face recognition methods and their respective individual performances on the visible and IR spectrum.

Method	Description	Spectrum	TOP1	EER
1	PCA Euclidian	IR	46.06	24.55
2	PCA Euclidian	Visible	87.92	14.11
3	PCA Mahalanobis	IR	87.74	8.87
4	PCA Mahalanobis	Visible	96.84	6.01

In order to predict the performance of the fusions, it was obtained the *Q-statistic* measure of dependency between the classifiers [49]. Given two classifiers i and k, the *Q-statistic* is

$$Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \tag{6.1}$$

where $\{N^{ab}|a = b\}$ denotes the number of cases which both classifiers agreed in their decisions and $\{N^{ab}|a \neq b\}$ denotes number of cases where the decision differs. The term N^{10} stands for the number of occurrences where a correctly matched a probe while b misclassified it. On the other hand, the term N^{01} represents the number of occurrences where a misclassified a probe while b correctly matched it. The Q-statistic ($Q_{i,k}$) measurement ranges from -1 to 1. For statistically independent methods, $Q_{i,k}$ is 0. For statistically correlated methods, $Q_{i,k}$ tends to 1, and for inversely correlated methods, $Q_{i,k}$ tends to -1.

Considering the difficulties of performing information fusion in early steps (sensor or feature level), and also taking into account the biometric systems' proprietary nature issue mentioned in Section 3.2.2, the proposed fusions were carried out in the score level. Based on the remarks of Jain et al. [21], instead of converting the matching score into probabilities, the experiments are done straightly with the similarity scores.

Three score normalization approaches were assessed (Section 3.2.3). Because of its simplicity, Min-Max normalization do not need additional explanations.

With respect to the double sigmoid normalization technique, the parameters were chosen following the outlines of [21]: t was the center of the overlapping region between the genuine and impostor score distribution, and $r1$ and $r2$ were made equal to the extent of the overlap toward the left and right of the center, respectively. In other words, a matching score that is equally likely to be from a genuine user and an impostor is chosen as the center (t) of the region of overlap. Then $r1$ is the difference between t and the minimum of genuine scores, while $r2$ is the difference between the maximum of impostor scores and t. In order to make this normalization robust, approximately 2% of the scores at the extreme tails of the distributions were omitted when calculating $r1$ and $r2$. Although this normalization scheme transforms all the scores to a common numerical range $[0, 1]$, it does not retain the shape of the original distributions.

Regarding the tanh-estimators normalization technique, the parameters were set up in such a way that the values of a, b and c represented 70%, 85%, and 95% of the scores in the intervals $(m - a, m + a)$,$(m - b, m + b)$, and $(m - c, m + c)$ respectively, with m being the median score.

With the scores from all classifiers in the same domain, the fusion of the face recognition methods were carried out by simple binary operations between their scores, namely: (i) sum , (ii) product, (iii) maximum, and (iv) minimum.

During the experiments, the double sigmoid score normalization approach and the product fusion technique showed to be more regular (better mean improvement) than the others. Therefore, they were chosen to denote the overall performance of the 6 possible fusions on Table 6.2.

Table 6.2: Overall performance from the fusions of methods X and Y as identified in Table 6.1. The results were obtained through the double sigmoid score normalization and the product fusion technique - better mean improvement. The best and worst results are highlighted.

X	Y	*Q Statistic*	TOP1 XY	%Improv.	EER XY	%Improv.
1	2	0.29	86.47	-1.66	**11.78**	16.51
1	3	0.85	**84.41**	**-3.80**	10.32	**-16.26**
1	4	0.24	97.21	0.38	5.17	13.87
2	3	0.22	95.63	**8.76**	4.84	45.41
2	4	0.95	96.54	-0.31	5.92	1.48
3	4	0.14	**98.85**	2.07	**3.28**	**45.47**

As expected, it can be observed that the correlation between methods applied on different spectra is much smaller than the correlation of methods applied on the same spectrum, which indicates that they hit and fail in different situations for many probes. For instance, the Q-statistic measurement for methods 3 and 4 was 0.14, the lowest for different spectra methods, while for methods 1 and 3 was 0.85, the lowest (but much higher) between the fusions on the same spectrum.

