ATTENTION–DRIVEN PATTERN RECOGNITION

THÈSE N° 2153 (2000)

PRÉSENTÉE AU DÉPARTEMENT D'INFORMATIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

FABRIZIO SMERALDI

Laurea in Fisica, Università di Genova, Italie
de nationalité italienne

acceptée sur proposition du jury:

Prof. R. Hersch, président
Prof. W. Gerstner, directeur de thèse
Dr. S. Fischer, rapporteur
Prof. M. Unser, rapporteur
Dr. A. Verri, rapporteur
Prof. T. Vetter, rapporteur

Lausanne, EPFL
2000

Abstract

Computer Vision systems deal with an extremely rich and complex representation of the external world encoded in the form of static images or video streams. In order to make sense of the huge amount visual input it receives, a system must be able to identify and process the most informative and interesting parts of it. In other words, an *attentional mechanism* is needed that selects task-relevant information.

A natural source of inspiration for attentional mechanisms is given by the human visual system. The present Thesis aims at introducing a new biologically inspired vision paradigm, which we refer to as *Retinal Vision*. This is intended as a general framework that pursues the integration of low-level biologically inspired signal conditioning such as retinotopic sampling and the Gabor decomposition with models of higher-order visual processes that constitute the interface to cognitive processes in human beings, with particular reference to the saccadic system. The underlying idea is that of implementing attentional mechanisms that would allow a cognitive task to steer visual acquisition and processing.

Humans and primates do not explore a visual scene in a raster-like fashion. They rather perform large jumps, known as *saccades*, between the points of interest in the scene, which are *fixated* for a fraction of a second. Saccades are known to play a role in cognitive processes, as the saccadic pattern depends both on the visual scene and on the task to be performed.

In this Thesis we introduce a *Saccadic Search* strategy that allows an efficient detection of deformable objects in static images or video streams. A description of the visual scene is constructed by computing the responses of a set of modified Gabor filters on the sparse nodes of a Log-polar retinotopic graph. A priori knowledge about the search targets is included in the form of appearance-based models, that are implemented by Support Vector Machine classifiers. These models are employed to generate a sequence of saccades that eventually centres the retinotopic grid on the patterns of interest. The process does not require the computation of the Gabor decomposition over the entire image; the resulting increase in efficiency makes such a powerful tool as the Gabor decomposition suitable for an active vision scenario, despite the computationally demanding nature of Gabor features.

We introduce the *Retinal Vision* paradigm through a series of applications related to the *Face Authentication* problem, that combines the necessary level of generality and technical complexity with an evident scientific and practical relevance. As a first applica-

tion of the Saccadic Search strategy we present a *Real-time Head Detection and Tracking* system. To the best of our knowledge, this represents the first example of the use of the Gabor decomposition in such an active vision task. The second application we consider, *Facial Feature Detection*, demonstrates the suitability of our approach for a complex pattern recognition problem on static images.

The descriptiveness of the image representation, which is based on the computation of the responses of modified Gabor filters at the nodes of a sparse Log-polar graph, has been assessed applying it to *Face Authentication*. A consistently good performance is reported over the two largest authentication-oriented image databases available.

A separate Chapter is dedicated to the study of the dimensionality of the Face Authentication problem, a subject on which widely contrasting estimates can be found in the literature. Classical compression techniques such as Principal Components Analysis and Linear Discriminant Analysis are considered together with a newly developed algorithm, Support Vector Features. These techniques, in combination with both traditional and modern classifiers such as K Nearest Neighbours and Support Vector Machines, are used to investigate the effect of the dimensionality of the feature space on authentication performance.

Sommario

I sistemi di visione artificiale hanno accesso ad una rappresentazione estremamente ricca e complessa del mondo esterno, sotto forma di immagini statiche o di sequenze video. Per poter interpretare l'enorme quantità di input visivo che riceve, un sistema deve essere in grado di identificarne le parti più rilevanti e più ricche di informazione. In altri termini, si rende necessario un *meccanismo attenzionale* che selezioni specificamente l'informazione pertinente al problema da risolvere.

Il sistema visivo umano costituisce naturalmente una fonte privilegiata di ispirazione per lo sviluppo di meccanismi attenzionali. In quest'ottica, la presente Tesi introduce un nuovo paradigma di visione, che chiameremo *Retinal Vision*. Questo è inteso come un quadro generale nel cui ambito ricercare l'integrazione di un trattamento di basso livello del segnale di matrice biologica (campionamento non uniforme, filtri di Gabor) con modelli dei processi visivi più avanzati che costituiscono, negli esseri umani, l'interfaccia verso i processi cognitivi. Lo scopo finale è lo sviluppo di meccanismi attenzionali che permettano ad un processo cognitivo di controllare l'acquisizione ed il trattamento dei dati visivi. In particolare, questo lavoro trae ispirazione dai movimenti saccadici oculari.

Gli esseri umani, come del resto i primati, non analizzano un'immagine percorrendola sistematicamente da un lato all'altro come il pennello elettronico di un televisore. Al contrario, il percorso seguito dallo sguardo su una scena si compone di ampi balzi tra i diversi punti di interesse, ciascuno dei quali viene fissato per una frazione di secondo. Questi bruschi spostamenti dello sguardo sono noti come *movimenti saccadici oculari*. Dal fatto che il percorso dello sguardo sull'immagine dipende non solo dall'immagine osservata, ma anche dal fine che il soggetto si propone nell'osservarla, si evince che i movimenti saccadici svolgono un ruolo nei processi cognitivi stessi.

In questa Tesi introduciamo un meccanismo di *Ricerca Saccadica* che permette la localizzazione efficiente di oggetti deformabili in immagini statiche o in sequenze video. Una rappresentazione parziale della scena visiva, come è proiettata sulle due dimensioni, è costruita a partire dal calcolo delle risposte di un insieme di filtri di Gabor modificati. I filtri sono disposti sui nodi di un grafo Log-polare sparso (la "retina"). L'algoritmo fa uso di una conoscenza a priori dei pattern da individuare, una descrizione dei quali è codificata in modelli realizzati mediante classificatori Support Vector Machines nello spazio di Gabor. Questi modelli sono impiegati per generare una sequenza di movimenti saccadici che ha lo scopo di centrare la retina sui pattern d'interesse. Tale procedimento non richiede il calcolo della decomposizione di Gabor sull'intera immagine. Il conseguente

aumento di efficienza permette di applicare questo classico e potente strumento anche in un contesto di visione attiva.

Abbiamo deciso di introdurre il nostro concetto di *Retinal Vision* attraverso una serie di applicazioni connesse con l'autenticazione a partire da immagini del volto, un problema che unisce un sufficiente livello di generalità e di complessità tecnica con un evidente interesse scientifico e pratico. Come prima applicazione della *Ricerca Saccadica* presentiamo un sistema in grado di rilevare in tempo reale e di mantenere centrata nel campo di una telecamera mobile la testa di un soggetto. Questo sistema costituisce, a nostro avviso, il primo esempio dell'uso della decomposizione di Gabor in un siffatto problema di visione attiva. La seconda applicazione che consideriamo, la detezione di parti del volto, dimostra l'applicabilità del nostro approccio ad un problema complesso di Pattern Recognition su immagini statiche.

L'efficacia della rappresentazione impiegata, basata sull'estrazione delle risposte di filtri di Gabor modificati sui punti di un grafo Log-polare sparso, è confermata dall'uso di tali risposte come dati biometrici per l'autenticazione a partire da immagini del volto. Gli esperimenti effettuati sui due maggiori database di immagini specificamente raccolti per l'autenticazione hanno fornito risultati molto soddisfacenti.

Un capitolo separato è dedicato allo studio della dimensionalità dell'autenticazione a partire da immagini del volto, un argomento sul quale la letteratura riporta stime molto discordanti. In esso si considerano algoritmi classici come l'analisi in componenti principali (PCA) e i discriminanti lineari (LDA), come pure un algoritmo originale, denominato Support Vector Features. L'uso di queste tecniche in combinazione con classificatori "tradizionali" (KNN) e "moderni" (SVM) permette di investigare l'effetto della dimensionalità dello spazio delle features sulla qualità dell'autenticazione.

To my parents Francesco and Edda,
my brother Paolo
and my sister Elena

"Taking Three as the subject to reason about —
A convenient number to state —
We add Seven, and Ten, and then multiply out
By One Thousand diminished by Eight.

"The result we proceed to divide, as you see,
By Nine Hundred and Ninety and Two:
Then subtract Seventeen, and the answer must be
Exactly and perfectly true.

"The method employed I would gladly explain,
While I have it so clear in my head,
If I had but the time and you had but the brain —
But much yet remains to be said.

The Hunting of the sNark
Fit the Fifth: *The Beaver's Lesson*
Lewis Carroll

Contents

1 Introduction

1.1 Retinal Vision

A typical personal computer spends most of its time waiting for the user to enter some input through the keyboard. In the case of Computer Vision systems, the problem seems to be the opposite: it only takes a fraction of a second to flood a system with tenths of thousands of pixels from a static image acquired by a camera. More realistically, even a (relatively) simple automated system designed to take a well framed, nicely looking passport photo of a person and print it on an ID card must be able to cope with a real-time stream of images in order to select the nicest frame. The problem therefore appears, in this case, to be the inverse: one must find a way to discard the bulk of the input, while retaining and processing the most informative and interesting parts of it. In other words, an *attentional mechanism* is needed in order to allow the system to concentrate its computational power on task-relevant information.

Since human beings are able to master without apparent effort a huge amount of visual information, the human visual system naturally proposes itself as the primary source of inspiration for attentional mechanisms. The present Thesis aims at introducing a new biologically inspired vision paradigm, which we refer to as *Retinal Vision*. This is intended as a general framework that pursues the integration of low-level biologically inspired signal conditioning such as retinotopic sampling and the Gabor decomposition with models of higher-order visual processes that constitute the interface to cognitive processes in human beings. The underlying idea is that of implementing attentional mechanisms that would allow a cognitive task to steer visual acquisition and processing.

At an early stage of the chain that goes from visual stimuli to symbolic representations we can draw inspiration from the *Human Saccadic System*. Humans and primates do not scan a visual scene in a raster-like fashion: they rather perform large jumps, known as *saccades*, between the different points of interest in the scene, on which the gaze is kept centred for a fraction of a second (*fixations*). Saccades are known to play a role in the underlying cognitive processes; the saccadic pattern, as has been demonstrated by the Russian psycho-physicist Yarbus already in the Fifties, depends both on the visual scene and on the task to be performed.

In this Thesis we introduce a *Saccadic Search* strategy that allows the efficient detection of deformable structures in static images or video streams. A sparse *Log-polar* retinotopic grid is centred at the current fixation point. The Gabor decomposition of the image is then sampled at the location of the points of the grid; the resulting Gabor responses are used in the planning of saccades. Computational models of saccades based on the responses of multi-scale orientation selective filters, such as Derivatives of Gaussians, have already been proposed, although no explicit description of a target was used to drive the saccades. The Saccadic Search strategy is task related in that it makes use of *a-priory knowledge* in the form of appearance-based models of the patterns of interest. The models, which are implemented by classifiers, encode a description of the appearance of the search targets in the form of a vector of Gabor filter responses. Different choices of classifiers including the recently developed Support Vector Machines are considered.

Three applications of the *Retinal Vision* paradigm, namely Real-Time Head Detection and Tracking, Facial Feature Detection and Face Authentication have been developed as part of the research work that lead to this Thesis. Each of them is discussed in a separate Chapter of the present work. Finally, one Chapter is dedicated to considerations on the dimensionality of the Face Authentication problem, with particular reference to the relations between Principal Component Analysis, Linear Discriminant Analysis and Support Vector Machines.

1.2 Real-time Head Detection and Tracking

We introduce the *Saccadic Search* strategy through an application to Real-time Head Detection and Tracking, which is meant to demonstrate the viability of this approach in an active-vision scenario. A Log-polar retinotopic sampling grid is considered to be rigidly attached to the sensor of a pan-tilt camera. Motion detection is used to identify regions of interest in the image, thus "activating" the relevant points of the sampling grid. Real-time performance is achieved by computing the Gabor responses only on (a subset of) the retinal points, which can be done by direct image domain filtering. After a frame has been processed, a saccade is performed by aiming the camera in the direction of the retinal point that best matches a pre-computed head model; this procedure is iterated to achieve tracking.

The pan-tilt unit employed consists in an eyeball-mounted camera explicitly designed to mimic the performance of the human eye. This special hardware was developed by O. Carmona and S. Menot at LAMI-EPFL.

To the best of the Author's knowledge, the Head Detection and Tracking system described represents the first example of the use of the Gabor decomposition in such an active vision task.

1.3 Facial Feature Detection and Face Authentication

Face Authentication is a fascinating problem in Computer Vision. From the technical point of view, the wide variability of facial appearance deriving from transient changes in pose, hair style and expression as well as from aging constitutes a formidable challenge.

Furthermore, in spite of the fact that research on Face Authentication is fostered by the growing, market-related need for reliable biometrics, the problem retains a fundamental scientific interest, that can be resumed by the question *"How do human beings recognise faces?"*.

These facts, along with the demonstrated effectiveness of multi-scale tools and non-linear sampling in recognising faces, make Face Authentication the ideal test-bed for the synthesis of bio-inspired computer vision algorithms proposed by the *Retinal Vision* paradigm.

The main regions of interest for the face recognition/authentication task are, as is well known, located around the eyes and the mouth of a subject. Detection of these facial features can be performed using the *Saccadic Search* strategy proposed in this Thesis. Models of the eyes and the mouth are implemented by means of Support Vector Machine classifiers. Saccadic patterns are iteratively generated by centring the retina at the point that maximises the output of the active target model.

Once the eyes and the mouth of the subject have been located, Face Authentication is achieved by using three independent machine experts to process the Gabor features extracted in the corresponding facial regions. Each expert, which is implemented as a classifier, outputs an authentication score. These are then combined in order to achieve the final decision on whether the identity claim being considered is true, a process that is known as Fusion of Experts. Different implementations of the machine experts as Support Vector Machines or other types of classifiers are discussed.

The Facial Features Detection and Face Authentication algorithms have been tested on the subset of frontal images from both the M2VTS and the Extended M2VTS (XM2VTS) databases, featuring images of 37 and 295 subjects respectively. These databases are explicitly intended for authentication purposes; in both of them images of each subject have been acquired on four distinct occasions to guarantee within-class variability for all persons. Authentication tests were performed according to the standard test protocols established by the European M2VTS consortium, thus allowing direct comparison of our results with those published by other research groups working on the same databases.

1.4 Dimensionality in Face Authentication

The investigation of the dimensionality of the Face Authentication problem has both a theoretical and a practical interest. From the theoretical point of view, it provides a measure of the complexity of the problem and of the intrinsic variability, so to speak, of human faces. This could prove useful in the analysis of the information theoretic aspects of human recognition of faces, as well as in the design of forensic systems that help witnesses draw an Identikit of a suspect.

From the practical point of view, knowing the number of features required for reliable authentication/recognition would help in the choice and dimensioning of a suitable media, such as a smart-card, on which to store the biometric data.

The existing Face Authentication/Recognition literature on the subject is largely contradictory. Estimates range across two orders of magnitude, from a minimum of about 30 features to over one thousand.

A study of the dimensionality of the Face Authentication problem using Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM) is presented. Starting from the Gabor feature space employed by the Face Authentication algorithm, an attempt to identify the most relevant subspace is performed using PCA, LDA and a novel dimensionality reduction algorithm that we call Support Vector Features (SVF). These techniques are shown to identify three different subspaces. The generalisation performance of the three solutions is assessed by using the subspaces obtained from one database (M2VTS) to authenticate clients from a different database (XM2VTS). PCA and SVF are shown to generalise best.

Experiments using two types of classifiers indicate that the number of Principal Components or Support Vector Features required for the authentication performance to saturate heavily depends on the choice of the classifier.

In particular, Support Vector Machine classification involves directions in the feature space that have weak variances and appear to be vulnerable to excessive PCA-based compression.

1.5 Organisation of the Thesis

The low-level signal conditioning stage employed in the *Retinal Vision* paradigm consists in the computation of a set of Gabor responses at each point of a sparse retinotopic grid, that is superimposed on the images. The procedure is outlined in Chapter 2, where references to similar approaches are also given. The Gabor filters employed are non-standard filters designed using the Log-polar mapping in the complex frequency plane. Their construction, which is both mathematically elegant and straightforward, is detailed in Appendix A.

The *Saccadic Search* strategy is introduced in Chapter 3 with an application to Real-time Head Detection and Tracking in an active-vision scenario. The system described probably represents the first attempt to use the Gabor decomposition in such an active-vision framework.

The application of *Saccadic Search* to more complex tasks benefits from an improved modelling of the search targets, that can be achieved by advanced classification techniques. For the Facial Feature Detection and Face Authentication applications described in this Thesis, Support Vector Machine classifiers have been used. An introduction to Support Vector Machines is provided in Chapter 4.

The main application developed so far for the *Saccadic Search* strategy is Facial Feature Detection, as discussed in Chapter 5; experimental results over the set of facial images from the M2VTS and Extended M2VTS (XM2VTS) databases are presented.

The Face Authentication results presented in Chapter 6 confirm the descriptiveness and efficiency of the feature extraction procedure introduced in Chapter 2 as part of the *Retinal Vision* paradigm. Authentication experiments are performed on the M2VTS and Extended M2VTS (XM2VTS) databases, that constitute the two largest authentication oriented databases available. Details on the databases and on the standardised experimental protocols employed are given in Appendix B.

A comparative study of the dimensionality of the Face Authentication problem using Principal Component Analysis can be found in Chapter 7. A novel SVM-based compression technique, that we call Support Vector Features, is also presented in this section.

Finally Chapter 8 concludes the Thesis with a few remarks on the contributions and results of this work and suggests possible directions for a further development of the *Retinal Vision* approach.

2 Feature Extraction

2.1 Computational models of low-level biological vision

Although very little is known about the way the human brain analyses images, the *primary visual cortex*, where some of the low-level visual processing occurs, is indeed the best understood of all the cortical regions [43, 42]. The analysis and transformation of the visual information that take place in these 15 square centimetres of cortex are known in some detail. Mathematical models have been suggested for at least two important aspects of such transformation, namely the *retino-cortical mapping* and the responses of the so-called *simple cells*.

One of the most widely known facts about the human visual system is that visual acuity is not uniform, but decreases from the centre of the retina (the *fovea*) towards the periphery. The regions of highest resolution occupy relatively more cortical area than the rest; the central part of the retina has a representation some 35 times more detailed than the far peripheral part [42]. Conversely, cortex cells corresponding to the outer visual field are associated with larger receptive fields, that is, they respond to stimuli coming from wider areas of the retina.

It has been suggested that the two-dimensional structure of the cortical representation of a visual stimulus is best described by a complex logarithmic function of the retinal coordinates. This hypothesis has been found in good agreement with the measured global structure of the cortical mapping of several primates and of the cat [84]. An attractive feature of such a model is that it provides a possible cue to visual invariance with respect to scaling and rotation. If the point of fixation is held constant, the complex logarithmic (Log-polar) mapping transforms rotations and scale changes in the visual scene into translations. Therefore, the cortical pattern of activity corresponding to a large or to a small stimulus, for a given fixation point, would be invariant up to a shift across the cortex. The same property would hold for rotations.

It is also commonly known that in the primary visual cortex orientation-sensitive cells can be found that only respond to lines or edges that cross their associated receptive fields in the retina with a specific orientation [43, 41, 42]. These so-called *simple cells* map stimuli characterised by similar orientations to adjacent vertical slabs in the visual cortex. It has been argued [22, 69] that a suitable computational model for such cells is represented by Gabor filters [5, 37, 74], that make allowance both for the orientation selectivity and for the variation of the sizes of the receptive fields across the visual cortex.

Before discussing the way the above mentioned mathematical models of biological low-level visual processing are integrated in the *Retinal Vision* paradigm, we will briefly review the use of retinotopic sampling and the Gabor decomposition in Computer Vision.