It can also be observed in Tables 6.1 and 6.2 that there is a relationship among the Q-Statistic, the TOP1 individual rates, and the performance of the fusion. When the Q-Statistic is low (lower than 0.5) and the TOP1 individual rates are high (greater than 50%), the performance of the fusion compared with the best individual rates always increases. Hence, the overall best performance (98.85% for TOP1, and EER=3.28) was obtained with the fusion of two good individual methods, 3 and 4, that present the lowest correlation rate (0.14).

6.2 Support Vector Machines

The face recognition experiments using Support Vector Machines were aimed at improving recognition rates achieved by classifiers which take the CH features as inputs. Moreover, by estimating performance according to the FERET *fa/fb* protocol, the intention was to compare the proposed method with results available in the literature.

The investigation was conducted in collaboration with the Cognitive System department at the Fraunhofer Institute [1], which provided the Census Histograms algorithm applied in the experiments as well as the image normalization techniques. To the best of our knowledge, there exist no published study regarding face recognition with Census Histograms features. Therefore, the performance target to be improved by the SVM was also provided by the Institute. The method which achieved the best recognition performance (target) consists on the summation of the CH features as an assembled L1 distance. Hence, the L1 benchmark is provided in all comparative tables as our baseline. Because it does not take into account a previous learning process, its training time is 0.

The basis SVM implementation was provided by Chang and Lin [50]. The algorithm used in the experiments takes equation (4.46) as its objective function, and equation (4.45) as constraints. The decision fusion, however, refers to the modified equation (4.52), a key issue of the experiments.

In order to summarize, from the face images to the trained SVMs the steps can be enumerated as follows:

1. Geometrical normalization and cutting based on previously annotated positions of the eyes;

[1] http://www.iis.fraunhofer.de/EN/bf/bv/kognitiv/index.jsp

2. Selection of matching pairs to build up the positive dataset - the referred within-class differences set (Section 4.3.5);

3. Random selection of unmatching pairs to build up the negative dataset - the referred between-class differences set (Section 4.3.5);

4. Census transformation of all selected images together with the Census Histogram computation (Section 4.2);

5. Support Vector Machine and kernel function parameters grid searching as described by Chang and Lin [50];

6. Support Vector Machine classifier training.

To evaluate the trained classifiers, the just mentioned steps 1, 2, 3, and 4 are carried out in the FERET database. Afterwards, the resulting pool of within- and between-class samples is submitted to the SVMs predictions. For each FERET probe sample, one-to-many predictions are performed according to the closed-set benchmark (Section 2.1.1).

Regarding the photometrical normalization, the images were just converted to gray scale, but no illumination processing was considered due to the lightning independence of the Census Transform.

The VALID database, as well as the BANCA database, were employed especially for the SVMs learning process. To avoid very large training sets, which would lead to slow and less effective training stages, just part of these databases were used in the experiments. The results are then expressed by means of the training sets:

VALID: From the VALID database, 4 images per subject were chosen. It was spontaneously decided to take the 1^{th} frame from session 1, the 10^{th} frame from session 2, the 50^{th} frame from session 3, and the 1^{th} frame from session 4. The 4 images from each of the 106 subject combined 2 by 2 leads to a training set with 636 positive samples;

BANCA: With respect to the BANCA database, the experiments took into account the English part of the database from where 5 images per subject in the Matched and Controlled (MC) scenario were selected. To establish the size of this training set, again it is necessary a 2 by 2 combination of the 5 images per each of the 52 subject. Therefore, each individual provides $(5 \times 4)/2 = 10$ positive samples, which leads to a training set of 520 positive samples;

BANCA20: Like BANCA, but considering only the first 20 individuals, i.e., 200 positive samples.

These groups of images formulate the within-class differences sets C_1 defined in Section 4.3.5. Furthermore, the negative samples sets C_2 (between-class differences) were generated, in each case, from its respective database. In this case, the difference space refers to the Census Histograms features from pairs of randomly chosen images. The random selection insures that each pair of faces belongs to distinct persons, and the number of negative samples (which would be very large) is restricted to the number of positive samples.

Most of the investigation was conducted by submitting all Census Histograms features to the SVM optimization problem. However, in some experiments, the AdaBoost learning algorithm [8, 27, 28] was applied to retain only the most discriminant features. Thereof, optimal position/size of scanning windows were chosen by the boosting approach. For the sake of our scope, the AdaBoost approach adopted in this thesis is briefly mentioned in Section 3.3.3 due to it does not belong to the core of the experiments.

With the training sets in mind, the next description refers to the types of SVM. Two attributes have to be considered to distinguish between the possible SVM setups: (i) the kernel functions employed, and (ii) whether the boosting approach was utilized for feature selection. A *Plain* SVM performs no feature selection, while a *Boosted* SVM selects optimal features prior the SVM training. A *Linear Kernel* SVM uses equation (4.41) as its kernel function, whereas a *Radial Basis Function* SVM uses (4.49). The variations between these attributes leads to the following combinations:

PSLK: Plain SVM with Linear Kernel;

PSRK: Plain SVM with RBF Kernel;

BSLK: Boosted SVM with Linear Kernel;

BSRK: Boosted SVM with RBF Kernel.

The FERET database (Section 5.2) was adopted in the evaluations, having the fa (frontal) images as gallery and the fb (different expression) images as probes. This is commonly called a FERET *fa/fb* test.

The performance was measured in terms of TOP1 and full retrieval (Section 2.1.2). In addition, time in seconds for the training and testing processes was assessed. Tables 6.3, 6.4, and 6.5 show the results.

Table 6.3: Evaluation performances for the VALID training set. The performance was measured in terms of TOP1 and full retrieval (Section 2.1.2).

	Rates		Time Consuming (seconds)	
Method	TOP1	Full Retrieval	Training	Evaluating
Baseline L1	97.2	197	0	299
PSLK	95.8	407	2465	1407
PSRK	44.3	321	578	1292
BSLK	87.4	1025	1653	382
BSRK	30.6	391	286	602

With respect to the following two evaluations, the parameters of the CH scanning window were changed and many rounds of benchmarking were performed in order to better explore CH's potential. Thus, each method is represented by its best performance along the different CH setups. As the scanning window changed, the number of CH features also changed. In order to provide additional information about the training/evaluation time consuming, the column "number of features" was added to the tables. Note that Table 6.3 do not presents the number of features because for the experiments with the VALID database, all evaluations were done with the same CH scanning window setup.

Table 6.4: Evaluation performances for the BANCA20 training set. The performance was measured in terms of TOP1 and full retrieval (Section 2.1.2).

	Rates		Time Consuming (seconds)		
Method	TOP1	Full Retrieval	Training	Evaluating	# Features
Baseline L1	97.2	197	0	299	154
PSLK	97.7	195	15	9138	2268
PSRK	95.2	456	30	331	154
BSLK	97.2	132	54	233	54
BSRK	97.4	240	65	330	60

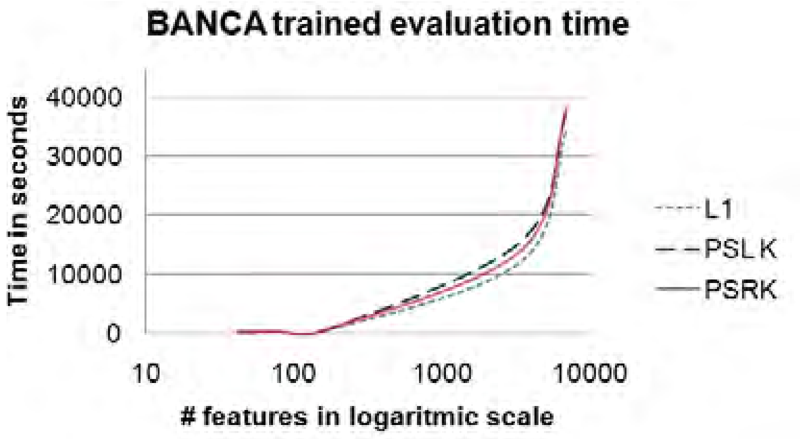
The first observation that can be done about these results is that the recognition based on RBF kernels is not superior to that based on linear kernel. In the best situation, it equals. Regarding the baseline L1 performance with a TOP1 of 97.2%, it is also possible to state that, thanks to the descriptiveness of CH features, the data in its input feature space is almost linearly separable and this could be the reason of the linear kernel superiority.