2.2 Retinotopic sampling in Computer Vision

It is generally assumed that, in order to recognise an object, it is necessary to have an "accurate" representation of its most salient features (in the simplest case, "accurate" can be interpreted as a synonym for "high resolution") [99]. On the other hand, efficiency considerations suggest that such a high level of detail should be exclusively reserved to the most relevant parts of the image. This can be achieved, for instance, by limiting the high resolution representation to localised interest windows in the images [16]. However, it is not evident that a collection of highly descriptive frames deprived of context information constitutes a good representation of an object. One should also provide some global information on the image. This can be accomplished, for instance, by constructing a hierarchy of overlapping windows with sampling rates inversely proportional to their sizes [16]. A valid alternative is provided by a space-variant, "retinotopic" sampling strategy. In this way, the low-resolution peripheral vision provides a context for the detailed information captured by the fovea.

The main advantage of this approach comes from the fact that it affords a consistent data reduction with respect to uniform sampling while preserving a wide field of view. For this reason, retinotopic sampling has an established role in Computer Vision and is considered particularly attractive for speeding up heavy computational tasks. Examples are optical flow calculations applied to robot navigation [27, 100], binocular vergence control [80] or, as in the case of the present Thesis, Gabor-based pattern recognition in real time (see Chapter 3).

Many different strategies have been proposed for space-variant sampling. Among the others, we cite the reciprocal wedge transform (RWT) [101] and the Log-polar or complex logarithmic mapping [46, 79]. The Log-polar mapping is receiving a growing attention [27, 75, 99, 100], also due to the fact that it constitutes a good model for the retino-cortical mapping [84]. Special camera sensors have also been developed that have their pixels arranged in a Log-polar lattice [25, 80]. However, there is an important difference between these camera sensors and the Log-polar mapping as it is found in biological vision. Retinotopic pixel arrays acquire an *intensity* map of varying resolution; on the contrary, the retino-cortical mapping refers to the topology of the *orientation* chart encoded by the simple cells. The same criticism applies to the way nonuniform sampling is used in most of the existing applications. Exceptions include [95], where the retinotopic sampling is applied to the responses of a set of Gaussian filters that provide some degree of orientation selectivity (filtering is performed over a standard, square pixel image). Computation of the responses of multi-scale Derivative of Gaussian filters on the points of a Log-polar grid has been shown in [75] to allow 100% accurate recognition of the 20 objects in the COIL-20 database. Always in [75], the different issue of applying multi-resolution orientation selective filters to an image acquired by a Log-polar sensor is discussed.

2.3 The Gabor decomposition: history and applications

The Gabor decomposition also constitutes a classical and ubiquitous tool in Computer Vision. Since the pioneering work of Gabor in the context of data transmission [37], this powerful instrument has undergone a certain evolution. In its original form the Gabor decomposition closely resembles a windowed Fourier transform (or short-time Fourier transform) [2, 49], but for the fact that the window function applied to the data is narrow as compared to the size of the acquisition window and that it is translated along the (one-dimensional) signal. In practice, the signal is projected onto a set of functions of the form

$$\mathcal{G}_{mn}(x,\omega_0) = g(x - nx_0)e^{im\omega_0 x} \tag{2.1}$$

obtained by modulating the complex sinusoidal waves that characterise the Fourier transform with a Gaussian window

$$g(x) = e^{-\frac{x^2}{2\sigma^2}}. \tag{2.2}$$

The window is displaced along the signal in steps of length x_0 (which is chosen to give $x_0\omega_0 = 2\pi$) [20]. Perhaps the most advantageous property of the Gaussian window function is that it is optimally concentrated in both time and frequency, which makes it suited for the case in which spatial and spectral localisation are both important (for the two dimensional case, this concept has been developed in [23]).

If we drop the translation term in Equation (2.1) by choosing $n = 0$, we get the family of "filters"

$$G_m(x,\omega_0) = \mathcal{G}_{m0}(x,\omega_0) = e^{-\frac{x^2}{2\sigma^2}}e^{im\omega_0 x}. \tag{2.3}$$

Since the standard deviation of the Gaussian envelope σ is constant it is easy to see that, as m increases, the number of cycles of the sinusoidal wave under the region of the envelope that is significantly different from zero also increases.

Early in the '80s, the geophysicist J. Morlet introduced the wavelet decomposition in Signal Analysis through what he called "wavelets of constant shape" [39, 65]. The main element of novelty with respect to the Gabor filters consisted in the fact that the Gaussian window function was now dilated or shrunk so as to keep the general shape of the "wave packet" constituting the filter unaltered at the different scales [38]:

$$M_m(x,\omega_0) = G_m(xm,\omega_0/m) = e^{-\frac{x^2}{2(\sigma/m)^2}}e^{im\omega_0 x}. \tag{2.4}$$

It follows that $M_m(x, \omega_0) = G_0(mx, \omega_0)$, and therefore Morlet wavelets are obtained from scaling a single Gabor filter[1]. For this reason these filters are often referred to, in the Computer Vision community, as Gabor "wavelets" [24, 55, 59]. The present work is mostly concerned with yet another family of filters analogous to Gabor wavelets, that are introduced in Section 2.4.2. These we shall indicate as "modified Gabor filters". We shall use the generic expression "Gabor filters" for either Gabor wavelets or these modified filters, as will be made clear by the context. We shall also use "Gabor decomposition" as a catch-all term for Gabor-based techniques, among which we include Gabor wavelets and the modified Gabor filters.

The applications of the Gabor decomposition in Computer Vision are countless. Among them we cite texture analysis [10, 14, 45]; unsupervised text segmentation [44]; image analysis and compression [24]; image feature detection for pattern recognition, image registration and motion tracking [59]. Gabor feature vectors have also been widely used to label the nodes of sparse graphs for object recognition and face recognition/authentication [28, 55].

Other multi-scale techniques for the analysis of local orientation charts have also been proposed. Among these, we cite linear symmetries [12] and derivatives of Gaussians [77]. These techniques trade a somewhat less informative representation of the image against a lower computational complexity. Gabor-based algorithms naturally compare also with those relying on the wavelet transform with other families of wavelets [3]. Here again, for instance, Haar wavelets afford a reduced computational load at the expense of poor spectral selectivity [2] (for an application see [71]). To date, there is no definite evidence of the superiority of one particular family of wavelets for Pattern Recognition tasks. Gabor wavelets remain a valid choice because of the underlying biological analogy and of the optimal space–frequency localisation property of the Gaussian window.

2.4 Feature extraction in Retinal Vision

2.4.1 The retinotopic sampling grid

The low-level biological signal processing stages described in Section 2.1 are implemented in the *Retinal Vision* paradigm through their mathematical models, the complex logarithmic (Log-polar) mapping and the Gabor decomposition.

The complex logarithmic mapping is employed in the context of sparse labelled graphs

[1] The exact expression of a Morlet wavelet actually is

$$\mathcal{M}(x, \omega_0) = e^{-\frac{x^2}{2\sigma^2}} e^{i\omega_0 x} - e^{-\frac{x^2}{2\sigma^2}} e^{-\frac{(\sigma\omega_0)^2}{2}}$$

The second term makes the filter DC-free and is required for the function to satisfy the so-called *admissibility* condition [3]. In practice, if the value of $\sigma\omega_0$ is large enough this correction term can be discarded. Intuitively, this means that the Gaussian envelope must be large enough to accommodate at least a few cycles of the sine wave.

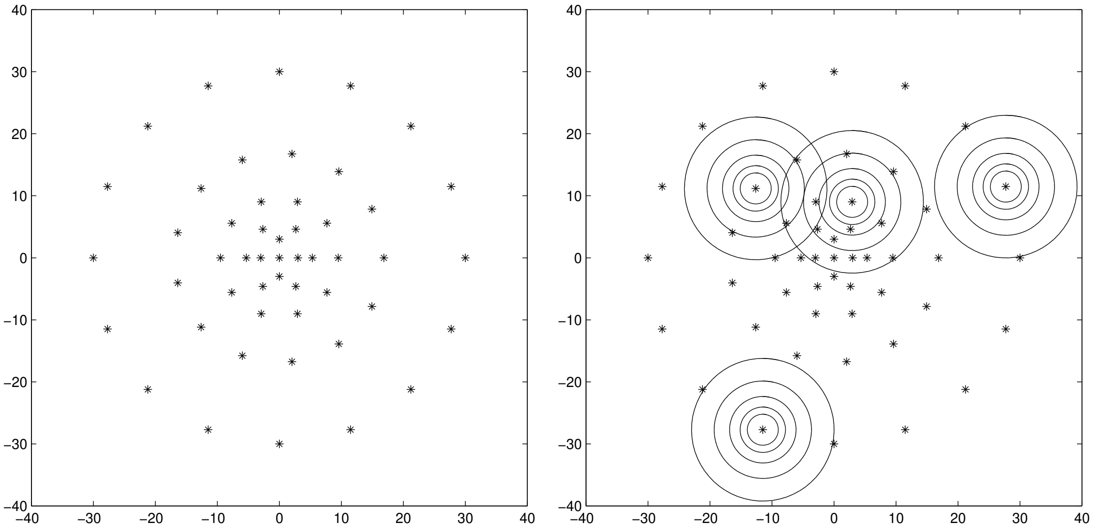


Figure 2.1: The retinotopic sampling grid (left) employed for Facial Features Detection (Chapter 5) and Face Authentication (Chapter 6). Axes are graded in pixels. A few receptive fields are represented as sets of five concentric circles, one for each Gabor frequency channel (right; details can be found in Section 5.3.1). The circles delimit the region in which the magnitude of the corresponding filters is larger than $1/\sqrt{2}$ times its maximum magnitude.

rather than for sampling the gray-level values of the image. We construct a retinotopic grid with a Log-polar arrangement of points. The points of the grid (henceforth also "the retina") are used to sample the Gabor decomposition of the image, each retinal point being associated with a *receptive field* represented by the support of a set of (modified) Gabor filters (Figure 2.1). This constitutes a simple computational model of both the cortical mapping and the responses of the simple cells in the primary visual cortex. From the mathematical point of view, the construction of the retinotopic grid is straightforward. If the 2D spatial coordinates of the grid points are represented by means of complex numbers $s_{c\gamma}$, the sampling point lying on circle $c = 1, 2, \ldots$ and with orientation index $\gamma = 0, 1, \ldots, N_c$ has coordinates $(x_{c\gamma}, y_{c\gamma})$ given by

$$s_{c\gamma} = x_{c\gamma} + iy_{c\gamma} = r_{min}\delta_r^{(c-1)}e^{i\frac{2\pi}{N_c}\gamma}. \tag{2.5}$$

Here r_{min} is the radius of the innermost circle, δ_r is the (constant) ratio between the radii of two consecutive circles and N_c is the number of grid points on circle c. We define s_{00} to be the centre of the retina, which coincides with the origin of the complex plane.

The actual values of these parameters vary with the application and are determined in relation with the dimension of the Gabor filters that constitute the associated receptive fields, as well as on the basis of empirical evidence. Details concerning each application are given in the respective Chapters; general guidelines for the design of Gabor filters can be found in Appendix A.

2.4.2 Log-polar sampling and the Fourier domain

The local power spectrum of the image is sampled at each retinal point by the vector of Gabor filters that constitute the receptive field associated with that point. In order for the filter responses to be descriptive of the image neighbourhood and to carry a maximal information content, i.e. to be uncorrelated, the local frequency plane must be covered as uniformly as possible.

A common set of Gabor filters used in Computer Vision (including Face Recognition) is constructed starting from the two-dimensional Gabor function

$$G(x, y|\sigma_x, \sigma_y, \omega_0) = e^{-\frac{x^2}{2\sigma_x^2}} e^{-\frac{y^2}{2\sigma_y^2}} e^{i\omega_0 x}. \tag{2.6}$$

This equation represents a Gaussian centred at the origin, modulated by a complex planar wave with its wave vector parallel to the x axis. When used as a filter on an image, it will selectively respond to vertical structures that have non-zero frequency components at approximately $\omega_0/2\pi$.

The Fourier transform $\hat{G}$ of G is given by

$$\hat{G}(\omega_x, \omega_y|\sigma_x, \sigma_y, \omega_0) = Ne^{-\frac{(\omega_x-\omega_0)^2}{2\sigma_{\omega_x}^2}} e^{-\frac{\omega_y^2}{2\sigma_{\omega_y}^2}}, \tag{2.7}$$

where $\sigma_{\omega_x} = 1/\sigma_x$, $\sigma_{\omega_y} = 1/\sigma_y$ and N is a normalisation factor. This is again a Gaussian function shifted along the horizontal frequency axis by the modulation frequency ω_0.

An entire set of two-dimensional Gabor filters (more specifically, Gabor wavelets) can be obtained by scaling and rotating the Gabor function (2.6). Since rotations of G correspond to rotations of $\hat{G}$ and a scaling of G by a factor s results in $\hat{G}$ being scaled by $1/s$, the same set of Gabor filters can equivalently be constructed in the Fourier domain. To this end, it is sufficient to apply the same rotations and the reciprocal scalings to the Fourier transform $\hat{G}$ of the Gabor function G.

Gabor filters are optimal in that they minimise the joint image and frequency plane spread, yielding compact filters that are nevertheless well tuned to a frequency [23]. When only a small number of frequency channels is employed, however, pure Gaussian filters tend to display an excessive overlap towards the (densely sampled) origin of the frequency plane, while the high frequency regions are poorly covered. This is due to the fact that each Gaussian weights high as well as low frequencies in its support in a perfectly symmetric manner, whereas the decomposition itself is coarser at high frequencies. This negatively affects the independence and descriptiveness of the features. In order to compensate for that, we chose to employ a set of *modified Gabor filters* that are represented by Gaussians in the *Log-polar* frequency plane [9, 51]. That is, a filter tuned to orientation ϕ_0 and angular frequency $\omega_0 = exp(\xi_0)$ has the following function form:

$$\hat{G}'(\xi, \phi) = Ae^{-\frac{(\xi-\xi_0)^2}{2\sigma_\xi^2}} e^{-\frac{(\phi-\phi_0)^2}{2\sigma_\phi^2}} \tag{2.8}$$

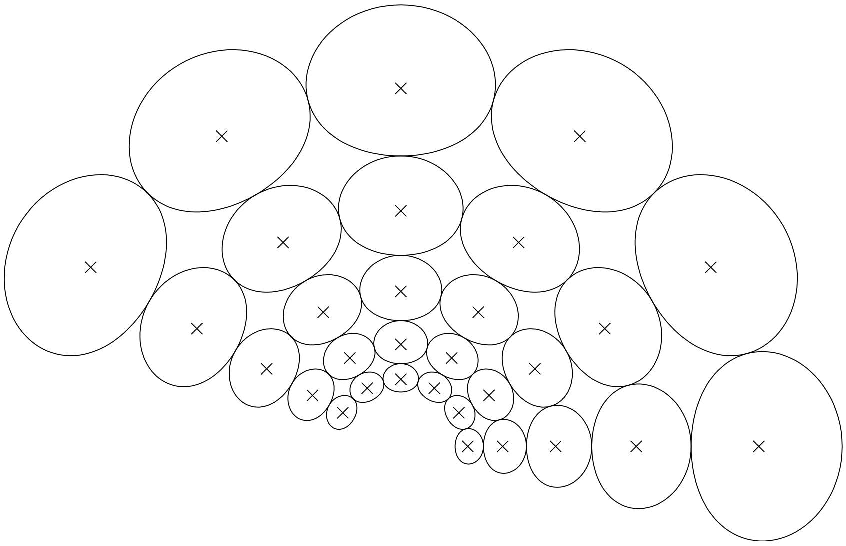


Figure 2.2: A set of modified Gabor filters in the Fourier domain. Maxima are marked with crosses. Contours correspond to level curves of the Gaussians in the Log-polar frequency plane. As can be seen, the effect of the mapping is to push the inner part of the level curves towards the maximum, thus reducing the overlap at lower frequencies.

where A is a normalisation constant and (ξ, ϕ) are the Log-polar frequency coordinates:

$$(\xi, \phi) = \left(\log \|(\omega_x, \omega_y)\|, \tan^{-1}(\omega_y/\omega_x)\right). \tag{2.9}$$

The expression for $\hat{G}'$ is the analogue of Equation (2.7) in the Log-polar frequency plane. The new coordinate system has the property that translations along the ϕ axis are mapped back to rotations in the image domain, while translations along the ξ axis represent a scaling of the image. A complete bank of filters can thus be designed by arranging a set of *identical* Gaussians in a *rectangular lattice* in the Log-polar frequency plane. When seen in the standard Fourier domain, the filters reproduce the familiar "daisy" structure (Figure 2.2), but the overlap towards low frequencies is significantly reduced (Figure 2.3).

These filters have been previously used in texture analysis problems, showing high discrimination power [8]; details on their design and implementation are given in Appendix A.

2.4.3 Magnitude versus phase

The Gabor decomposition is complex-valued. In analogy with the Fourier transform, it makes sense to think of it as carrying a *magnitude* and a *phase* information.

The magnitude of Gabor responses represents the local image spectrum, i.e. the local repartition of the energy among frequency components, much in the same way as the spectrum obtained with a windowed Fourier transform [49] (the only difference lies in the

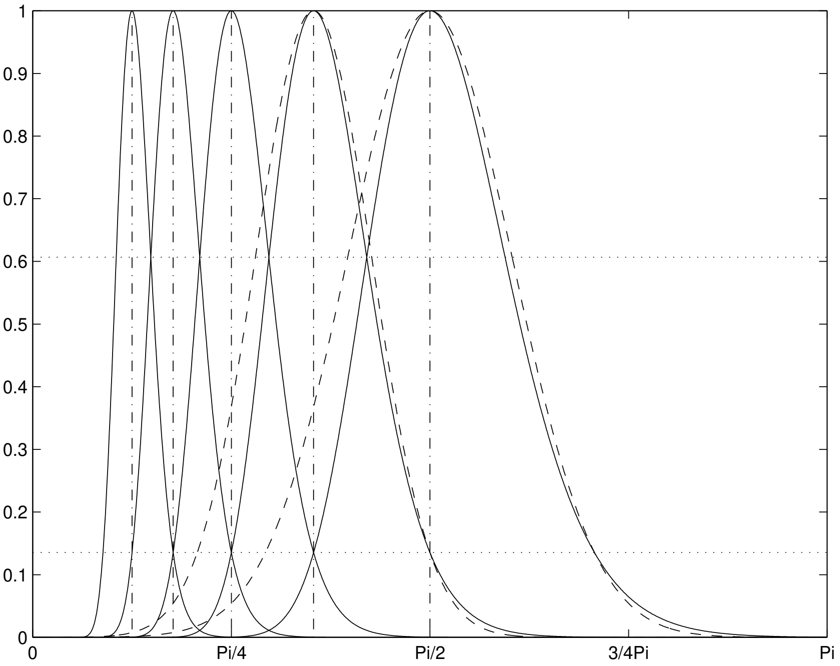


Figure 2.3: This cross-sectional plot of the Fourier domain shows the magnitude of our set of modified Gabor filters designed in the Log-polar frequency plane (solid lines, Equation 2.8) superimposed with two of the standard Gabor filters (dashed lines, Equation 2.7). The standard deviations have been set to half the separation between the centres of the filters in the respective coordinate systems; the horizontal dotted lines intersect the filters at σ and 2σ. The abscissa is labelled with the angular frequency. The graph exhibits the asymmetric shape of the modified filters, that have a steeper cut-off on the low frequency side. This effectively reduces the overlap between adjacent channels.

fact that in the case of the Gabor decomposition the spatial resolution varies together with the frequency, see Section 2.3).