Table 6.5: Evaluation performances for the BANCA training set. The performance was measured in terms of TOP1 and full retrieval (Section 2.1.2).

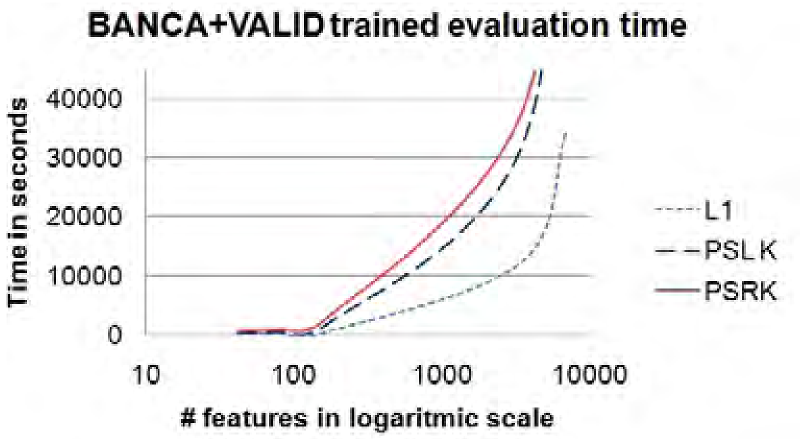
	Rates		Time Consuming (seconds)		
Method	TOP1	Full Retrieval	Training	Evaluating	# Features
Baseline L1	97.2	197	0	299	154
PSLK	97.6	503	37	10018	2016
PSRK	97.2	791	13432	38327	6930
BSLK	97.7	273	109	318	72
BSRK	97.7	218	164	425	72

With respect to the training databases, one can observe a significant improvement on the BANCA-trained performance compared to the VALID-trained SVM. Taking into account the fact that the VALID training set was built up from images of uncontrolled office type scenarios and the BANCA training set was based in a subset of constrained images, it is reasonable to remark in our experiments one of the "learning from data" fundamentals, which states that the closer the training data to the real testing data, the better. This is exactly what happened considering the similarity between the BANCA images and the FERET fa/fb datasets, and it also explains the poor performance of the VALID-trained RBF methods showed in Table 6.3.

Another clear attribute of the assessed SVM is the link between the number of features and the evaluation time consuming. As the predictions are based on linear combinations of support vectors, the input space dimensionality is strictly related to the testing time performance, which in some cases becomes prohibitive. Such an observation is better illustrated in Figure 6.2, where the number of features in the input space is plotted against the benchmark time in seconds. The difference in the charts is related to the training set complexity. In (a), it is possible to note a straight link between the input space dimension and the evaluation time consuming. In addition, chart (b) shows that not only the input space dimensionality, but also the training set size and complexity (nonlinearity) considerably affects the evaluation. The BANCA+VALID is merely the concatenation of the BANCA and the VALID sets, which leads to a more intricate training problem with 1156 positive and 1156 negative samples. Therefore, the optimization of such a training set requires much more support vectors to discriminate the separating hyperplane.



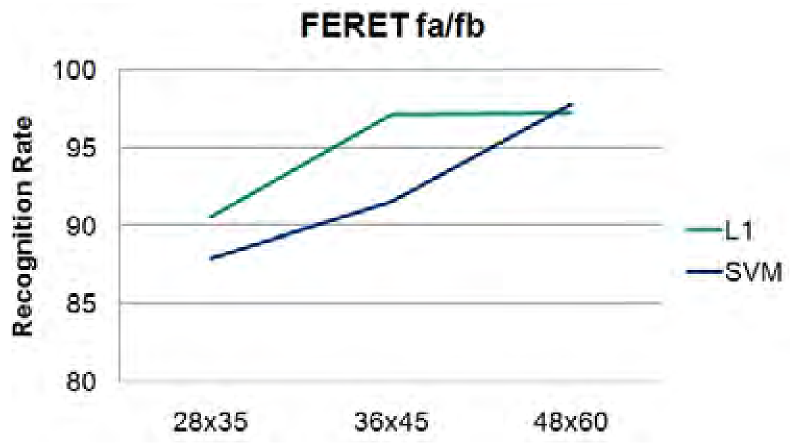
(a)



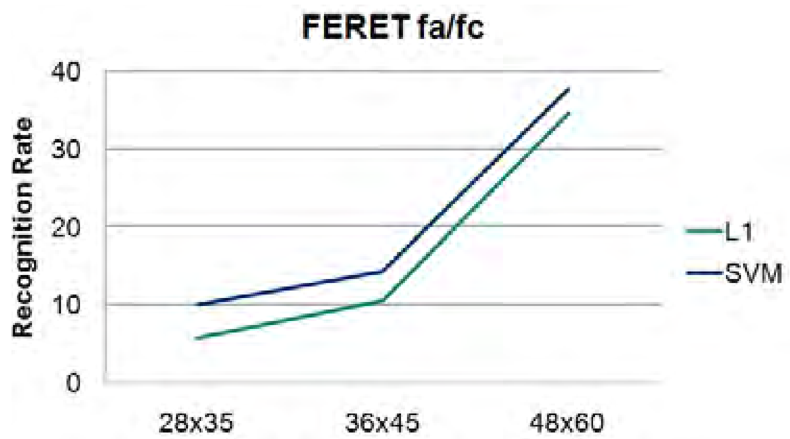
(b)

Figure 6.2: Comparison between the amount of features and the SVM evaluation time consuming. A straight relation between the input space dimension and the evaluation time consuming can be noted in (a), while (b) shows that not only the input space dimensionality but also the training set size and complexity (nonlinearity) considerably affects the evaluation.

Along Chapter 2, one of the highlighted robustness that a face recognition algorithm must have was pose invariance. Such an independence is not only related to the viewpoint, but also to the face scale due to different image capture distances. Therefore, another group of SVM experiments is based on rescaling the training and testing sets in three different setups to assess recognition accuracy with image size increasing. In this analysis, not only the FERET fa/fb protocol was evaluated, but also the FERET fa/fc (Section 5). Figure 6.3 (a) shows the expected correlation between size and recognition performance increasing, where the SVM algorithm just surpassed the baseline L1 with resolutions bigger then 48×60. Once again, the linearity on the Census Histograms FERET fa/fb representation is believed to be in favor of L1. On the other hand, the FERET fc dataset was recorded with different cameras and under different light conditions, leading to a more complex recognition problem. The L1 and SVM performances were also compared in such a case, and are illustrated in Figure 6.3 (b). As one might note in this case, SVM overcame L1 in all three different image size setups, which attests its advantage dealing with less linear and nonconvex data.



(a)



(b)

Figure 6.3: Correlation between face image sizes and recognition performance. As expected, as the image size increases, recognition accuracy also increases. (a) Due to linearity in FERET fa/fb, the SVM algorithm just surpassed the baseline L1 with resolutions bigger then 48×60. (b) However, in the more complex FERET fa/fc problem, SVM overcame L1 in all different image size setups.

As presented, the results are in terms of two distinct kernel functions. Both linear and RBF functions reached similar results in most cases. Hence, it was not possible to conclude about which type of function is preferred for face recognition purposes. Such an assertion is also in concordance with the previous narrowly related work [40], whose conclusion stated this issue as open problem up to now.

At last, it can not be discarded the intuition that if the problem was more intricate, then SVM could lead to even better results. Moreover, the yielded performance of 97.7% in FERET fa/fb (Tables 6.4 and 6.5) can be considered competitive compared to results available in the literature [27].

Chapter 7

Conclusion

As emphasized in the beginning, this master thesis was aimed at improving current face recognition performance rates by investigating the potential advantages of three different subjects related to this research field.

The multispectral fusion enabled us to experiment the first two of these aspects: new sensor modalities and multibiometrics. The results obtained with the multispectral fusion experiments suggest that combining two good and not correlated classifiers leads to significant improvement, both on performance (TOP1) and on system errors (EER). This was the case of all multispectral fusions described in Section 6.1. In the best case, this approach enabled an improvement of 2.07% in the recognition rate while reduced the EER in 45.47%. On the other hand, if the fusion is carried out with highly correlated methods, even both presenting good individual rates, the performance may decrease. Another observation that can be done is that, if one of the classifiers present a prior low performance, the performance might decrease regardless the correlation.