The Gabor phase is related to the "sign" of intensity variations or discontinuities (light to dark or vice-versa) [32]. In analogy with the shift property of the Fourier transform it is sensitive to translations, and because of its cyclic nature it has an inherently local character. For these reasons, the Gabor phase has mainly been used in texture segmentation [8, 14, 33], which involves local comparisons between adjacent image regions, as well as in motion estimation [4, 36] and stereo matching [35], two tasks that require detecting (small) translations in similar images.

For most applications, including pattern matching tasks that potentially involve large deformations [59] and algorithms that make use of deformable graphs to encode geometric relationships [28, 55], the magnitude of Gabor responses is generally preferred.

Throughout this work, only the magnitude of the responses of modified Gabor filters is considered. For simplicity, we will refer to such quantities as *Gabor features* or *Gabor filter responses*, without any additional qualification.

3 Saccadic Search

3.1 The human saccadic system

Humans and primates do not scan a visual scene in a raster-like fashion, as it would be drawn on a television screen. They rather perform large jumps, known as *saccades*, between the points of interest [17, 43, 105]. Saccadic eye movements are characterised by a high angular speed (700°/s) and essentially constitute ballistic movements, meaning that they are pre-planned and cannot be modified by visual feedback during their execution. In fact, all the visual processing takes place during the roughly 300 ms long *fixations* that separate a saccade from the next. The visual system actually spends much of this time planning the following saccade.

The study of saccades is technically demanding, since it requires the capability to track the gaze of a subject. Early techniques were invasive and uncomfortable. Already in the Fifties, however, the Russian psycho-physicist A. L. Yarbus managed to demonstrate that the saccadic pattern depends both on the visual scene and on the cognitive task to be performed [105]. Furthermore, the saccadic pattern appears to have a certain stability: each subject will tend to reproduce a certain pattern each time he or she is presented with a given cognitive task over a given image, and this even after an interval of a few days.

Because of their close connection with cognitive tasks, eye movements are not to be considered just as a way to compensate for reduced visual acuity in the peripheral field. Rather, saccades seem to play an important role in the selection of task-relevant information. Findings from experimental research on a visuo-motor task [72] suggest that, at least for the type of task considered in the study, we do not store in our brain a complete description of the outside world. We rather seem to maintain in our working memory a set of pointers to the areas of the visual field in which the information required to complete the task is located; saccades are then used to actually retrieve the information the moment it is needed.

3.2 Computational models of saccades

Computational models of saccades have been proposed both in the framework of biological modelling and in an applied scenario.

Simple models of attention generate saccadic paths in a purely input-driven fashion, relying on bottom-up saliency criteria provided by motion or colour cues. In [66], for instance, an active camera is employed to perform reactive capture saccades prompted by image motion. Saccadic planning is based on optical flow calculations.

Such models do not incorporate any task-specific knowledge that might clarify the relation of saccadic eye movements to cognitive tasks. A computational model of human saccades that also considers top-down models of the objects to be located is presented in [75] (comparison with human performance is given in [76]). This work makes use of multi-scale orientation-selective filters (Derivatives of Gaussians) to build iconic representations of the search targets, which are encoded as sets of vectors of filter responses. Object location is achieved by matching filter responses from the visual scene with the model vectors. Scene analysis and matching are performed starting from the largest scale. This allows building a first, coarse saliency map. A series of increasingly refined saliency maps is generated as the system goes through all the available scales, in a coarse to fine order. The polygonal line obtained by joining the maxima of adjacent saliency maps in this sequence constitutes an estimate of the saccadic pattern a human being would follow to perform the same object location task. While in good agreement with biophysical data, this algorithm effectively performs exhaustive search on the images, that are entirely rescanned at each resolution. The use of spatiochromatic filters in place of the intensity filters is discussed in [77].

A model of saccadic camera control based on the use of Gabor filters and its application to robot navigation are presented in [52]. An example of an application specific model can be found in [96], where saccades are considered in the framework of Face Recognition. The responses of a set of anisotropic Gaussian filters are used to perform reactive saccades that build up a saliency map including the most relevant regions of the image. In a second step, tuned filters are directly applied to the image intensity values in order to locate the relevant facial landmarks. This second stage, in which task-related knowledge is introduced, therefore abandons the biological analogy in that it bypasses the orientation-based representation.

3.3 Saccadic Search in Retinal Vision

The *Retinal Vision* paradigm features the *Saccadic Search* strategy as a general purpose attentional mechanism for the recognition of deformable objects.

The model proposed is task-related in that it makes use of *a-priori knowledge* about the search targets. Such knowledge is encoded by classifiers that are trained to recognise the relevant patterns based on their Gabor signature, as it is extracted by the retinotopic sampling grid introduced in Section 2.4. The orientation decomposition is used both for saccade planning and for complex pattern recognition tasks such as Facial Features Detection and Face Authentication, the two main applications considered in this Thesis. From this point of view, the *Retinal Vision* paradigm carries the biological analogy further in that it does make use of an image representation inspired by the workings of the primary visual cortex to perform complex cognitive tasks.

From the computational point of view, the use of an attentional mechanism in conjunction with the Gabor decomposition is desirable in order to reduce the computational load. Despite being a ubiquitous tool in Pattern Recognition, Gabor features are generally considered impractical for active-vision applications. However, the use of a sparse Log-polar retina with a relatively low number of sampling points makes feature extraction affordable even for real time systems. Gabor responses can be extracted by direct filtering in the image domain.

In the next few sections, we introduce the *Saccadic Search* strategy through an application to Real-time Head Detection and Tracking. The system described is, to the best of the Author's knowledge, the first example of the application of the Gabor decomposition in such an active-vision scenario. The results we present have been published in [93, 92].

3.4 Real-time Head Detection and Tracking

Head Detection and Tracking is an application of Computer Vision that is taking on a growing importance, since it represents a first stage in many face-based authentication and surveillance systems. The task in itself is up to date unsolved in the general case and presents large difficulties due to the wide variability of the search target in appearance (hair, spectacles, occlusions), scale, pose, illumination and to the constraints of real-time processing.

We introduce a moderately constrained Head Detection and Tracking system designed to demonstrate the applicability of the *Saccadic Search* strategy in an active-vision scenario. The target application is an automated login system in which a computer terminal user is automatically authenticated and logged into the system when he or she sits in front of a terminal. The user is then automatically logged out on leaving the site. Such a device would find a practical application in a dynamic working environment in which several operators alternate in front of a computer, a scenario in which an active control of the user's identity is evidently needed.

The prototype system has been implemented using an eyeball-mounted camera developed at LAMI–EPFL by O. Carmona and S. Menot. This pan-tilt unit was expressly designed to mimic the dynamic performance of the human eye.

3.4.1 Coding the a-priori knowledge

The *Saccadic Search* algorithm is driven by an appearance-based model of the search target. This is obtained from the image representation given by the feature extraction stage described in Section 2.4.

A training set of images was acquired by having 6 different subjects sit in front of the (static) camera. A short sequence of 30 non-consecutive frames was collected from each subject, for a grand total of 180 frames. Each subject was asked to move slightly and change his/her expression during the acquisition. The images were acquired at a resolution of 320×240 pixels (half of standard SECAM resolution).

For this application we chose a simple implementation of the head model, based on the response vector $\vec{v}$ of a set of 24 (modified) Gabor filters $G_i(x, y)$ placed on the middle point (x_m, y_m) between the eyes of the subjects:

$$\vec{v}_i = \left| \iint I(x_m - x, y_m - y) G_i(x, y) \, dx \, dy \right| \tag{3.1}$$

where $I(x, y)$ is a training image. Note that each component $\vec{v}_i$ is the magnitude of a complex filter response. A Gabor response vector $\vec{v}$ has been computed for each image in the training set; the average response vector $\vec{v}_{av}$ constitutes the head model. Similarity with respect to the model is measured by the Euclidean distance from the average vector $\vec{v}_{av}$ in the 24-dimensional Gabor feature space.

The filters employed are organised in three frequency channels with wavelengths ranging from 8 to 12 pixels; there are 8 orientation channels. Due to the wide-angle optics adopted, the average distance between the eyes in a typical image is about 15–20 pixels, which is comparable with the wavelength of the filters. This is necessary if we want the filter responses to encode a signature of the whole head rather than of its subparts. Under these conditions, we can expect the map of the Euclidean distance between the head model $\vec{v}_{av}$ and the Gabor response vectors extracted at all the points of an image containing a head to present a single pronounced minimum in correspondence of the head itself.

3.4.2 Motion detection

The feature extraction stage described in Section 2.4 has been slightly modified in order to meet the demands of real-time processing on a relatively low-performance unspecific hardware. Specifically, a simple motion detection algorithm has been added that rules out uninteresting points of the Log-polar retinotopic grid.

As the first step in the tracking procedure, two consecutive frames are acquired and the absolute value of their difference is computed. Thresholding is applied to identify pixels that differ significantly between the two images. The field of view is then partitioned according to a 10×10 rectangular grid. Each grid cell can be labelled as a region of motion or not according to the percentage of significantly differing pixels it contains (Figure 3.1). Only the parts of the image that have been marked as regions of motion are considered for further processing. This step also allows us to make sure that the camera has stopped moving before a frame is considered for processing. If more than 70% of the image appears to contain motion, the current frames are discarded and two new ones are immediately acquired.

The Log-polar retinotopic grid (Figure 3.2) is centred in the image and can be thought of as if it were "attached" to the camera, in the sense that it moves rigidly on the scene when the camera moves. According to the assumptions of the *Retinal Vision* paradigm,

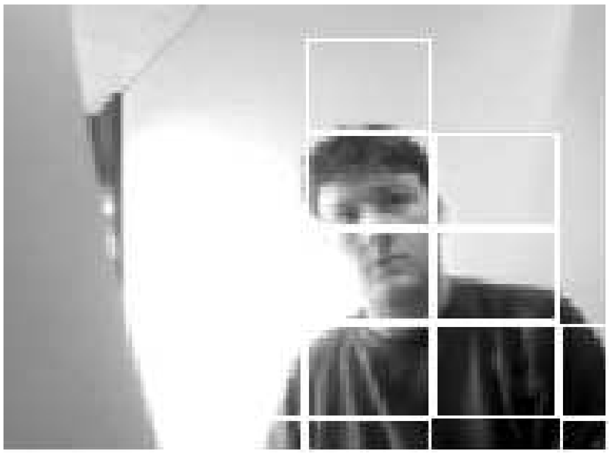


Figure 3.1: The output of the motion detection stage is a set of rectangles surrounding moving objects in the scene.

the information contained in the grey level of the image pixels is accessible to the system only to the extent that it contributes to the response of a Gabor filter centred at one or another of the retinal points.

Access to a more detailed description of a particular image area is obtainable only by centring the retina at that position. Since the number of points in the grid is small (29 in our experiments), the required responses can be computed directly by filtering in the image domain, with a reduced computational effort.

A further selection among the retinal points is in this case performed based on the information from the motion detection stage: the set of 24 Gabor responses is only computed at the grid points that fall into regions of motion in the image. The central point of the retina is always considered in the computation regardless of motion, so that the system is allowed to keep fixating the same spot if nothing has changed in the scene.

3.4.3 Performing saccades

Finding the position of a head in an image requires, in our framework, comparing the Gabor feature vectors extracted at each retinal point with the camera pointing in a given direction against the pre-computed head model.

A saliency value is assigned to each retinal point based on the Euclidean distance between the feature vector extracted at that point and the head model. A saccade is then performed that reorients the camera in the direction of the point at which the best match has been found.

Each frame is processed independently of the preceding ones, i.e. there is no time integration. Nevertheless, the structure of the retina, which is denser near the centre, helps to concentrate the computational effort in the region in which the head was found at the previous iteration. Also, as the system homes in on a head the resolution at which

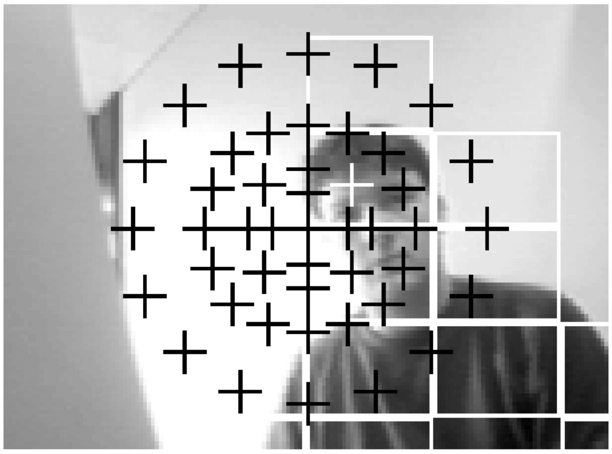


Figure 3.2: Gabor features are extracted only at the retinal points that happen to fall inside a region of motion. The retinal point at which the best match with the head model is found (marked white in the image) is selected as the target for the next saccade.

the Gabor decomposition is sampled automatically becomes higher, allowing for finer adjustments. On the other hand, motion detection can draw the attention of the system to the periphery of the retina, thus allowing for quick recovery from tracking failures or rapid detection of subjects entering the field of view.

3.4.4 Hardware setup

The algorithm was implemented using the Vision Sphere developed at the Microprocessor and Interface Laboratory of the Swiss Federal Institute of Technology, Lausanne, by O. Carmona and S. Menot (Figure 3.3). This system acquires high resolution images from a standard CCD camera that can tilt and pan under computer control. Its innovative principle is close to a "reversed" computer mouse: two orthogonally placed motors, replacing the encoders in a standard mouse, move the sphere. The position of the camera is known using two orthogonally placed incremental optical encoders. Absolute desired camera positions are transmitted through a serial line to a Motorola 68336 micro-controller card. This cards generates 20 KHz pulse-width modulated 8-bit precision signals that drive the motors. The controller was elaborated by a robust-control pole-placing method. The low inertia of the Vision Sphere affords a fast response to reach any point of interest in the view space (Table 3.1). Control is made easy by the fact that the two rotation axes and the optical axis of the camera intersect in a single point.

Vision computations are carried out on a 200 MHz Pentium PC equipped with a PCI Matrox Meteor frame grabber. The average cycle time, including the time required by the camera to move, is about 500 ms; it is generally less for small corrections, since in that case the motion detection stage activates fewer retinal points and the camera settling time is shorter. For comparison, human beings typically perform 300 ms long fixations.

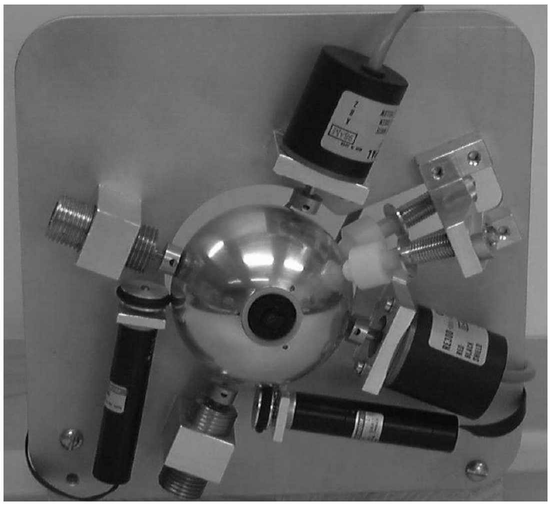


Figure 3.3: Front view of the Swiss Institute of Technology Vision Sphere. The two narrow black cylinders are the DC motors; the larger ones are the optical encoders (Courtesy of O. Carmona, EPFL–LAMI).

3.4.5 Experimental results

The performance of the system has been tested by tracking the head of 10 persons sitting in front of the camera. Subjects were asked to move freely. The system was programmed to acquire a full SECAM resolution frame (640×480) each time it believed the head of a person to be centred in the image. A 110×130 "passport photo" sub-frame was then cropped at the image centre and stored for visual inspection. The sub-frame represents 5% of the image area. Its size has been computed from the average head size (75×90) by adding to each side the intrinsic system accuracy of 10 pixels (the radius of the inner circle of the retina) plus another 10 pixels of tolerance.

Video Signal	Composite NTSC
Pixels	542 (H) $\times$ 492 (V)
Field of View	43° (H) $\times$ 32° (V)
Excursion	$\pm$ 40°
Speed	$500°/s$
Acceleration	$40.000°/s^2$
Dimensions	60 $\times$ 170 $\times$ 170 mm^3
Power	12 V DC, 0.5 A up to 1,2 A

Table 3.1: Characteristics of the Swiss Federal Institute of Technology Vision Sphere (EPFL–LAMI).

Figure 3.4: A “passport photo” sub-frame covering 5% of the image size is grabbed each time the system believes a head is centred in the field of view.

Out of 500 images acquired from the 10 persons, 446 (89%) turned out to represent the head as expected (Figure 3.4). Three error typologies were found. In the first, subjects move outside the range of operation of the camera. In the second, subjects move too quickly, causing the frames to be blurred. The third typology is erroneous head detection, which normally happens in the presence of sharp geometrical patterns with contrasting lines moving in the image (Figure 3.5). This type of failure is partially corrected by the motion detection stage and is due to the simplicity of the head model employed (represented by 24 filter responses at a single point).

3.5 Final remarks

In the present Chapter we have introduced the *Saccadic Search* strategy through an application to Real-time Head Detection and Tracking. A simple motion detection stage prompts the computation of Gabor filters responses on the points of a retinotopic grid. The distance between such feature vectors and a precomputed reference is taken as a measure of the saliency of the corresponding image regions. A saccade is then performed that displaces the focus of attention to the point that minimises such distance. To the best of the author's knowledge, this constitutes the first attempt to apply the Gabor decomposition in such an active-vision task [93, 92].

The system described displays a certain robustness to scale changes, a good recovery speed and the ability to effectively discriminate between a head and other objects, such as for example a hand waving in front of the camera. Robustness could have been increased

Figure 3.5: The top two images show errors due to blurred frames. The two at the bottom represent errors induced by sharp geometric patterns or head-shaped objects (the helmet).

by implementing a simple contrast test to rule out geometric patterns. Also, a more descriptive model of the search target could have been used in a post-processing stage to give the system feedback about the fact that the object being tracked actually is a head. This second level of modelling will be introduced in the framework of Facial Features Detection in Chapter 5. In general, target modelling also benefits from a more powerful classifier. In the rest of this Thesis, Support Vector Machines are employed for this purpose. An introduction to SVM classifiers is provided in the following Chapter.

4 Support Vector Machines

4.1 Why Support Vector Machines?

The use of the *Saccadic Search* strategy for the localisation of complex deformable objects such as facial features requires a more articulate description of the search target than is provided by the simple average response vector employed in the Head Detection system presented in Chapter 3. A reliable classifier is also required for high-performance Face Authentication, that is discussed in Chapter 6.

Out of the wide variety of classifiers that has been made available by research on classification and learning theory, we have decided to rely on Support Vector Machines (SVMs). The main advantages of SVMs for our purpose are the fact that they do not require a large training set and that they offer an easy way to control the complexity of the decision surface through the choice of the kernel function (see Section 4.7). Apart from this, however, the choice of SVM classifiers is widely arbitrary. In principle, there is no reason why *Saccadic Search* strategy and the authentication algorithm presented in this Thesis should not work in conjunction with any other reasonably powerful classifier. On the contrary, as will become evident at the end of Chapter 6, the low-level signal processing employed in the *Retinal Vision* paradigm (Chapter 2) turns out to be so descriptive that the use of SVMs might almost appear as "overkill", at least as far as authentication is concerned.

Support Vector Machine classifiers constitute, in their own right, a wide research topic, a complete discussion of which lies beyond the scope of the present Thesis. Extensive tutorials can be found in [15] and [81]; the "classical" reference [19] provides a concise introduction. A presentation of the theory underlying SVMs can be found in [103]. For a good perspective of the most recent developments and research trends, we refer the reader to [82].

In the present Chapter we provide a quick, concise overview of the principles of SVM classifiers, based on the sources quoted above. While it has no pretence at being complete, this introduction should give the reader the grasp of Support Vector "machinery" that is required for the rest of this Thesis.