The fusion of IR and visible images for face recognition has been studied in the past years, as described along this thesis. Nevertheless, to the best of our knowledge, such comparisons about the recognition results and the levels of independence between the methods were not carried out up to now. Therefore, this is one of the contributions of this thesis.

The third highlighted aspect to overcome face recognition deficiencies consists on the application of promising machine learning techniques to judge the similarity between faces. To this end, the framework which combines Census Histograms features with Support Vector Machines was proposed.

Concerning the results obtained with this technique, it can be concluded that the approach was effective in recognizing faces, with 97.7% of the FERET fa/fb face images correctly identified. In addition, this study enabled a deeper understanding about the SVM behavior when dealing with problems of different size, convexity, and dimensionality. The fundamental contribution introduced by Phillips [5] regarding the intuition of similarity as the distance between the test patterns and the separating hyperplane was successfully replicated in this thesis. The results attested the viability of a single C-SVC SVM in charge of a inherently multi-class problem. However, no conclusion could be done about which type of kernel function is better for the face recognition problem.

Still on SVM, a topic which deserves further clarification is its computation burden. To tune the optimization and the kernel function parameters, a number of trainings in accordance to the tuning setup will have to be conducted. If the problem being optimized already demands substantial time, the optimal parameters' searching may become prohibitive. Once trained, however, the prediction cost is related with the number and the dimensionality of the Support Vectors, and in most cases did not use to be a shortcoming. What really takes time in some situations is the training, which can be discarded in the classification operational scenario.

A broad, but not new, conclusion of this master thesis is that most of automatic face recognition techniques works well enough for cooperative frontal views without large expressions changes and with controlled illumination. Nevertheless, for an unconstrained daily life application, without cooperation from the users, such as to recognize someone in the street at a distance, face recognition is still a challenging task.

Chapter 8

Future Work

An interesting future work with respect to the IR face imaging technology is the extraction of thermal minutia points from the branches of the face vascular network under the skin. The sensitiveness of current IR sensors together with recently proposed image processing techniques enable the extraction of these minutiae [17]. This physiological information capitalizes on the permanency of innate characteristics that are under the skin and can be used as features to reliable predict face identities. Considering the effectiveness of minutia-based biometrics, e.g. fingerprints, this is a quite promising subject to be investigated.

With respect to the SVM experiments, an interesting future investigation would be the employment of more complex face databases both for training and evaluating. As the Census Histograms showed to be a good face representation technique, the hope is that, in such scenario, the kernelized SVMs will present an even higher superiority compared to linear classification methods.

The employment of CH features in the multispectral fusion experiments is also an additional planned work, as well as the possibility to compare SVMs to other promising recognition approaches like Optimum-Path Forest [51].

References

[1] W. Zhao, R. Chellapa, P. J. Phillips, and A. Rosenfeld. Face Recognition: A Literature Survey. *ACM Computing Surveys*, 35(4):399–458, December 2003.

[2] S. G. Kong et. al. Recent advances in visual and infrared face recognition - a review. *Computer Vision and Image Understanding 97*, pages 103–135, 2005.

[3] S. Z. Li and A. K. Jain. *Handbook of Face Recognition.* Springer, 2004.

[4] D. T. Jarvis and B. K. Cromwell. Thermography and Radiometry in the Range Community. Technical report, FLIR Systems, Inc., December 2008. http://www.rangerats.org/therm.html.

[5] P. J. Phillips. Support Vector Machines Applied to Face Recognition. *Advances in Neural Information Processing Systems 11*, (11):803–809, 1999.

[6] R. Zabih and J. Woodfill. A Non-Parametric Approach to Visual Correspondence. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1996.

[7] A. K. Jain, A. Ross, and S. Prabhakar. An Introduction to Biometric Recognition. volume 14, January 2004.

[8] C. Kublbeck and A. Ernst. Face Detection and Tracking in Video Sequences using the Modified CensusTransformation. *IVC*, 24(6):564–572, June 2006.

[9] N. A. Fox, B. A. O'Mullane, and R. B. Reilly. VALID: A New Practical Audio-Visual Database, and Comparative Results. In *Int. Conf. on Audio- and Video-based Biometric Person Authentication*, 2005.

[10] P. J. Phillips, P. J. Flynn, and T. Scruggs. Preliminary Face Recognition Grand Challenge Results. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pages 15–24, 2006.

[11] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 Large-Scale Results. Technical Report NISTIR 7408, National Institute of Standards and Technology, 2007.

[12] X. Chen, P. J. Flynn, and K. W. Bowyer. IR and Visible Light Face Recognition. *Computer Vision and Image Understanding 99*, page 332358, 2005.

[13] K. W. Bowyer, K. I. Chang, P. J. Flynn, and X. Chen. Face Recognition Using 2-D, 3-D, and Infrared: Is Multimodal Better Than Multisample? In *Proceedings of the IEEE*, volume 94, pages 2000–2012, November 2006.

[14] X. Chen, P. J. Flynn, and K. W. Bowyer. Visible-light and Infrared Face Recognition. *ACM Workshop on Multimodal User Authentication*, pages 48–55, December 2003.

[15] X. Lu and A. K. Jain. Deformation Modeling for Robust 3D Face Matching. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2006)*, pages 1377–1383, August 2006.

[16] Simple3D. 3D Scanners, Digitizers, and Software for making 3D Models and 3D Measurements. http://www.simple3d.com, December 2008.

[17] P. Buddharaju, I. T. Pavlidis, P. Tsiamyrtzis, and M. Bazakos. Physiology-Based Face Recognition in the Thermal Infrared Spectrum. *IEEE Trans. Pattern analysis and Machine Intelligence*, 29(4):613–626, April 2007.

[18] D. Socolinsky and A. Selinger. A comparative analysis of face recognition performance with visible and thermal infrared imagery. *Proceedgins of International Conference on Pattern Recognition*, pages 217–222, 2002.

[19] S. G. Kong, J. Heo, F. Boughorbel, Y. Zheng, B. R. Abidi, A. Koschan, M. Yi, and M. A. Abidi. Multiscale Fusion of Visual and Thermal IR Images for Illumination-Invariant Face Recognition. *Int. Journal of Computer Vision, Special Issue on Obj. Tracking and Classif. Beyond the Visible Spectrum*, 71(2):215–233, February 2007.

[20] A. A. Ross, K. Nandakumar, and A. K. Jain. *Handbook of Multibiometrics*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[21] A. K. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition 38*, pages 2270–2285, 2005.

[22] C. Sanderson and K. K. Paliwal. On the Use of Speech and Face Information for Identity Verification. *IDIAP Research Report 04-10*, March 2004.

[23] A. Ross and A. Jain. Information fusion in biometrics. *Pattern Recognition Letters*, (24):2115–2125, 2003.

[24] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, pages 71–86, 1991.

[25] A. M. Martinez and A. C. Kak. PCA versus LDA. *Pattern analysis and Machine Intelligence*, 23:228–233, 2001.

[26] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[27] G. Zhang, X. Huang an S. T. Li, Y. Wang, and X. Wu. Boosting Local Binary Pattern (LBP)-Based Face Recongnition. *Advances in Biometric Person Authentication*, pages 179–186, 2005.

[28] B. Fröba and A. Ernst. Face Detection with the Modified Census Transform. In *Int. Conf. on Automatic Face and Gesture Recognition*. IEEE, 2004.

[29] K. Delac, M. Grgic, and M. S. Bartlett. *Recent Advances in Face Recognition*. In-Teh, 2008.

[30] A. R. Calvo, F. T. Ruiz, J. Rurainsky, and P. Eisert. *2D-3D Mixed Face Recognition Schemes*, chapter 10. In-Teh, 2008.

[31] S. Romdhani, V. Blanz, C. Basso, and T. Vetter. *Morphable Models of Faces*, chapter 10, pages 217–245. Springer, 2004.

[32] G. Shakhnarovich and B. Moghaddam. *Face Recognition in Subspaces*, chapter 7, pages 141–168. Springer, 2004.