4.2 Classifiers and training

Following a standard notation, we describe a generic two-class classifier on an N-dimensional (real) feature space $\mathbb{R}^N$ as a family of functions

$$\{f_\alpha : \alpha \in \Lambda\}, f_\alpha : \mathbb{R}^N \to \{\pm 1\} \tag{4.1}$$

This notation expresses the fact that a given classifier (e.g. a Linear Perceptron) can implement a certain type of decision functions f_α (e.g. hyperplanes). Which of these functions actually is implemented depends on a set of tunable parameters represented by the index α (e.g. the weights of the connections). The training process has the purpose of selecting a suitable value for α in the allowed parameter range specified by Λ. The values ± 1 are conventionally selected to indicate the two classes, also called the "accept" $(+1)$ and the "reject" (-1) class.

Training is performed on a set of *training examples* for the two classes,

$$(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_l, y_l) \tag{4.2}$$

where $\vec{x}_i$ is a vector in $\mathbb{R}^N$ and $y_i \in \{\pm 1\}$ is its corresponding label. We assume that each training example is drawn from a distribution $P(\vec{x}, y)$ that specifies the dependency the classifier is required to model.

The aim of the training process is to identify the set of parameters $\bar{\alpha}$ that minimises the expectation under P of a suitable loss function $L(f_\alpha(\vec{x}), y)$ that expresses the cost of a misclassification:

$$E_P(L) = \int L(f_\alpha(\vec{x}), y)\, dP(\vec{x}, y). \tag{4.3}$$

This formulation is quite general, as it takes into account the fact that the two possible types of error, i.e. False Acceptance and False Rejection, will in principle have different costs depending on the application. A company baking biscuits, for instance, might find it preferable to discard good product rather than package a broken biscuit; a bank, on the contrary, might be willing to accept a certain amount of fraudulent transactions rather than deny a legitimate client access to telematic services to which he or she is entitled.

However, it is here convenient to restrict ourselves to the case in which the two types of errors are equally undesirable. In this case, training the classifier amounts to minimising the expected number of errors under P, a quantity that is known as the *risk*:

$$R(\alpha) = \int \frac{1}{2}|f_\alpha(\vec{x}) - y|\, dP(\vec{x}, y). \tag{4.4}$$

The resulting training algorithm, and indeed the classifier that we will be inclined to use, depends at this point on the type of assumptions that we are willing to make about the distribution function $P(\vec{x}, y)$. If, for instance, there are reasons to believe that the distributions $P(\vec{x}, +1)$ and $P(\vec{x}, -1)$ are Gaussians the optimal solution, i.e. Nearest

Neighbour classifiers, is provided by Bayesian decision theory. In the case that no parametric model is available for P, Equation (4.4) can be minimised only by resorting to some *induction principle*.

The most direct and simple approach consist in minimising the *empirical risk*

$$R_{emp}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} \frac{1}{2} |f_\alpha(\vec{x}) - y|, \tag{4.5}$$

which represents the fraction of errors over the points of the training set. However, this is not guaranteed to produce a good generalisation ability if the number of training examples is small with respect to the "capacity" of the classifier, intended as some measure of the complexity of the decision functions it can generate (e.g. the number of adjustable parameters or the number of neurons in a Neural Network). This loss of generalisation ability due to an "oversized" classifier is known as *over-fitting*. Although a number of regularization techniques has been devised to overcome this problem [13], they often find their place only as afterthoughts in traditional Neural Network approaches. One of the basic ideas of the Statistical Learning Theory developed by V. Vapnik and co-workers consists in incorporating the control of the complexity of the classifier in the induction principle, which is accomplished by the *Structural Risk Minimisation* principle.

4.3 Structural Risk Minimisation

The Structural Risk Minimisation induction principle is based on an upper bound on the risk which is independent of the probability distribution $P(\vec{x}, y)$. For the learning problem stated above, for any $\alpha \in \Lambda$ and $l > h$, the following inequality holds [103] with probability $1 - \eta$:

$$R(\alpha) \leq R_{emp}(\alpha) + \varphi\left(\frac{h}{l}, \frac{\log(\eta)}{l}\right). \tag{4.6}$$

The *confidence term* φ is defined as

$$\varphi\left(\frac{h}{l}, \frac{\log(\eta)}{l}\right) = \sqrt{\frac{h\left(\log\frac{2l}{h} + 1\right) - \log\frac{\eta}{4}}{l}}. \tag{4.7}$$

The parameter h is called the Vapnik-Cervonenkis (VC) dimension of the family of functions $\{f_\alpha\}$; l is the size of the training set.

The VC dimension is a measure of the capacity of the family of functions, in a sense that will presently be specified. Given a set of l points in $\mathbb{R}^N$, we consider all the possible 2^l assignments of each of these points to the "accept" and "reject" classes. Now, if for each such assignment there exists a decision function $f_{\bar{\alpha}}$ that respects the division among the two classes, we say that the set of points is *shattered* by the set of functions. The VC dimension of the set of functions $\{f_\alpha\}$ is defined as the maximum number of points that it can shatter. Note that for a set of functions to have VC dimension l it is sufficient that it will shatter one particular set of l points.

The probabilistic bound (4.6) is far from being tight. Indeed, there exist classifiers (such as KNN classifiers) that have an infinite VC dimension but in practice have a good generalisation ability. Nevertheless, the bound does indicate that, in order to obtain a good generalisation ability, one should be able to control both the empirical risk $R_{emp}(\alpha)$ and the complexity of the classifier as measured by $h(\{f_\alpha : \alpha \in \Lambda\})$. The empirical risk depends on the particular decision function chosen and can be minimised by appropriately tuning the parameter set α. The VC dimension depends on the set of decision surfaces available to the classifier, which can be controlled by restricting the choice of α to a subset of Λ. In particular, one can construct a series of classifiers of increasing complexity by introducing the series of nested sets $S_k = \{f_\alpha | \alpha \in \Lambda_k\}$, so that

$$S_1 \subset S_2 \subset \ldots \subset S_n \subset \ldots \tag{4.8}$$

and therefore

$$h_1 \leq h_2 \leq \ldots \leq h_n \leq \ldots . \tag{4.9}$$

For a given set of training data $(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_l, y_l)$ the *Structural Risk Minimisation* principle chooses the function f_{α_n} in the subset S_n for which the right end side of the risk bound (4.6) is minimal. That is to say that a balance is struck between the search for a good approximation of the data and the complexity of the family of functions used for such approximation.

4.4 Linear Support Vector Machines

Suppose we are given a linearly separable set of observations

$$(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_l, y_l), \tag{4.10}$$

meaning that there exists an (oriented) hyperplane

$$f(\vec{x}) = \vec{w} \cdot \vec{x} + b = 0 \tag{4.11}$$

such that $f(\vec{x}_i) > 0$ if and only if $y_i = +1$. Among all such hyperplanes, we define the *Optimal Separating Hyperplane* (OSH) as the one which has the maximal distance d_{max} from the closest example. This requirement alone only defines $f(\vec{x})$ in Equation (4.11) up to a multiplicative factor. In order to remove this ambiguity, we define a *canonical form* for the OSH by rescaling $\vec{v}$ and b so that $\|w\| = 1/d_{max}$. It follows that

$$\begin{cases} \vec{w} \cdot \vec{x}_i + b \geq +1 & \text{if } y_i = +1, \\ \vec{w} \cdot \vec{x}_i + b \leq -1 & \text{if } y_i = -1 \end{cases} \tag{4.12}$$

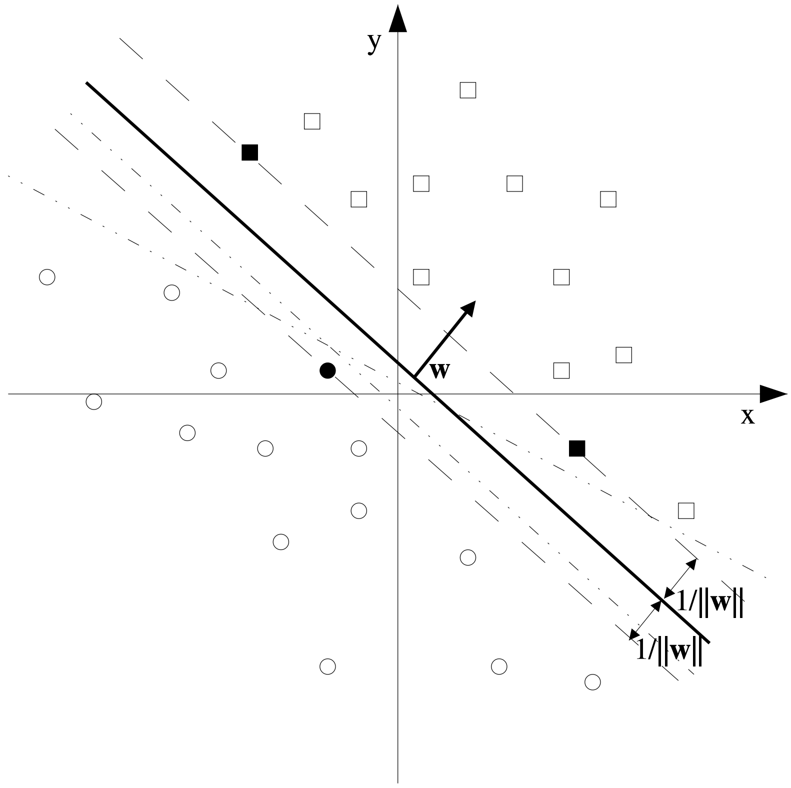


Figure 4.1: A linear Support Vector Machine classifier in $\mathbb{R}^2$ (separable case). Training examples for the (-1) and $(+1)$ classes are represented as circles and squares respectively. Filled symbols denote the support vectors. The thick, slanted line represents the Optimal Separating Hyperplane. The dash-dotted lines are examples of suboptimal separating hyperplanes that do not maximise the margin.

and that the projections of the closest examples from the two classes along the direction of $\vec{w}$ are at least $2/\|w\|$ apart (Figure 4.1). Support Vector Machines assign data to the "accept" $(+1)$ or "reject" (-1) classes based on the (canonical) Optimal Hyperplane decision function $f(\vec{x}) \geq 0$.

From a practical point of view, the Optimal Separating Hyperplane can be determined by minimising the functional

$$\tau(\vec{w}) = \frac{1}{2}\|\vec{w}\|^2 \tag{4.13}$$

subject to the constraints (4.12). This is a convex optimisation problem. A convenient way of accounting for the inequality constraints consists in introducing a Lagrangian

$$L(\vec{w}, b, \vec{\alpha}) = \tau(\vec{w}) - \sum_{i=1}^{l} \alpha_i [y_i(\vec{w} \cdot \vec{x}_i + b) - 1], \tag{4.14}$$

where the α_i are non-negative multipliers. The solution of the constrained optimisation problem then corresponds to the saddle point that maximises (4.14) with respect to the α_i and minimises it with respect to $\vec{w}$ and b. The conditions on the partial derivatives at the saddle point:

$$\frac{\partial L(\vec{w}, b, \vec{\alpha})}{\partial \vec{w}} = 0, \quad \frac{\partial L(\vec{w}, b, \vec{\alpha})}{\partial b} = 0 \tag{4.15}$$

lead to the two equalities

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{4.16}$$

and

$$\vec{w} = \sum_{i=1}^{l} \alpha_i y_i \vec{x}_i. \tag{4.17}$$

The normal vector to the OSH therefore is a linear combination of the training examples (note that while the solution to the convex problem is unique, the coefficients α_i need not be unique, since the training vectors are not constrained to be linearly independent). The Kuhn-Tucker conditions [56, 58] guarantee that, at the saddle point, the following relation holds:

$$\alpha_i [y_i(\vec{w} \cdot \vec{x}_i + b) - 1] = 0, \quad i = 1, \cdots, l. \tag{4.18}$$

Therefore, the non-zero multipliers α_i correspond to training examples for which the equality sign applies in the constraints (4.12), i.e. those lying exactly at the margin. Such training examples are called the *Support Vectors* and will be indicated as $\vec{s}_j,\ j = 1, \ldots, n$. The equation of the Optimal Separating Hyperplane (4.11) can therefore be rewritten explicitly in terms of the Support Vectors, leading to the following expression for the decision function:

$$f(\vec{x}) = \sum_{j=1}^{n} \alpha_j y_j \vec{s}_j \cdot \vec{x} + b \geq 0. \tag{4.19}$$

The training examples $\vec{x}_i$ other than the Support Vectors do not appear in this expression. They automatically satisfy the constraints (4.12) and the solution would be unaffected if they were removed from the training set.

The sparseness of the solution in terms of the training vectors can be used in a leave-one-out argument to obtain an upper bound on the probability of test errors. Namely, the *expected* probability of test errors can be shown to be bounded by the *expectation* of the number of Support Vectors obtained during training on a training set of size l, divided by $l - 1$.

4.5 SVMs and Structural Risk Minimisation

The relation between the *Structural Risk Minimisation* induction principle and Support Vector Machine classifiers is clarified by the following bound on the VC dimension of classifiers whose decision functions are hyperplanes in $\mathbb{R}^N$.

Let R be the radius of the smallest ball B_R in $\mathbb{R}^N$ containing the training examples $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_l$, and let

$$f'_{\vec{w},b}(\vec{x}) = \operatorname{sign}(\vec{w} \cdot \vec{x} + b) \tag{4.20}$$

represent a canonical hyperplane decision function in $\mathbb{R}^N$. For the VC dimension h of the set $\{f'_{\vec{w},b} : \|\vec{w}\| \leq A\}$, the following bound holds [15]:

$$h < R^2 A^2 + 1. \tag{4.21}$$

By maximising the margin, i.e. minimising $\|\vec{w}\|$, we therefore minimise the confidence term in the risk bound Equation (4.6). So far, we have made the assumption that the data are linearly separable, i.e. $R_{emp} = 0$; this condition will be lifted in the following Section.

4.6 Soft margin hyperplanes

In the case that no hyperplane decision function exists that correctly separates the vectors of the training set, an Optimal Separating Hyperplane can still be found by making allowance for a few classification errors.

From the formal point of view, this is accomplished by introducing a set of l *slack variables* $\xi_i \geq 0$ and by rewriting the bounds in Equation (4.12) in the following way:

$$\begin{cases} \vec{w} \cdot \vec{x}_i + b \geq +1 - \xi_i & \text{if } y_i = +1, \\ \vec{w} \cdot \vec{x}_i + b \leq -1 + \xi_i & \text{if } y_i = -1. \end{cases} \tag{4.22}$$

The number of classification errors can be controlled by modifying the objective function for the minimisation:

$$\tau'(\vec{w}) = \frac{1}{2}\|\vec{w}\|^2 + C \sum_{i=1}^{l} \xi_i \tag{4.23}$$

which is minimised subject to the new constraints. The solution is the same as in the separable case, the only difference being the existence of an upper bound on the coefficients α_i: we now have $0 \leq \alpha_i \leq C$. In intuitive terms, this means that no training vector is allowed to modify the solution "too much".

The term $\sum_{i=1}^{l} \xi_i$ represents an upper bound on the number of classification errors, that is on the *empirical risk* R_{emp} in the risk bound (4.6). However, the sum of the ξ_i

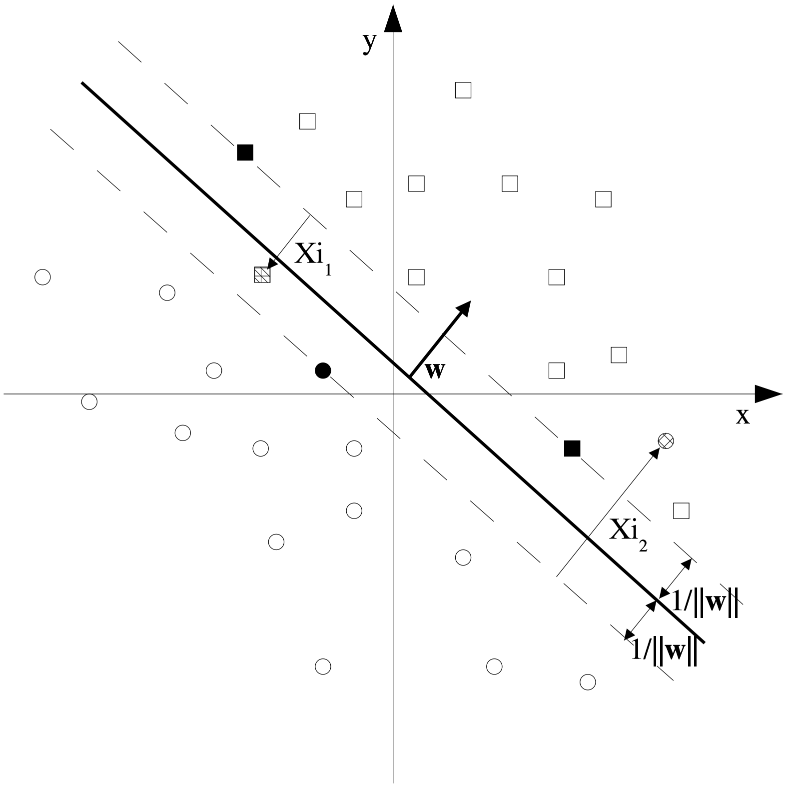


Figure 4.2: A linear Support Vector Machine classifier in $\mathbb{R}^2$ (non-separable case). Training examples for the (-1) and $(+1)$ classes are drawn as circles and squares respectively; the thick, slanted line represents the Optimal Separating Hyperplane. The hatched circle and square constitute two classification errors. Note how $\xi_1 \approx 1$ (i.e., ξ_1 "counts" one error), while $\xi_2 > 2$.

can significantly exceed the number of classification errors if many of the slack variables have large values (Figure 4.2). For this reason, Support Vector Machines only implement the Structural Risk Minimisation principle *approximately*. A study of the dependence of the solution on the constant C, which sets the balance between the control of the VC dimension (maximisation of the margin) and the minimisation of classification errors, can be found in [73]. However, from the practical point of view, classification errors can be quite rare when the number l of training examples is very small as compared with the dimensionality N of the vector space, which is generally the case for the applications considered in the present Thesis.

4.7 The nonlinear case

Support Vector Machine classifiers can afford more complex (nonlinear) decision functions by re-mapping input vectors to a higher dimensional space in which classification is performed using the Optimal Separating Hyperplane decision function.

The SVM optimisation problem only involves the dot products among the training vectors $\vec{x}_i$ (see Equations 4.13, 4.14 and 4.17). Therefore, the function $\Phi : \mathbb{R}^N \to \mathcal{H}$ that maps the vectors $\vec{x}_i$ onto the new space (a Hilbert space, in the general case) does not need to be given explicitly. One only needs to specify the dot product of any two image vectors $\Phi(\vec{x})$ and $\Phi(\vec{y})$ in $\mathcal{H}$ through a *kernel function* K defined over $\mathcal{C} \times \mathcal{C}$, where $\mathcal{C}$ is a compact subset of $\mathbb{R}^N$ that includes the training and test vectors:

$$< \Phi(\vec{x}), \Phi(\vec{y}) >:= K(\vec{x}, \vec{y}). \tag{4.24}$$

In order for this definition to be well-posed, K must satisfy the conditions of Mercer's theorem of Functional Analysis. More specifically, $K(\vec{x}, \vec{y})$ must be symmetric and continuous over $\mathcal{C} \times \mathcal{C}$. Furthermore, the integral operator over $L^2(\mathcal{C})$

$$(\mathcal{K}f)(\vec{y}) = \int_{\mathcal{C}} K(\vec{x}, \vec{y}) f(\vec{x})\, d\vec{x} \tag{4.25}$$

must be positive, meaning that

$$\int_{\mathcal{C}\times\mathcal{C}} K(\vec{x}, \vec{y}) f(\vec{x}) f(\vec{y})\, d\vec{x}\, d\vec{y} \geq 0 \tag{4.26}$$

for all $f \in L^2(\mathcal{C})$.

When these conditions are met, it is possible to find a map Φ of the input vectors onto a suitable Hilbert space $\mathcal{H}$ such that (4.24) defines a scalar product in $\mathcal{H}$.

The Optimal Separating Hyperplane in $\mathcal{H}$ can be written in terms of the input vectors in $\mathbb{R}^N$, giving the following expression for the decision function:

$$f(\vec{x}) = \sum_{j=1}^{n} \alpha_j y_j K(\vec{s}_j, \vec{x}) + b \geq 0. \tag{4.27}$$

This is the exact analogue of Equation (4.19), but for the fact that the dot product $\vec{s}_j \cdot \vec{x}$ in $\mathbb{R}^N$ has been substituted by the value $K(\vec{s}_j, \vec{x})$ of the kernel function.

Since the kernel function is in general nonlinear, the decision surface $f(\vec{x}) = 0$ is also nonlinear. Common choices for K are the family of kernel functions

$$K_d(\vec{x}, \vec{y}) = (\vec{x} \cdot \vec{y} + 1)^d, \tag{4.28}$$

that lead to polynomial classifiers, and the Gaussian kernel

$$K(\vec{x}, \vec{y}) = \exp\left(-\frac{||\vec{x} - \vec{y}||^2}{\sigma^2}\right) \tag{4.29}$$

that implements the same decision functions as Radial Basis Function (RBF) networks.

4.8 Note on implementation

The Support Vector Machine engine employed in the experiments reported in this Thesis was developed by the author and N. Capdevielle following the ideas in [47] and [70].

5 Facial Feature Detection

5.1 Why detect Facial Features?

The importance of the reliable detection of relevant structures, or features, in facial images is closely related to the growing demand for Face Authentication/Recognition systems, that often require Facial Feature Detection as a pre-processing stage. Many Face Authentication techniques (see for instance [96, 99]), including the algorithm presented in Chapter 6 of this Thesis, make a direct use of facial features. However, even authentication algorithms such as Eigenfaces [102] that make no explicit use of facial landmarks show a greatly improved performance if the images are normalised by applying suitable transformations that (approximately) compensate for variations in scale, pose, rotation and translation. Although registration techniques for facial images do not necessarily require the detection of the facial features [34, 48], knowledge of the position of the eyes and the mouth allows the direct estimation of a suitable affine transformation.

A variety of Facial Feature Detection algorithms has been proposed in the recent years. A detailed survey goes beyond the scope of the present work; the references we report are only meant to give the reader a feeling for the diverseness and the ingenuity of the existing approaches. Colour information has often proved a useful cue for fast detection [78, 94]. At a more abstract level of description, finite state automata driven by a genetic algorithm have been used to locate the eyes based on gray level statistics [40]. Hough-transform based approaches [6, 67] and model-based algorithms [7] rely on the shape properties of the facial features. Finally, in [96], orientation-tuned Gaussian filters arranged in a retinotopic grid have been employed in conjunction with iconic visual filters that are applied to the intensity values of the image.

Despite the considerable research effort, the detection of facial features remains a challenging problem. The good results achieved by applying the *Saccadic Search* strategy to eye and mouth detection constitute both a confirmation of the validity of the *Retinal Vision* paradigm and an advancement in this important field of Computer Vision. The material presented in this Chapter has been published in [87, 91].

5.2 Eye and mouth detection

The Facial Feature Detection procedure described in this Thesis consists in a saccadic exploration of the image. The saccadic pattern is generated on the basis of a-priori knowledge about the targets to be detected, which is supplied by *local* models of the eyes and the mouth. These encode the appearance of a facial feature (search target) as a vector of Gabor responses; they are conceptually equivalent to the average model vector employed for Head Detection and Tracking (Section 3.4.1), the only difference consisting in the introduction of Support Vector Machines classifiers. The retinotopic grid is initially placed at a random position in the image, and a Gabor response vector is extracted at all of its points. The similarity between such vectors and the signature of the looked-for facial feature is evaluated, and the "retina" is displaced to the point where the best match is found. Due to the retinotopic arrangement of sampling points, the search is automatically refined as the sensor converges to a likely target (Section 5.2.3). A more selective *extended model*, based on the complete set of Gabor responses extracted by centring the retina on an eye or on the mouth, is then used to refine the position of the match. Once one of the facial features has been found, a probabilistic model is used to perform a saccade to the assumed location of one of the others.

A final evaluation stage selects the best set of facial features based on a competition scheme (Section 5.2.4).

5.2.1 Modelling the Facial Features

The *Saccadic Search* strategy requires the use of two levels of modelling for each of the targets [91]. The coarser, or *local*, model is computed for each facial feature from the vector of Gabor responses extracted at the location of that feature in the images of the training set. The finer, or *extended* model is obtained by placing the whole retina at the location of the facial feature on the training images (Figure 5.1) and collecting the set of Gabor filter responses from all of the retinal points. These Gabor features are then arranged into a single vector.

In both cases, the example vectors for each facial feature are complemented with negative examples and used to train a Support Vector Machine classifier (see Chapter 4). In our experiments we have used a training set consisting of 202 frontal images of 13 subjects, that was extracted from the first version of the M2VTS database (Appendix B). This allowed us to obtain 202 positive examples for each facial feature. Negative examples have been collected by extracting Gabor vectors (or, respectively, positioning the retina) at 10 random locations in each training image. Also, right eyes and mouths have been included as negative examples in the models of the left eye, and so on.

The training set is also used to estimate the expected relative position of each facial feature relative to one of the others.

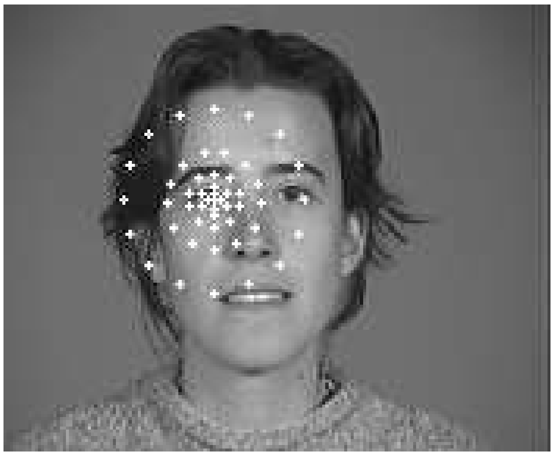


Figure 5.1: The extended model of a facial feature is obtained by positioning the retina at that feature in the images of the training set.

5.2.2 Selecting the kernel function

In order to determine the most appropriate SVM kernel function (see Section 4.7) for the local and extended models, we used the 349 images from the remaining 24 subjects in the M2VTS database as a probe set. Probe images were used to feed the classifiers with the appropriate set of Gabor responses extracted at the location of the facial feature they were supposed to model, together with a set of negative examples they were supposed to reject. The negative examples included, for each classifier, the remaining two facial features plus ten random points per image.

Since the optimisation process depends on K, training had to be performed anew for each choice of the kernel. In order to speed up the procedure, the Gabor response set for the extended models was "pruned" by eliminating the high frequency channels at the periphery of the retina and the low frequency channels in the region of the fovea.

Due to the wider availability of negative examples during the training process, all classifiers showed a higher tendency towards making errors when classifying a positive probe (false rejection) rather than a negative probe (false acceptance). In order to compensate for that, we modified the decision function (4.27) with the introduction of a threshold τ:

$$\sum_j \alpha_j y_j K(\vec{s}_j, \vec{v}) + b \geq \tau. \tag{5.1}$$

We adjusted the threshold in order to obtain an Equal Error Rate (EER) on the probe set; we also recorded the value of τ both at the EER and at the point at which the false rejection goes to zero. The dependence of these quantities on the kernel type is displayed in Table 5.1, 5.2 and 5.3 respectively. It must be emphasised that, since our algorithm employs a competition scheme rather than a fixed decision surface (see Section 5.2.4), the behaviour of the classifiers as the decision surface is shifted by varying τ is of greater

EER (%)	**Kernel type**			
Local Model	Linear	Poly deg. 2	Poly deg. 3	**Poly deg. 4**
Left eye	11 ± 2	11 ± 2	11 ± 2	$\mathbf{10 \pm 2}$
Right eye	13 ± 2	11 ± 2	11 ± 2	$\mathbf{11 \pm 2}$
Mouth	6 ± 1	8 ± 1	8 ± 1	$\mathbf{7 \pm 1}$
Extended Model	Linear	Poly deg. 2	Poly deg. 3	**Poly deg. 4**
Left eye	0.5 ± 0.4	0.3 ± 0.3	0.3 ± 0.3	$\mathbf{0.3 \pm 0.3}$
Right eye	0.8 ± 0.5	0.5 ± 0.4	0.4 ± 0.3	$\mathbf{0.3 \pm 0.3}$
Mouth	2.1 ± 0.8	1.8 ± 0.7	1.7 ± 0.7	$\mathbf{0.3 \pm 0.3}$

Table 5.1: The table displays the value of the EER of the local and extended models as a function of the kernel type employed. Notice how especially the extended mouth model benefits from a higher order kernel. Differences between the two eye models are not statistically significant. The uncertainty of the estimates has been computed based on the Binomial distribution.

interest to us than the precise value of the EER. A small value of τ both at the EER and at the point where the false rejection becomes zero indicates that the outputs of the SVM classifier are "appropriately grouped". Also, it is of interest to our algorithm that the classifiers trained on the three facial features have approximately the same behaviour. As evidenced by Table 5.2 and 5.3, these requirements are best met by classifiers based on a fourth order polynomial kernel of the form

$$K(\vec{v}, \vec{w}) = (1 + \vec{v} \cdot \vec{w})^4. \tag{5.2}$$

We therefore decided to employ this kernel function in both the local and the extended models. We can here observe that the local models are much cruder than the extended models, which turn out to be approximately 30 times more accurate. This is in line with our expectations, since local models describe a facial feature based only on a concentric neighbourhood of a single point (a receptive field), whereas each extended model employs the entire set of receptive fields that make up the retina.

5.2.3 The Saccadic Search strategy

Facial Feature Detection begins with the retinotopic sensor randomly placed on an image. Search is initially aimed at one randomly selected feature among the three possible targets. The Gabor response vector $\vec{v}_{c\gamma}$ extracted at retinal point γ on circle c is rated by the local model of the selected target according to the output of its associated SVM classifier:

$$f_{loc}(\vec{v}_{c\gamma}) = \sum_j \alpha_j y_j K(\vec{s}_j, \vec{v}_{c\gamma}) + b. \tag{5.3}$$

The retina is subsequently centred at the position of its sampling point $s_{c\gamma}$ that maximises f_{loc}. This procedure is iterated until the centre s_{00} of the retina obtains the best

τ at EER	**Kernel type**			
Local Model	Linear	Poly deg. 2	Poly deg. 3	**Poly deg. 4**
Left eye	-5.0	-2.0	-2.0	$-\mathbf{1.9}$
Right eye	-7.0	-2.3	-2.0	$-\mathbf{1.9}$
Mouth	-3.5	-1.8	-1.7	$-\mathbf{1.6}$
Extended Model	Linear	Poly deg. 2	Poly deg. 3	**Poly deg. 4**
Left eye	-0.30	-0.15	-0.05	$-\mathbf{0.03}$
Right eye	-0.43	-0.30	-0.25	$-\mathbf{0.22}$
Mouth	-0.90	-0.82	-0.80	$-\mathbf{0.25}$

Table 5.2: The table shows the value of τ at the EER as a function of the kernel type, for both the local and the extended models. The value of τ steadily increases towards zero with kernels of higher order, meaning that the tendency towards false rejection is reduced.

score, meaning that a local maximum has been reached. Note that one advantage of this optimisation strategy is that the search automatically becomes finer as a local maximum is approached, since the artificial retina is denser at the centre (*fovea*) than at the periphery. As a consequence of this and of our experimental results, we think that, in the topology of the human retina, one of the effects of the acuity gradient existing between the peripheral and the foveal vision is that it permits to achieve fast convergence (*homing*) of the saccades.

After saccades have converged to a local maximum, the retinotopic grid is displaced in a pixel-by-pixel fashion to maximise the output f_{ext} of the more accurate extended model for the detected facial feature. Feature candidates with a score lower than the score of support vectors of the reject class ($f_{ext} < -1$) are discarded at this stage; the others are pushed onto a stack along with their score. After that, a saccade is performed to the most probable location of the next facial feature on the list of targets (in a cyclic fashion) and an attempt at detection is made directly with the corresponding extended model. If this succeeds, another saccade is performed based on the expected relative position of the eyes and the mouth to reach the third facial feature; otherwise, saccadic search is restarted at random to look for the feature for which the fewest candidates have been detected.

Saccadic search is continued until a complete set of facial features which has a very high score (see Section 5.2.4) is found or until the retina has been centred on the maximum allowed number of different locations. Examples of the resulting saccadic patterns are given in Figure 5.2.

5.2.4 Selecting the best feature set

The final decision on which feature candidates constitute the best facial feature set is based on a competition scheme. The left eye, right eye and mouth candidates present in the stack are grouped in all the possible ways and are assigned a score according to the

τ at FR = 0	**Kernel type**			
Local Model	Linear	Poly deg. 2	Poly deg. 3	**Poly deg. 4**
Left eye	−14.0	−5.5	−5.0	**−5.0**
Right eye	−19.0	−8.0	−7.0	**−6.5**
Mouth	−8.0	−9.0	−8.0	**−6.0**
Extended Model	Linear	Poly deg. 2	Poly deg. 3	**Poly deg. 4**
Left eye	−0.5	−0.2	−0.1	**−0.2**
Right eye	−0.7	−0.6	−0.5	**−0.5**
Mouth	−1.4	−1.3	−1.2	**−0.4**

Table 5.3: The table shows the maximum value of τ for which FR = 0 as a function of the kernel type, for both the local and the extended models. Although this measure is sensitive to the presence of outliers, its marked tendency to increase with the kernel order indicates that the fluctuations of the output of the classifiers become smaller with a higher order kernel.

expression

$$\text{Score}(i, j, k) = \mathcal{S}(i)\mathcal{S}(j)\mathcal{S}(k). \tag{5.4}$$

For each candidate feature i, $\mathcal{S}(i)$ is computed by transforming the output $f_{ext}(i)$ of the corresponding extended SVM model through a sigmoid function

$$\sigma_f(x) = \frac{1}{2}(\tanh \kappa x + 1) \tag{5.5}$$

that maps the scores to the interval $(0, 1)$. The steepness factor κ is such that SVM scores in the interval $(-1, 1)$ fall in the linear region. We chose to combine the three scores through a product rule because of its higher sensitivity to the presence of a single misdetection.

If none of the possible feature sets gives a satisfactory score, saccadic search is restarted from scratch for a second attempt (we have allowed a maximum of four independent trials in our experiments). Sets of fewer than three facial features are considered if no candidates have been detected for one or more of the search targets.

5.3 Experimental results

5.3.1 Sensor configuration

The retinotopic grid we used consists of 49 points arranged in 5 concentric circles. The centre of the circles is also sampled, for a total of 50 points. In our experiments, we fixed the parameters in the Equation (2.5) of the retinotopic grid as follows: we set the radius of the innermost circle r_{min} to 3 pixels and we adjusted δ_r so that the outermost radius

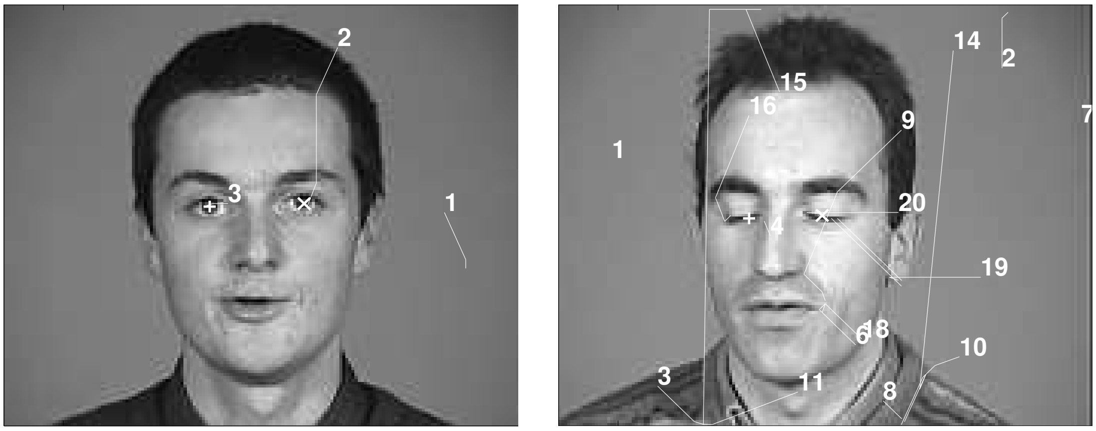


Figure 5.2: Examples of the saccadic patterns generated during eye detection. The + and × signs denote the best match with the right and left eye models respectively. Numbers identify successive starting points for saccades. Eye detection over the image on the left required 51 fixations, the great majority of which are part of search refinement (only saccades are displayed here). Note how the saccadic pattern starting from (1) ended up at a local maximum that was considered uninteresting and therefore discarded. A random restart (2) then led to detection of the left eye, after which saccadic search resumed (3) near the location of the right eye. Information from the surrounding facial region (eyebrows, outline of the eye orbit) allows correct detection of the eyes even in the case they are shut (right). During this trial 99 fixations were made. These images were generated using simple local and extended eye models based on the Euclidean distance from an average vector (see Section 3.4.1).

of the retina is 30 pixels, which is about the average inter-eye distance in the images we used for testing.

Receptive fields are modelled by computing a vector of 30 Gabor filter responses at each point of the retina. The filters are organised in 5 frequency channels and 6 equally spaced orientation channels. Filter wavelengths span the range from 4 to 16 pixels in half-octave intervals. The size of the receptive fields is constant; therefore, their overlap increases towards centre of the retina, where the grid points are closer to each other. Since the responses from widely overlapping filters would no longer be independent, we found it convenient to reduce the number of sampling points from 16 on the outermost circle to 4 on the innermost circle (Figure 2.1).

The configuration of the retinotopic grid and of the Gabor filters has been kept unchanged for the Face Authentication experiments described in Chapter 6.

For increased speed, Gabor feature vectors are memorised to avoid computing them more than once; furthermore, it is assumed that the magnitude of the response of a Gabor filter of wavelength λ does not change significantly in a circular neighbourhood of radius $\lambda/4$ around the centre of the filter. The filter response in the entire neighbourhood is therefore approximated with the value it assumes at its centre (for Face Authentication, the exact values are employed). We have confirmed the validity of this approximation and observed that it does not affect the accuracy of the algorithms negatively, while it significantly improves the speed.

5.3.2 Detection performance

Facial Feature Detection tests were run on a total of 349 images from the M2VTS database and on 2388 images from the XM2VTS database (we also included in the test set the images of a few subjects who are not part of the official XM2VTS distribution). Training was performed in both cases on the training set from the M2VTS database (see section 5.2.1); all parameters were left unchanged for tests on the M2VTS and XM2VTS databases.

Detection performance was evaluated by visual inspection. Experimental results show reliable eye and mouth detection performance, with comparable results on both databases. The error rate for the M2VTS test set is 0.7% for each eye and 0.0% for the mouth, which is always detected correctly. Therefore, 5 images out of 349 present a misdetection of one eye; the three facial features are correctly detected in all the others (98.6%). Error rates on the XM2VTS database are 1.0% for each eye and 1.3% for the mouth. This last figure reflects the higher variability in the appearance of the mouth due to changes in expression. Completely erroneous detection has occurred in 0.3% of the images (i.e. 6 images out of 2388); in 99.5% of the cases at least two features have been detected correctly. Correct detection of both the eyes and the mouth has been achieved on 97.4% of the test set. The system appears to be robust to the presence of eyeglasses (Figure 5.3, 1st row), partial occlusion (2nd row) and even significant pose changes (3rd row). It also generalises quite well to the presence of beards (Figure 5.3, 4th row) and of non-Europeans (5th row), that are not represented in the training set.