[33] D. Bolme, R. Beveridge, M. Teixeira, and B. Draper. The CSU Face Identification Evaluation System: Its Purpose, Features and Structure. In *International Conference on Vision Systems*, pages 304–311, Graz, Austria, April 2003.

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

[35] B. Dasarathy. *Nearest Neighbor (NN) Norms-NN Pattern Classification Techniques*. IEEE CS Press, 1991.

[36] H. Qiao, S. Zhang, B. Zhang, and J. Keane. Face recognition using SVM decomposition methods. In *Int. Conf. on Intelligent Robots and Systems*, volume 2, pages 2015–2020, September 2004.

[37] X. He, J. Tian, Y. He, and X. Yang. Face Recognition with Relative Difference Space and SVM. *The 18th International Conference on Pattern Recognition*, (0-7695-2521-0/06), 2006.

[38] B. Heisele, Y P. Ho, and T. Poggio. Face recognition with support vector machines: global versus component-based approach. In *In Proc. 8th International Conference on Computer Vision*, pages 688–694, 2001.

[39] G. Guo, S. Z. Li, and K. Chan. Face Recognition by Support Vector Machines. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 196–201, 2000.

[40] W. CAO, L. LI, and X. LV. Kernel Function Characteristic Analysis Based on Support Vector Machine in Face Recognition. In *Sixth International Conference on Machine Learning and Cybernetics*, pages 19–22, Hong Kong, August 2007.

[41] Y. Rodriguez. *Face Detection and Verification using Local Binary Patterns*. IDIAP-RR, École Polytechnique Fédérale de Lausanne, 2006.

[42] V. N. Vapnik. An Overview of Statistical Learning Theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999.

[43] B. Schölkopf, R. C. Williamson, and P. L. Bartlett. New Support Vector Algorithms. In *Neural Computation*, pages 1207–1245. MIT, 2000.

[44] C. Cortes and V. Vapnik. Support-Vector Networks. In *Machine Learning*, pages 273–297. Kluwer Academic Publishers, 1995.

[45] Ralph Gross. *Face Databases*, chapter 13, pages 301–327. Springer, 2004.

[46] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi. The FERET Evaluation Methodology for Face-Recognition Algorithms. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 22, October 2000.

[47] E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J. Thiran. *The BANCA Database and Evaluation Protocol*, pages 625–638. Springer, 2003.

[48] Equinox Corporation. Human Identification at a Distance Database, 2009. http://www.equinoxsensors.com/products/HID.html.

[49] L. I. Kuncheva and C.J. Whitaker. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*, pages 181–207, 2003.

[50] C. C. Chang and C. J. Lin. *LIBSVM: a Library for Support Vector Machine*, October 2008.

[51] J.P. Papa, A.X. Falcão, C.T.N. Suzuki, and N.D.A. Mascarenhas. A discrete approach for supervised pattern recognition. In *12th International Workshop on Combinatorial Image Analysis*, volume 4958, pages 136–147. Springer Berlin/Heidelberg, 2008.

Appendix A

Related Publications

The following publications are related to this thesis:

I. **Chiachia, G.** ; Marana, A. N. . Fusion of Infrared and Visible Spectra Face Recognition Methods. *In: Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI) - Technical Posters*, p. 11-12, Campo Grande, 2008.

II. **Chiachia, G.** ; Penteado, B. E. ; Marana, A. N. . Fusão de Métodos de Reconhecimento Facial através da Otimização por Enxame de Partículas. *In: Workshop de Visão Computacional (WVC)*, Canal6 Editora, Bauru, 2008.

III. **Chiachia, G.** ; Marana, A. N. ; Papa, J. P. ; Falcão, A. X. . Infrared Face Recognition by Optimum-Path Forest. *In: IWSSIP 2009 - International Workshop on Systems, Signals and Image Processing (accepted)*, Chalkida - Greece, 2009.

IV. Marana, A. N. ; Papa, J. P. ; **Chiachia, G.** . Introdução ao Reconhecimento de Padrões. *In: DINCON'09 - 8th Brazilian Conference on Dynamics, Control and Applications*, v. 8, Bauru, 2009.