Facial Feature Detection results on the M2VTS database have been published by another research group in [53]. An 86.5% success rate for simultaneous detection of eyebrows/eyes, nostrils/nose and mouth over 37 frontal images from the database is reported. Thus our results are significantly better [87].

Figure 5.3: Examples of Facial Feature Detection (XM2VTS database).

6 Face Authentication

6.1 A brief introduction

The spreading of Internet commerce, tele-banking and similar services which require privileged and possibly remote access to resources on the part of accredited users has put a strong emphasis on the development of reliable biometrics for person authentication. A wide choice of techniques has been proposed to meet this demand, ranging from fingerprints to voice analysis all the way to complete multi-modal systems. These get an edge in robustness through the fusion of several experts [11]. However, despite its intrinsic complexity, face-based authentication, admittedly a "natural" choice, still remains of particular interest because of its nature, which is perceived as psychologically and/or physically "non-invasive" by users.

A variety of Face Authentication techniques has been developed in recent years. While certain approaches such as Eigenfaces [86, 102] and Fisherfaces [57] are entirely statistic in nature, others are influenced by theories of human perception [18]. Several appearance-based algorithms show some degree of biological analogy in that they make use of *retinotopic sampling* [95, 99] or of *Gabor filters*, which as we know model the responses of the simple cells in the visual cortex [69].

The use of the Gabor decomposition for face recognition purposes has been introduced by C. v. d. Malsburg in the context of elastic graph matching [55, 68]. This technique has undergone further independent developments in the framework of Face Authentication [11, 28]. Selective weighting of single Gabor responses using Fisher's Linear Discriminants has been shown to be more effective than elastic deformation of the graph [28]. Also, a mixture of experts approach has been developed which delegates a first decision to each node of the grid and achieves authentication through the fusion of the different experts' opinions [1]. The use of the Gabor decomposition on sparse graphs therefore has a well-established role in face recognition and authentication.

Following the *Retinal Vision* paradigm, the present work pioneers the use of retinotopic sampling applied to the Gabor decomposition. The experimental results presented over the two largest authentication-oriented facial image databases available confirm the validity of this approach. Our results have been published in [87, 90].

6.2 Authentication with three localised experts

Face authentication is performed using a multiple expert approach. Three machine experts are employed to independently authenticate each client. Two of them focus on the eye regions, the third on the mouth region; this choice is motivated by perceptive studies showing that certain parts of the face, and in particular the area surrounding the eyes, tend to be more relevant than others in the characterisation of a face [50].

The input to the machine experts is obtained from the final stage of the Facial Feature Detection algorithm. The latter provides the three sets of Gabor responses extracted with the retinotopic sensor centred on the eyes and the mouth of the subject (see Chapter 5); each of these feature sets can be seen as a vector in a separate copy of a feature space $\mathcal{G}$. We recall that the set of Gabor features employed for Facial Feature Detection consists of 30 filters arranged in 5 resolution (frequency) and 6 orientation channels. The retinotopic grid consists of 50 points; the dimensionality of the feature space $\mathcal{G}$ therefore is $30 \times 50 = 1500$.

The three machine experts, that in the following will be implemented by various types of classifiers, independently authenticate a client based on the feature vectors in $\mathcal{G}$; more precisely, each expert processes the Gabor response vector describing the facial feature of its competence. The final decision about the identity claim being true or false is obtained by fusing the opinions of the experts using a simple rule, and therefore is a map from $\mathcal{G} \times \mathcal{G} \times \mathcal{G}$ to the set $\{True, False\}$ of the truth values of an identity claim. This multiple expert procedure introduces a certain amount of robustness. We recall, however, that each machine expert independently authenticates the client based on a single copy of $\mathcal{G}$.

For the classification step we have experimented with different implementations of the experts, namely K Nearest Neighbours (KNN) classifiers with $\mathrm{K} = 1, 3, 5$ and SVM classifiers with various choices of the kernel function (see Sections 6.4 and 6.5).

All classifiers are used in the two-class context of Face Authentication; examples from the client represent the "accept" class and examples from all the training impostors are grouped into the "reject" class. KNN experts yield a discrete score, with $2\mathrm{K} + 1$ possible values equal to the signed difference between the number of positive and negative examples that appear among the K nearest neighbours; this in a way gives a measure of "how sure" the expert is feeling about its output. When $\mathrm{K} = 1$ (Nearest Neighbour), this reduces to a binary score of $+1$ (acceptance) or -1 (rejection). In the case of SVM experts, we take the value of the left hand side of the decision function Equation (4.27) as a graded score:

$$f(\vec{x}) = \sum_{j=1}^{n} \alpha_j y_j K(\vec{s}_j, \vec{x}) + b. \tag{6.1}$$

This makes sense, since the sign of $f(\vec{x})$ gives Optimal Separating Hyperplane classification, while its absolute value represents a normalised distance from the hyperplane that can be taken as a measure of the degree of "confidence" of the classifier in making the decision.

In the case of KNN classifiers, expert fusion is achieved by considering the sign of the total score (which amounts to a majority vote for $\mathrm{K} = 1$). When fusing together the output of SVM experts, we have found it convenient to use a nonlinearity to limit the influence of any single expert $\mathcal{E}_j$ on the global authentication score $\mathcal{S}$ by gradually saturating the scores outside the $(-1, 1)$ interval. The decision function is therefore expressed by the inequality

$$\mathcal{S}(I) = \frac{\sum_{j=1}^{3} \tanh\left(\kappa\, \mathcal{E}_j(I)\right)}{3} \geq \tau \tag{6.2}$$

where $\mathcal{E}_j(I)$ stands for the score assigned by expert $\mathcal{E}_j$ to the identity claim I according to Equation (6.1), κ is a constant and τ can be varied to obtain FA/FR curves and determine the Equal Error Rate (EER). These two concepts are introduced in the following Section.

Optimal performance is already achieved using linear SVM classifiers, which separate the client and impostor classes with a hyperplane.

6.3 Evaluating authentication algorithms

Identity verification algorithms are prone to two complementary types of errors, known as False Acceptance (FA) and False Rejection (FR). False Acceptance consists in failing to detect a false identity claim (or impostor claim), which results in an impostor accessing the system. False Rejection stands for refusal of a true identity claim (or client claim), with the consequence that a legitimate client is denied access to the system. These two errors are empirically quantified by the ratios

$$\mathrm{FA} = \frac{\text{Number of erroneously accepted claims}}{\text{Total number of impostor claims}} \tag{6.3}$$

and

$$\mathrm{FR} = \frac{\text{Number of erroneously rejected claims}}{\text{Total number of client claims}}. \tag{6.4}$$

Depending on the application of interest, one of these two typologies of errors may be more penalising than the other.

In the absence of a tunable threshold, the performance of a system is characterised by the (fixed) values of its False Acceptance and False Rejection. This is the case for our system when KNN experts are used. In the case of SVM experts, the decision on an identity claim I is taken by comparing the global authentication score $\mathcal{S}(I)$ against a threshold τ (see Equation 6.2). This is a standard choice for authentication systems that make use of a continuous score function. The threshold can be used to bias the system towards False Acceptance or towards False Rejection, thus determining its operation point according to the demands of the specific application.

It is customary, as we did here, to assign high scores to claims that are likely to be true and should be accepted, and conversely to assign low scores to claims that, in the opinion of the system, are likely to be false and should therefore be rejected. With this

KNN Experts	FA	FR	$\sqrt{\mathrm{FA} \cdot \mathrm{FR}}$
K=1	1.1%	2.0%	1.5%
K=3	0.5%	4.0%	1.3%
K=5	0.2%	8.0%	1.4%

Table 6.1: False Acceptance and False Rejection with KNN experts (M2VTS database).

convention, False Acceptance decreases as the value of τ is increased. As soon as τ is larger than the largest score attributed to an impostor claim, FA becomes zero because all impostors are rejected. At this point, however, unless perfect authentication is achieved, a few of the scores assigned to the client claims will also lie below the threshold, implying a positive value of the False Rejection rate. Increasing τ even more eventually results in FR = 1, since all the client claims will fall below the threshold and be rejected. False Acceptance and False Rejection can be plotted against τ, thus giving the so-called FA and FR curves. Let $\bar{\tau}$ be the value of the threshold for which these two curves intersect. The Equal Error Rate (EER) is defined as

$$\mathrm{EER} = \mathrm{FA}(\bar{\tau}) = \mathrm{FR}(\bar{\tau}). \tag{6.5}$$

This is considered the single most significant figure of merit for an authentication system.

The False Acceptance and False Rejection curves are not readily comparable across different systems, since they depend on the parametrisation of the scores. A more convenient representation can be obtained by eliminating the threshold τ and plotting the False Rejection versus the False Acceptance. The resulting graph is known as FA/FR curve or ROC (Receiver Operating Characteristic) curve. The EER is given by the intersection of the FA/FR curve with the diagonal of the 1st quadrant.

For increased readability, the FA/FR curves reported in this Thesis have been plotted on a Normal Deviate scale [60] using the software provided by the American National Institute of Standards and Technology (NIST) [104].

6.4 Authentication on the M2VTS database

Face authentication experiments on the M2VTS database were conducted according to a “leave-one-out” rotation scheme also known as “Brussels Protocol” [11, 26, 30]. The database includes frontal images from 37 persons. Four series of 3–4 portraits each are available for each subject. These image series have been acquired on different occasions separated by a significant time interval (see Appendix B).

The protocol requires that three of the four image series be labelled as “training series”, the fourth being the “test series”. At each step, one person is entirely removed from the database to act as impostor. The 36 persons remaining act as registered clients; their respective training series are used for training the system, while the test series are set aside for client tests.

SVM Experts	EER
Linear	0.17%
Poly - deg 2	0.15%
Poly - deg 3	0.15%
Poly - deg 4	0.13%

Table 6.2: False Acceptance and False Rejection with SVM experts based on linear and polynomial kernels (M2VTS database).

A single image from the test series of each client is used to generate a client claim; in this way, 36 client tests are performed. Likewise, a single image from the impostor's test series is tested against all the clients to give 36 impostor tests. By rotating through all the 37 possible choices of the impostor and the 4 independent choices of the test series, a total of $37 \times 4 \times 36 = 5328$ client tests and the same number of impostor tests can be generated.

The use of the training series of the registered clients is not constrained by the Protocol. In the training procedure we adopted, each expert (independently of its implementation) uses the images in the training series of the client it should authenticate as positive examples. All of the images from the training series of the remaining 35 registered clients are used as negative examples.

Since part of the M2VTS database has been used as the training set for the Facial Feature Detection algorithm, we only present authentication results obtained by manually identifying the position of the eyes and the mouth of the subjects in both the training and the test images, so that the performance of the authentication algorithm itself can be gauged.

Experiments indicate that a very low error rate can be achieved with both SVM based and KNN based experts. Results are summarised in Tables 6.1 and 6.2; an example of client and impostor claims is reported in Figure 6.1. As can be observed, KNN classifiers have a marked tendency towards False Rejection, due to the low number of training images available for each client as compared to impostors. In the case of SVM experts, this has been compensated by tuning the threshold τ in Equation (6.2) so as to determine the EER (Table 6.2). To provide some form of comparison, we listed the geometric average error given by KNN experts in the third column of Table 6.1. Variations of the EER due to the choice of different kernel functions for the SVM classifiers are not statistically significant. Although one could argue that the lack of sizeable improvements using higher order kernels is due to the scarcity of training examples, the fact that remarkably low Equal Error Rates can be achieved by means of *linear* decision surfaces demonstrates the discriminating power of the (orientation) features employed. This is confirmed by the consistently good performance of linear SVM classifiers on the larger XM2VTS database (see Section 6.5).

Summing up, (linear) Support Vector Machine experts yield the best performance with an Equal Error Rate around 0.15%. The corresponding FA/FR curve is reported in Figure 6.2. For comparison, we refer the reader to published results on the M2VTS database obtained using Elastic Graph Matching (EER = 6.1%, Brussels Protocol) [28] and the Morphological Dynamic Link Architecture (EER = 3.7%, Brussels Protocol) [97]. Equal Error Rates of 5.4% for frontal image authentication and 3.1% for image sequence based authentication are reported in [61], where a warping and correlation algorithm is employed. Finally, an approach combining the Morphological Dynamic Link Architecture and Support Vector Machines is documented in [98] (EER = 2.4%, Brussels Protocol).

6.5 Authentication on the XM2VTS database

Face Authentication experiments on the XM2VTS database were carried out according to the standard test protocol named Lausanne Protocol (Configuration II) by the M2VTS Consortium [62]. This protocol specifies 200 identities to be used as clients and 70 identities for impostors. For each client, two sessions (4 images) are available for training; one is used for testing (2 images), and two images are left unused. All 8 images of each impostor are used to generate impostor tests. This gives a total of 400 client tests and $70 \times 8 \times 200 = 112'000$ impostor tests.

As was the case with the M2VTS database, we trained each expert using the available images from the client it is supposed to authenticate as positive examples, and the images of all the other registered clients as negative examples (training impostors).

Experiments are performed in an *entirely automatic* fashion, that is we let the Facial Feature Detection algorithm (which has not been trained on any image from the XM2VTS database) find the coordinates of the eyes and the mouth both in the training images and in the images of the test set. An Equal Error Rate (EER) of 0.50% has been achieved using linear Support Vector Machine experts (dash-dot line, Figure 6.3). This corresponds to two client tests failed out of 400.

This result can be further improved by adding the remaining two images of each client (that are left unused in the Lausanne Protocol) to the training set, which then contains six images of each registered subject. In that case system operation is described by the solid line in Figure 6.3. The Equal Error Rate (0.25%) corresponds to only one client test failed out of 400. Comparison of these results with those obtained on the M2VTS database by manually identifying the position of the eyes and the mouth (EER = 0.15%) shows that the use of automatic facial feature detection both in the training and in the test phase does not significantly degrade authentication performance. This is due both to the reliability of the Facial Feature Detection procedure and to the intrinsic robustness of the authentication algorithm, which is achieved through the use of multiple training images and of a mixture of experts approach. Independent results on face authentication on the XM2VTS database are presented in [54], where an EER of 5% is obtained using the Morphological Dynamic Link Architecture (Lausanne Protocol, Configuration II).

(a) Training set for subject FL (images collected during 3 acquisition sessions).

(b) A correctly authenticated image of FL (4th session)

(c) A plausible impostor for FL that has been rejected by the system.

Figure 6.1: Example of a client claim and an impostor claim from the M2VTS database.

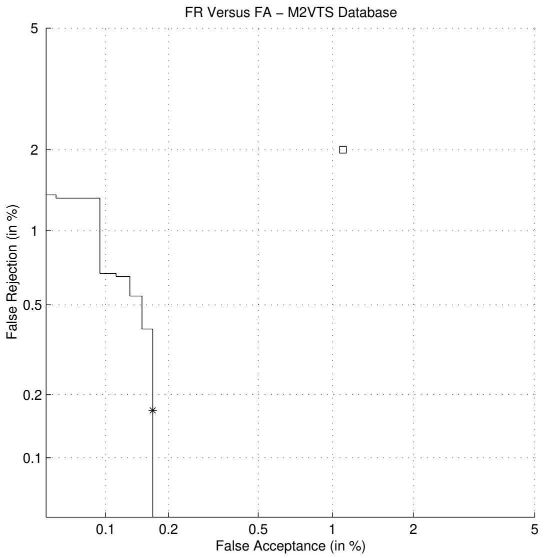


Figure 6.2: False Rejection versus False Acceptance on the M2VTS database (SVM classifiers with linear kernel). The Equal Error Rate ($\sim 0.15\%$) is indicated by a star. The square marks the fixed operation point of the system when K Nearest Neighbours experts are used (K = 1). The presence of sizeable "steps" in the curve is due to the very low number of errors it represents.

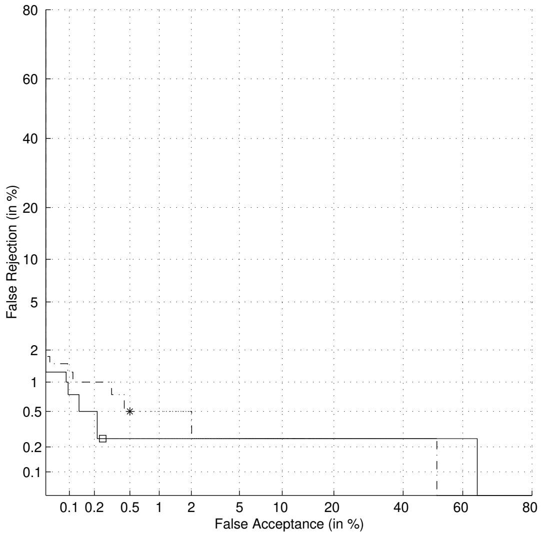


Figure 6.3: False Rejection versus False Acceptance on the XM2VTS database. The dash-dot line shows performance on Lausanne Protocol, Configuration II tests (4 training images). The EER (0.50%) is marked with a star. The solid line is obtained using 6 training images for each client. The square indicates the EER (0.25%). The FR plateau to the right of the square is due to the false rejection of a single image.

7 Dimensionality in Face Authentication

7.1 Introduction

The Face Authentication/Recognition problem we have dealt with in the preceding Chapter can be described as a classification problem in a high dimensional vector space. The dimensionality is initially defined by the size of the digitised images, that typically is of the order of ten thousand pixels. Much research effort is devoted, as in Pattern Recognition in general, to determining a subspace that would carry all the relevant information in a form that is convenient for classification [29]. The mapping between the image space and the feature space in which classification is performed can involve several steps, with linear as well as nonlinear transformations being applied at each stage [18]. The feature extraction procedure we propose in the context of the *Retinal Vision* paradigm (Chapter 2) constitutes an elaborate example of such mapping. However, the transformation involved can range all the way from simple averaging or down-sampling to fitting a complex geometrical model with hundreds of parameters.

Among these techniques, Principal Component Analysis (PCA, see Section 7.4) aims at a reduction of the dimensionality of the input space by projecting the intensity images onto a set of basis vectors (Eigenfaces) [86, 102]. Part of the literature indicates a number of Principal Components of the order of 30 to 50 to be sufficient [57, 102], while other studies such as [64] and [106], where PCA is used in conjunction with Linear Discriminant Analysis (LDA, see Section 7.4), raise this number to a few hundred components. LDA also performs a dimensionality reduction in its own right by identifying the most discriminative direction for each specific two-class problem. This is achieved by making explicit use of the labels of the training vectors in a supervised setting.

On the other hand, multi-resolution techniques such as the Gabor decomposition map (regions of) the image to a higher dimensional space which, in the continuous case, corresponds to an over-complete representation [21] of the image. Discretization of the local spectrum typically leads, as in our case, to a number of features of the order of one thousand or more (see Section 5.3.1 and [28, 55]).

We know from Chapter 4 that Support Vector Machine classifiers also implicitly involve mapping the input space onto a possibly infinite dimensional Hilbert space. A regularization technique is employed in such a Hilbert space to determine the one-dimensional subspace that is optimal for classification, in the sense specified by the Structural Risk Minimisation principle (Section 4.3). As we have seen in Chapter 6, the joint use of a sparse subset of Gabor features corresponding to the region around the eyes and the mouth and linear Support Vector Machine classifiers yields very low error rates in Face Authentication tests [87, 90].

As shown by the Face Authentication experiments presented in Sections 6.4 and 6.5, the sets of Gabor responses corresponding to the eye and mouth regions of images from different subjects turn out to be linearly separable, so that the Optimal Separating Hyperplanes of linear SVM classifiers are sufficient to discriminate between clients and impostors. Each such hyperplane determines, through its normal vector, a direction in the Gabor decomposition space. Consequently client specific authentication requires no more than one discriminant vector per client.

We show that the discriminant directions identified by SVM classifiers on a set of training subjects can effectively be used to reduce the dimensionality of the feature space used for authenticating a completely different set of subjects, thus achieving a *data-independent compression* of the biometric features. This SVM-based compression technique, which we refer to as Support Vector Features (SVF), yields a performance comparable to that obtainable by performing PCA in the Gabor decomposition space. The regularization properties of Support Vector Machines appear to play a role in this result, since an attempt to extract similar directions without resorting to Structural Risk Minimisation yields a significantly poorer performance. A higher error rate also occurs if the directions determined by performing LDA on a set of training subjects are used to compress biometric data from a disjoint set of subjects. If such a low discrimination performance on a different dataset can be expected from LDA because of its supervised nature, the generalisation ability of SVF, which is likewise a supervised technique, is all the more noteworthy.

Remarkably SVF, PCA and LDA select essentially different subspaces in the Gabor decomposition space, with SVF appearing to rely also on directions that carry a small contribution to the variance of the training set and therefore form severe angles with the most significant Principal Components. The Linear Discriminants identify, in their turn, an entirely different set of directions.

Concerning the relation between the dimensionality of the feature space, the type of classifier employed and the resulting performance, we find the following results. The error rate of simple threshold-based classifiers applied after either PCA or SVF saturates to its minimum value already in a 30 dimensional subspace. These classifiers are similar to those that have been used in earlier reports [102]. In contrast, more complex and reliable classifiers such as Support Vector Machines appear to profit from a higher dimensionality. This is confirmed by the fact that a set of 200 linear SVM classifiers trained to authenticate 200 clients in the uncompressed Gabor decomposition space appears to make full use of a 200-dimensional subspace. Again, many of the directions involved only account for a small fraction of the variance of the data, a fact that carries some analogy to Linear Discriminant Analysis. Our results therefore indicate that the classical criterion of discarding the Principal Components associated with the smallest Eigenvalues must be used with utmost care in conjunction with complex classifiers, such as the SVM classifier. The material presented in this Chapter has been published in [88, 89].

7.2 Organisation of the Chapter

The Face Authentication algorithm described in Section 6.2 is based on three machine experts that authenticate the clients in separate copies of a Gabor feature space $\mathcal{G}$. The set of Gabor filters employed consists of 30 filters arranged in 5 resolution (frequency) and 6 orientation channels. Given that the retinotopic grid consists of 50 points, the dimensionality of the feature space $\mathcal{G}$ turns out to be $30 \times 50 = 1500$.

In this Chapter we employ the Support Vector Features technique introduced in Section 7.3 and the Principal Component and Linear Discriminant analyses described in Section 7.4 to restrict the feature space to subspaces of $\mathcal{G}$. The descriptiveness of such subspaces is compared in Section 7.5. Different implementations of the machine experts and the requirements they set on the dimensionality of the feature space are discussed in Section 7.6.

7.3 Support Vector Features

The decision function of a linear Support Vector Machine classifier over the feature space $\mathcal{G}$ is a hyperplane in $\mathcal{G}$. For $\vec{v} \in \mathcal{G}$, the decision criterion is

$$\sum_j \alpha_j y_j \vec{s}_j \cdot \vec{v} + b = \vec{w} \cdot \vec{v} + b \geq 0. \tag{7.1}$$

The normal vector to the hyperplane is $\vec{w} = \sum_j \alpha_j y_j \vec{s}_j$ (see Section 4.4 and Figure 7.1 a).

Given a database of training images for an ordered set $\Gamma = \{1, 2, \dots, \#\Gamma\}$ of subjects and a corresponding set of Support Vector Machines trained to distinguish between the feature vectors belonging to person γ (a client) and those belonging to the $\#\Gamma - 1$ identities in $\Gamma \setminus \{\gamma\}$ (the training impostors), we get a normal vector $\vec{w}_\gamma$ for each client $\gamma \in \Gamma$. We consider the set $\{\vec{w}_\gamma\}_{\gamma \in \Gamma}$ of all the normal vectors. If $\vec{v}$ is an arbitrary vector in $\mathcal{G}$, we define the first n Support Vector Features $\vec{v}|_\Gamma^n$ associated to Γ and $\vec{v}$ as the ordered list $\{(\vec{w}_\gamma \cdot \vec{v}) | 1 \leq \gamma \leq n\}$ of scalar products:

$$\vec{v}|_\Gamma^n = (\vec{w}_\gamma \cdot \vec{v})_{1 \leq \gamma \leq n}. \tag{7.2}$$

The $\vec{v}|_\Gamma^n$ can be seen as n-dimensional vectors (SVF vectors) in a compressed feature space $\mathcal{G}|_\Gamma^n$.

Note that, even though several Support Vectors will result from training a SVM to authenticate a client, there is always a single normal vector which defines the hyperplane that separates this particular client from the others. Since classification can be performed directly in the compressed space $\mathcal{G}|_\Gamma^n$ (see Section 7.5), no nonlinear transformation is involved. Support Vector Features therefore compare directly with linear algorithms such as PCA and Linear Discriminant Analysis rather than with their kernel-based nonlinear extensions, kernel PCA [83] and kernel LDA [63]. This simplification is another reason for using the Gabor Space in the first place.

Clearly, if our problem is authenticating a subject in Γ there will be a Support Vector Feature that is particularly discriminating, namely the one obtained from the SVM that discriminates that subject from all others. However, we are here rather interested in determining whether the vectors $\vec{w}_\gamma$ obtained from a training set Γ consisting of a (small) number of subjects carry some general information about the structure of the Face Authentication problem itself. In other words, we would like to know if the directions identified by the $\vec{w}_\gamma$ span a subspace of $\mathcal{G}$ that accounts for the variation between any two different faces regardless of their identities. This is the question we will address in Sections 7.5 to 7.7.

7.4 Principal Components and Linear Discriminants

Principal Components are defined as the Eigenvectors of the covariance matrix $\mathsf{C} = \mathsf{X}\mathsf{X}^{\mathsf{t}}$, where X is a matrix defined by

$$\mathsf{X} = \frac{1}{\sqrt{N-1}} \left(\vec{x}_1 - \langle \vec{x} \rangle, \vec{x}_2 - \langle \vec{x} \rangle, \cdots, \vec{x}_N - \langle \vec{x} \rangle \right). \tag{7.3}$$

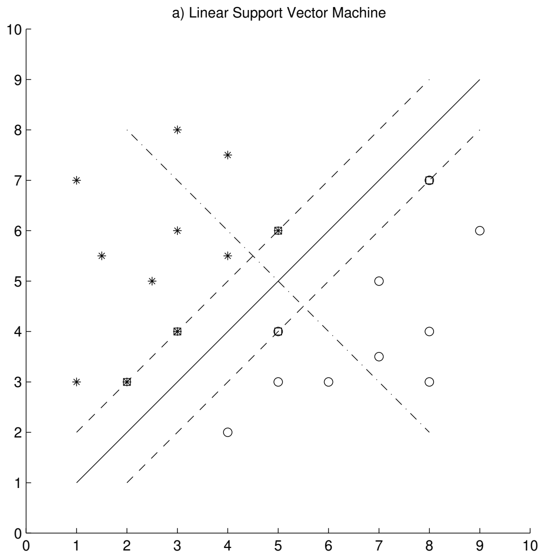
In other words, the columns of X are the training examples (feature vectors) for all classes (in our case, for all subjects in Γ) centred around their average $\langle \vec{x} \rangle$ (Figure 7.1 b). In practice, C is a large matrix (1500×1500 dimensional in our case). As long as the number N of training examples is lower than the dimensionality of the feature space, one would rather solve the simpler N-dimensional problem

$$\mathsf{X}^{\mathsf{t}}\mathsf{X}\vec{u}'_i = \lambda_i \vec{u}'_i. \tag{7.4}$$

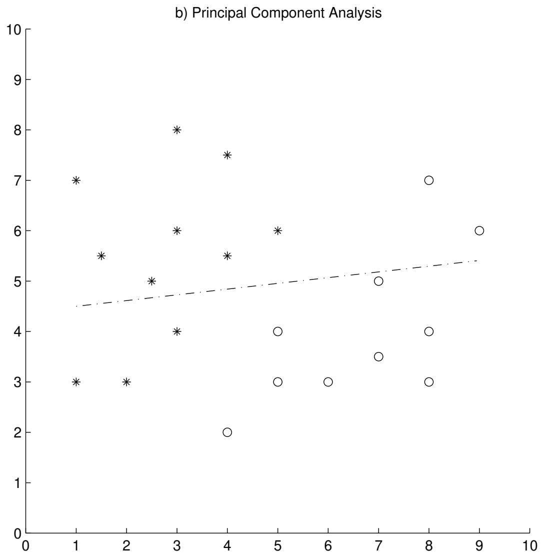
By left multiplying (7.4) by X, it is easily seen that $\vec{u}_i = (1/\sqrt{\lambda_i})\mathsf{X}\vec{u}'_i$ is a (normalised) Eigenvector of C associated with the same Eigenvalue λ_i. All of the other Eigenvalues of the covariance matrix C are zero. Since C is symmetric, the $\vec{u}_i$ are orthogonal. If we assume the λ_i to be sorted in descending order, the first few Eigenvectors account for most of the variance of the training set and can therefore be used to compress the feature vectors. The projection of a data vector on the set of Principal Components is also called Karhunen–Loève transform. In our experiments we found it useful to perform "sphering" of the data by further dividing each Eigenvector $\vec{u}_i$ by the square root of the associated Eigenvalue, which has the effect of normalising the variance of the data to 1 along all the Principal Components.

Contrary to PCA, which makes no use of class labels, and like SVF, Linear Discriminant Analysis is a supervised technique. Given a set of training vectors for the subjects in Γ, LDA consists in determining, for each two-class problem γ vs $\Gamma \setminus \{\gamma\}$, the direction $\vec{\delta}_\gamma$ that maximises the ratio

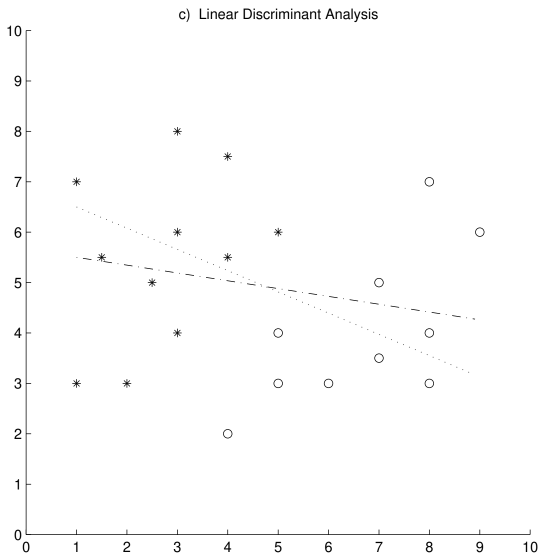
$$D_\gamma = \frac{\vec{v}^t \mathsf{B}_\gamma \vec{v}}{\vec{v}^t \mathsf{W}_\gamma \vec{v}}, \tag{7.5}$$



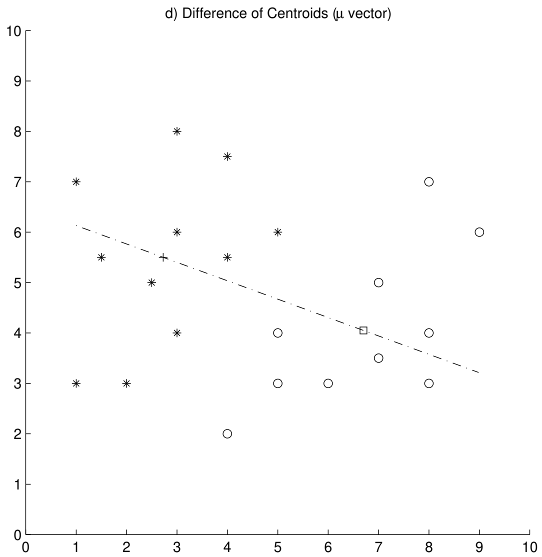
(a) The dash-dot line represents the direction of the normal vector $\vec{w}$. Support vectors are boxed.



(b) The dash-dot line represents the direction of the first Principal Component.



(c) Linear Discriminant for the "crosses" class. The dotted line represents the exact solution (Eq. 7.8); the dash-dot line is the approximate solution (Eq. 7.9).



(d) The dash-dot line represents the direction of vector $\vec{\mu}$. The centroids of the two classes are marked by a cross and a square.

Figure 7.1: Two dimensional "toy" examples of (a) Linear SVM, (b) Principal Component Analysis, (c) Linear Discriminant Analysis, (d) Difference of Centroids ($\vec{\mu}$ directions).

with respect to $\vec{v}$ (Figure 7.1 c). The between-class covariance matrix B_γ and the within-class covariance matrix W_γ can be expressed as $\mathsf{B}_\gamma = \mathsf{YY}^\mathsf{t}$ and $\mathsf{W}_\gamma = \mathsf{ZZ}^\mathsf{t}$. The definition of Y and Z is similar to the definition (7.3) of matrix X:

$$\mathsf{Y} = \frac{1}{\sqrt{P-1}} \left(\vec{x}_{\neg\gamma_1} - \langle\vec{x}_\gamma\rangle, \vec{x}_{\neg\gamma_2} - \langle\vec{x}_\gamma\rangle, \cdots, \vec{x}_{\neg\gamma_P} - \langle\vec{x}_\gamma\rangle\right) \tag{7.6}$$

and

$$\mathsf{Z} = \frac{1}{\sqrt{Q-1}} \left(\vec{x}_{\gamma_1} - \langle\vec{x}_\gamma\rangle, \vec{x}_{\gamma_2} - \langle\vec{x}_\gamma\rangle, \cdots, \vec{x}_{\gamma_Q} - \langle\vec{x}_\gamma\rangle\right). \tag{7.7}$$

with Y only including the P example vectors $\vec{x}_{\neg\gamma_i}$ for the training impostors in $\Gamma \setminus \{\gamma\}$ and Z only including the Q training vectors $\vec{x}_{\gamma_i}$ for client γ. Vector $\langle\vec{x}_\gamma\rangle$ is the average of the training examples for client γ.

By equating the ratio in (7.5) to a constant λ and performing partial derivation w.r.t. $\vec{v}$ it can be shown that, if W is positive definite, the ratio D_γ is maximised by the Eigenvector corresponding to the largest Eigenvalue in the Generalised Eigenvalue problem

$$\mathsf{B}\vec{u}_i - \lambda\mathsf{W}\vec{u}_i = 0. \tag{7.8}$$

Unfortunately, matrix W cannot be positive definite unless the number of training examples for client γ exceeds the dimensionality of the feature space, which is not the case in our present problem (minimising $1/D_\gamma$ would not help, since the number of examples for the training impostors in $\Gamma \setminus \{\gamma\}$ is also lower than the dimensionality of the feature space). One way to tackle this situation consists in splitting the feature space into orthogonal subspaces and applying LDA independently in each subspace [28]. We chose the different option of approximating the within-class covariance matrix W with a diagonal matrix Σ such that $\Sigma_{ii} = \mathsf{W}_{ii}$ and $\Sigma_{ij} = 0$ when $i \neq j$ [57]. In this way, the Generalised Eigenvalue problem (7.8) is approximated with the Eigenvalue problem

$$\Sigma^{-1}\mathsf{B}\vec{u}_i - \lambda\vec{u}_i = 0. \tag{7.9}$$

The standard procedure used to reduce the size of the matrix associated with the PCA Eigenvalue problem (Equation 7.4) is not directly applicable in this case, since $\Sigma^{-1}\mathsf{B} = \Sigma^{-1}\mathsf{YY}^\mathsf{t}$ is no longer symmetric. However, let $(\vec{u}'_i, \lambda_i)$ be the solutions of the Eigenvalue problem

$$\mathsf{Y}^\mathsf{t}\Sigma^{-1}\mathsf{Y}\vec{u}'_i = \lambda_i\vec{u}'_i. \tag{7.10}$$

It can be shown by direct substitution that the vectors $\vec{u}_i = (1/A_i)\Sigma^{-1}\mathsf{Y}\vec{u}'_i$, where A_i is a normalisation factor, along with the Eigenvalues λ_i are the solutions of (7.9). Therefore, the approximate problem (7.9) can be solved efficiently; the Eigenvector $\vec{u}_i$ associated with the largest Eigenvalue constitutes an estimate of the direction $\vec{\delta}_\gamma$ that maximises D_γ (Equation 7.5, Figure 7.1 c).

This process is repeated for all the possible choices of a client γ in Γ; for each client, a direction $\vec{\delta}_\lambda$ is obtained that discriminates him from the training impostors. Like the normal vectors $\vec{w}_\gamma$ of Section 7.3 the directions $\vec{\delta}_\lambda$, known as Linear Discriminants, are not necessarily orthogonal.

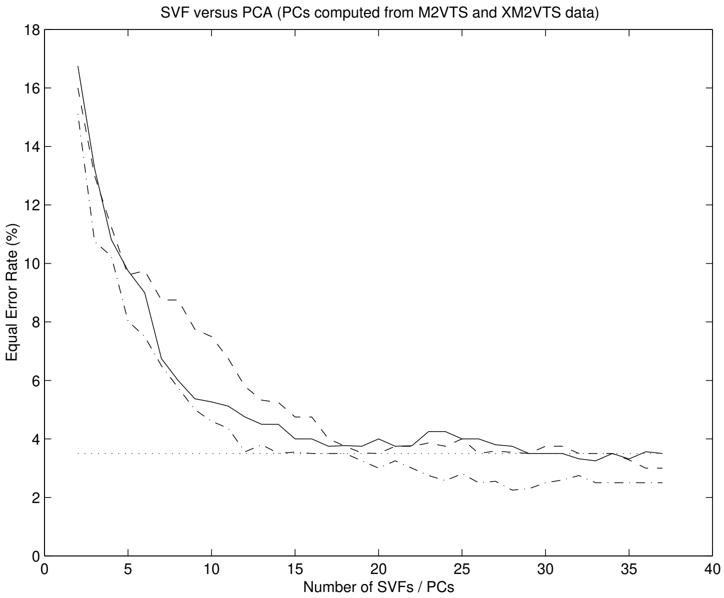


Figure 7.2: Equal Error Rate of XM2VTS Authentication tests as a function of the number of Support Vector Features (solid line), Principal Components obtained from the M2VTS database (dashed line) and PCs obtained from the XM2VTS database (dash-dot line). The dotted line indicates the performance of the simple threshold-based classifiers in the uncompressed 1500-dimensional Gabor decomposition space $\mathcal{G}$.

7.5 Support Vector Features versus PCA

The descriptiveness of Support Vector Features for Face Authentication has been evaluated with experiments involving both the M2VTS database (37 subjects) and the XM2VTS database (295 subjects; details on the databases can be found in Appendix B).

A set of 37 normal vectors $\vec{w}_\gamma$ has been computed for left-eye based, right-eye based and mouth-based authentication separately starting from the set Γ of 37 identities from the M2VTS database.

We have then tested the Face Authentication algorithm described in Chapter 6 on the entirely different set of subjects from the XM2VTS database[1] after replacing the uncompressed Gabor feature space $\mathcal{G}$ with the space $\mathcal{G}|_\Gamma^n$ of the first $2 \leq n \leq 37$ Support Vector Features associated to Γ (see Equation 7.2). More precisely, the feature space $\mathcal{G}$

[1] All the authentication results presented in this Chapter refer to the XM2VTS database and are obtained according to the Lausanne Protocol, Configuration II (see Section 6.5 and [62]), but for the fact that the images from the third session have been included into the training set. The protocol allows generating a total of 400 client tests from 200 clients, while $112'000$ impostor tests are obtained from 60 impostors. This allows estimating the Equal Error Rate (EER) with a resolution of 0.25%, corresponding to the false rejection of a single image of a client. The position of the eyes and the mouth has been manually detected on the images of the M2VTS database. Automatic detection has been used on the XM2VTS database (see Chapter 5); erroneous detections have been corrected manually.

has been compressed by computing the dot products between the Gabor feature vectors $\vec{v} \in \mathcal{G}$ obtained from *all* the images in the XM2VTS database and the first n vectors $\vec{w}_\gamma$ to give the first $2 \leq n \leq 37$ Support Vector Features. The machine experts that authenticate the clients have then been restricted to work on the resulting n-dimensional SVF vectors $\vec{v}|_\Gamma^n$.

For these experiments, the machine experts have been implemented as simple threshold based classifiers of the type commonly used in conjunction with PCA. An identity claim is accepted if the distance between the corresponding SVF vector $\vec{v}|_\Gamma^n$ and the closest training example (SVF vector) $\vec{x}_{\gamma_i}|_\Gamma^n$ for the claimed identity γ is lower than a fixed threshold τ, otherwise it is rejected. The resulting decision function is

$$\min_i \|\vec{x}_{\gamma_i}|_\Gamma^n - \vec{v}|_\Gamma^n\| \leq \tau. \tag{7.11}$$

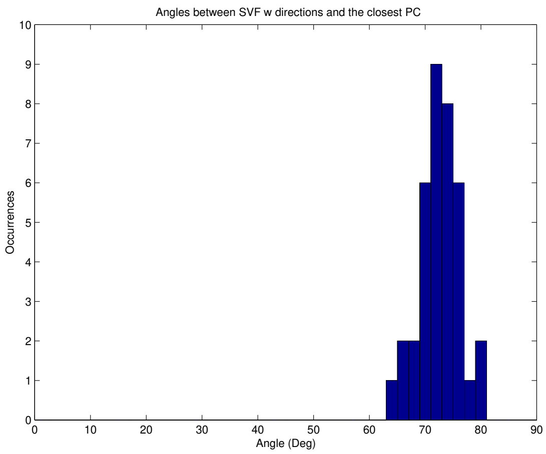
Figure 7.2 (solid line) shows the Equal Error Rate obtained in XM2VTS authentication tests as a function of the number n of Support Vector Features employed for classification. The Equal Error Rate obtained using the full-length, 1500-dimensional Gabor feature vectors is 3.5%. As can be seen, the Support Vector Features technique allows achieving the same result already in a 29-dimensional subspace, with a compression ratio in excess of 50. Even in a 15-dimensional subspace the Equal Error Rate is as low as 4.0%.

Analogous results are obtained by projecting the Gabor vectors onto a variable number of Principal Components determined by performing PCA over the M2VTS database (note that no statistical knowledge about the test database XM2VTS is employed). Again, authentication performance quickly saturates to the minimal EER allowed by the classifiers (Figure 7.2, dashed line).

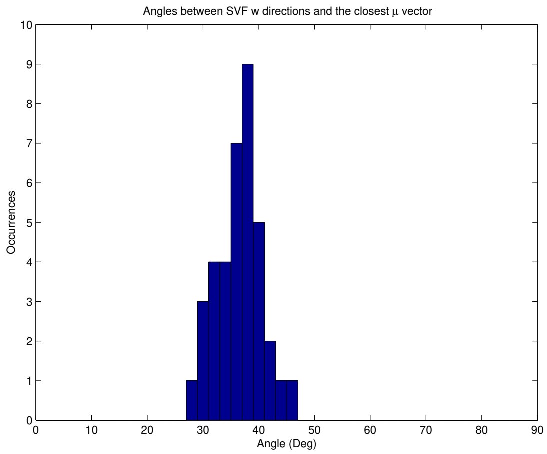
The dash-dotted line in Figure 7.2 refers to the Equal Error Rate obtained by extracting the Principal Components from the XM2VTS training data directly. As can be seen, the introduction of statistical knowledge about the very same identities the system is required to authenticate only results in a relatively small improvement in performance, at the expense of making the feature compression scheme data-dependent.

7.6 Is there an optimal subspace for Face Authentication?

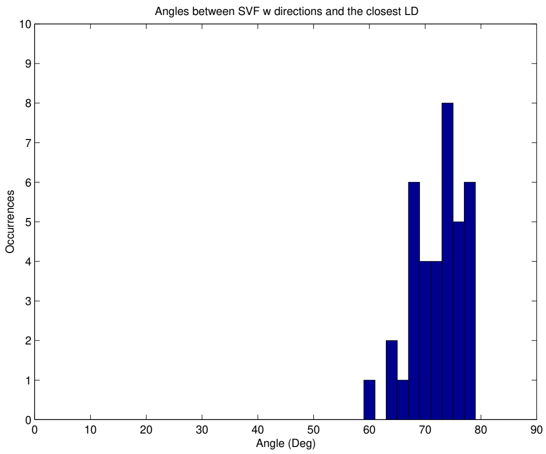
The experiments discussed in the preceding section suggest that a 30-dimensional subspace might be sufficient to perform face authentication, a result that would be in agreement with other studies on Principal Component Analysis using similar classification techniques [102]. Furthermore, such a subspace appears to be characterised by a certain stability, in that the directions $\vec{w}_\gamma$ or the Principal Components computed on the M2VTS database can be used to compress the Gabor feature vectors from the completely different set of subjects constituting the XM2VTS database with only a small performance degradation. However, the optimal subspace for authentication need not be unique. The



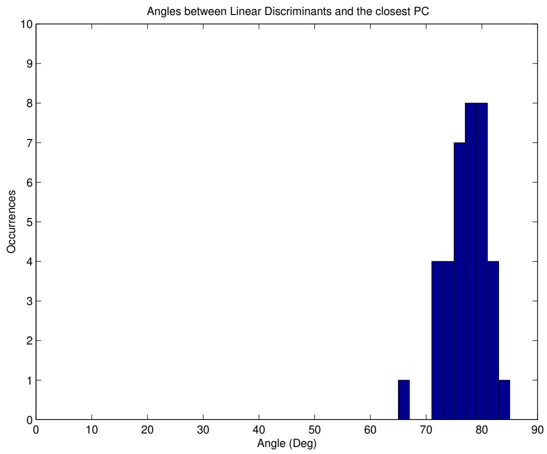
(a) Angle between each normal vector $\vec{w}_\gamma$ and the closest direction among the 37 most significant Principal Components (median: 73°).



(b) Angle between each normal vector $\vec{w}_\gamma$ and the closest direction among vectors $\vec{\mu}_\gamma$ (median: 37°).



(c) Angle between each normal vector $\vec{w}_\gamma$ and closest Linear Discriminant (median: 73°).



(d) Angle between each Linear Discriminant and the closest direction among the 37 most significant Principal Components (median: 78°).

Figure 7.3: Pairwise comparison of the preferential direction identified by (a) SVF and PCA, (b) SVF and the $\vec{\mu}$ vectors, (c) SVF and LDA, (d) LDA and PCA. All data refer to the M2VTS database, left eye expert.

subspace spanned by the Principal Components only accounts, on the average, for 36% of the norm of the $\vec{w}_\gamma$; therefore, PCs and SVF essentially map the Gabor vectors to two different subspaces. As is shown in Figure 7.3 (a), the angles formed by the direction of each $\vec{w}_\gamma$ with the most similar direction among the first 37 Principal Components from the M2VTS database are quite high (median: 73°). It follows that a large contribution to the SVF comes from directions that only carry a small fraction of the variance of M2VTS data, a result that is confirmed by the experiments reported in Section 7.7.

The comparable performance of SVF and PCA appears to be a consequence of the regularization properties of Support Vector Machines. This becomes evident if we compare the SVF dimensionality reduction technique with the results obtained by projecting the data onto the (intuitively similar) set of directions $\vec{\mu}_\gamma$ constructed by subtracting the average training vector $\langle \vec{v}_\gamma \rangle \in \mathcal{G}$ for identity γ from the average of all other identities in Γ:

$$\vec{\mu}_\gamma = \frac{\sum_{\lambda \neq \gamma} \langle \vec{v}_\lambda \rangle}{\#\Gamma - 1} - \langle \vec{v}_\gamma \rangle \tag{7.12}$$

where Γ is the set of subjects represented in the M2VTS database and $\#\Gamma = 37$ is their number (Figure 7.1 d). As shown in Figure 7.3 (b), the directions of the vectors $\vec{\mu}_\gamma$ are indeed closer to the directions of the $\vec{w}_\gamma$ than the Principal Components considered in Figure 7.3 (a) and their span accounts for an average 57% of the norm of the vectors $\vec{w}_\gamma$. However, they are not nearly as efficient as the Support Vector Features in compressing biometric data belonging to the (different) subjects from the XM2VTS database (Figure 7.4). It must be noted that since

$$\sum_{\gamma \in \Gamma} \vec{\mu}_\gamma = \sum_{\gamma \in \Gamma} \frac{\sum_{\lambda \neq \gamma} \langle \vec{v}_\lambda \rangle}{\#\Gamma - 1} - \sum_{\gamma \in \Gamma} \langle \vec{v}_\gamma \rangle = 0$$

the vectors $\vec{\mu}_\gamma$ span at most a $\#\Gamma - 1 = 36$ dimensional subspace. As Figure 7.4 shows, using the Linear Discriminants determined from the subjects of the M2VTS database to compress biometric data from the XM2VTS database also leads to poor results. This can be understood from the fact that while the Principal Components drawn from a database can be expected to be a small-sample approximation of the Principal Components for the set of "all faces", each Linear Discriminant encodes information that is specific to the particular two-class problem for which it has been derived. In this respect the good performance of Support Vector Features, that are also obtained from specific two-class problems, appears even more remarkable. We finally note that the Linear Discriminants computed from the M2VTS database span a subspace that only account for an average 9% of the norm of vectors $\vec{w}_\gamma$; their directions carry little resemblance both to the vectors $\vec{w}_\gamma$ and to the first 37 Principal Components extracted from the same database (Figure 7.3 c–d).

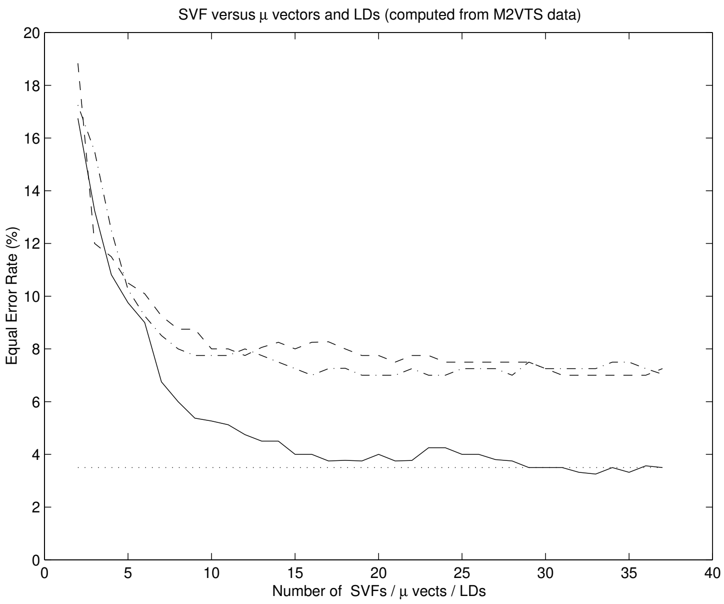


Figure 7.4: Equal Error Rate of XM2VTS Authentication tests as a function of the number of Support Vector Features (solid line, as in Figure 7.2), $\vec{\mu}_\gamma$ vectors (dash-dot line) and Linear Discriminants obtained from the M2VTS database (dashed line). The machine experts for authentication are implemented as simple threshold-based classifiers.

7.7 Low dimensionality and advanced classifiers

We might ask whether the apparent convergence of SVF and PCA based authentication to the performance achieved on the unaltered feature space $\mathcal{G}$ indicates that all the available information has been captured, or whether it is rather a consequence of the (simple) classification technique employed for authentication.

Substituting linear SVM classifiers for the simple threshold based rule on the entire Gabor decomposition space $\mathcal{G}$ causes the Equal Error Rate (EER) achieved on the XM2VTS database to drop from 3.5% to below the resolution limit of 0.25%, that corresponds to a single falsely rejected identity claim (see Section 6.5 and [87]). In Table 7.1 we show the EER obtained by using SVM experts in conjunction with the three choices of subspaces in $\mathcal{G}$ corresponding to 34 Support Vector Features, 34 Principal Components obtained from the M2VTS database and PCA over the XM2VTS database. We recall that the first two choices both lead to the minimal EER of 3.5% when the simple threshold classifiers are used (see Figure 7.2). As can be seen, authentication performance on the compressed Gabor decomposition space is considerably improved when SVMs are used for classification. In particular, the dimensionality boost provided by nonlinear kernel functions appears to somewhat compensate for the compression, yielding lower error rates. Performance does not differ significantly on the subspaces determined by SVF and PCA, with PCA on the XM2VTS database being (as expected) slightly better. However, the best results still fall

EER	34 SVF	34 PCs (M2VTS)	34 PCs (XM2VTS)	SVM on $\mathcal{G}$ (1500-dim)
Linear	2.3%	2.1%	1.5%	$\lesssim 0.25\%$
Best	1.6%	1.5%	1.0%	$\lesssim 0.25\%$
Degree	3	2	2	1–4

Table 7.1: Equal Error Rates of XM2VTS tests using 34 SVF, the first 34 PCs extracted from the M2VTS database and the first 34 PCs from the XM2VTS database. Performance of both linear kernels and the optimal polynomial kernel is shown. The degree of the polynomial is reported in the last line. The last column refers to SVM classification over the entire feature space $\mathcal{G}$. In this case, all the kernels tested give an EER below the resolution limit of 0.25%, corresponding to a single falsely rejected image.

short of the performance achieved using the entire Gabor space $\mathcal{G}$ in conjunction with Support Vector Machines. This suggest that the SVM experts exploit information that is lost in the compression procedure and that the dimensionality of the three subspaces actually is too low.

In order to get a feeling for what the correct dimensionality might be, we consider the set $\{\vec{w}_\xi\}$ of the normal vectors of the 200 linear SVMs trained over the set Ξ of the 200 clients from the XM2VTS database (classification takes place in the uncompressed Gabor decomposition space $\mathcal{G}$). On the average, only 28% of the norm of each $\vec{w}_{\bar{\xi}}$ lies in the subspace spanned by the remaining 199 vectors $\{\vec{w}_\xi\}_{\xi\in\Xi\setminus\{\bar{\xi}\}}$. This indicates that all of the $\vec{w}_\xi$ identify independent directions, meaning that authentication effectively involves a 200-dimensional subspace of $\mathcal{G}$. It is interesting to have an idea of what fraction of the norm of the $\vec{w}_\xi$ contributes an independent direction as the size of Ξ varies. An indicative answer can be obtained by normalising the $\vec{w}_\xi$, applying the Gram-Schmidt orthogonalisation procedure and displaying the successive residues as a function of the index ξ (Figure 7.5; the ordering of Ξ is arbitrary).

By looking at the distribution of the magnitude of the first 200 PCA Eigenvalues on the same set Ξ of clients we see that SVM classification necessarily involves many directions that carry only a small fraction of the global variance. Again, this is confirmed by the distribution of the angles between the $\vec{w}_\xi$ and the closest direction among the first 200 PCA Eigenvectors (Figure 7.6).

The capacity of SVM classifiers to extract relevant information from directions carrying little variance can possibly be traced to the "zooming" effect of the norm of the $\vec{w}_\xi$, which is set by the optimisation procedure so that the margin is normalised to 2 regardless of how close to the hyperplane the Support Vectors of the two classes are (see Subsection 4.4). An analogous effect might be present in [106], where the combination of PCA and Linear Discriminant Analysis is found to require 300 Principal Components for best performance. Since LDA determines the most discriminant directions according to the ratios of within-class and between-class variances, it does not necessarily select directions that carry a

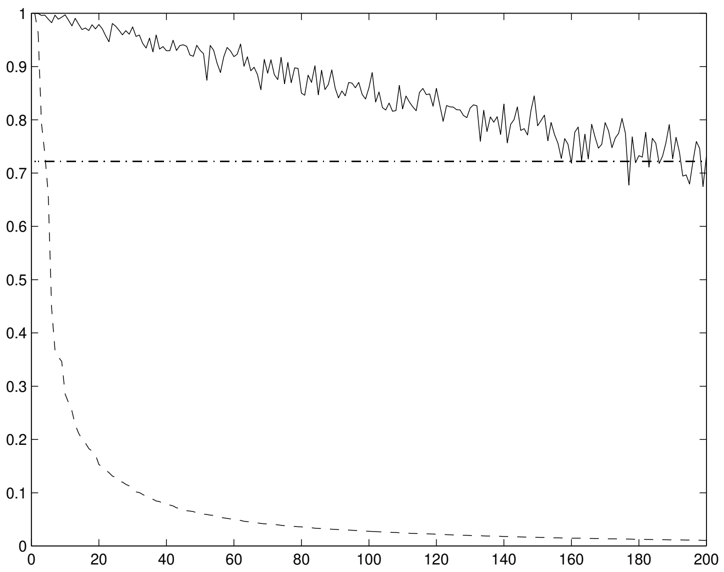


Figure 7.5: Solid line: norm of the residues obtained by applying the Gram-Schmidt orthogonalisation procedure to the list of (normalised) vectors $\vec{w}_\xi$. Dash-dot line: average percentage of the norm of each $\vec{w}_\xi$ that contributes an independent direction. Comparison with the relative magnitude of the first 200 Eigenvalues from PCA (dashed line) shows that most of these directions only carry a small fraction of the data variance. All data refer to the set of clients from the XM2VTS database, left eye expert.

large global variance. A further intuitive link between these two situations might be provided by the fact that SVM classifiers have been shown to be equivalent to LDA on the set of Support Vectors [85].

7.8 Final remarks

The Chapter presents a comparative study of the dimensionality of the Face Authentication problem. Three types of compression algorithms are employed to reduce the dimensionality of the feature space, namely PCA, Linear Discriminant Analysis and a novel Support Vector Machines based technique that we call Support Vector Features (SVF). We report experimental results using two classification techniques, thresholded distance classifiers and Support Vector Machines. PCA and SVF are shown to have a comparably good performance regardless of the classifier employed, although they identify different subspaces in the Gabor decomposition space.

The Support Vector Features technique, that relies on the vectors $\vec{w}_\gamma$ determined by the same set of two-class problems used to compute the Linear Discriminants, remarkably outperforms LDA based compression. However, a large arbitrariness is involved in the construction of such two-class problems. Instead of constructing “one-against-all” problems, one could group the M2VTS training identities according to other criteria; this

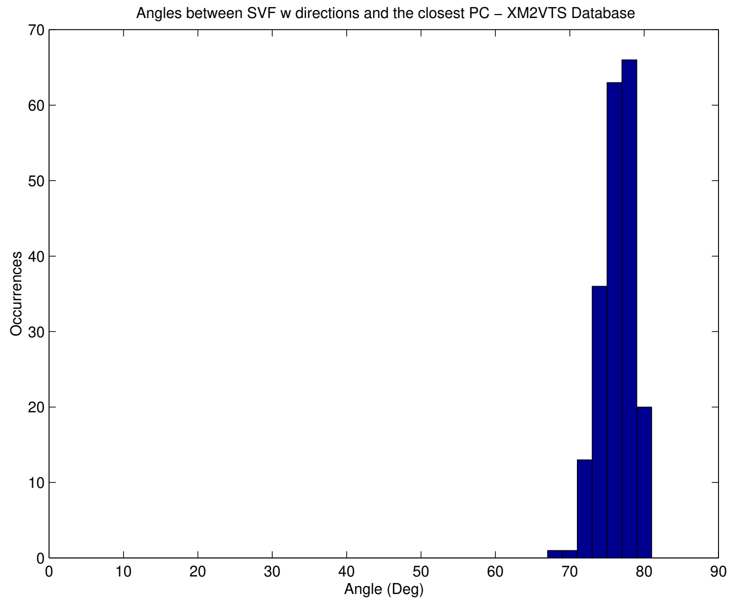


Figure 7.6: Angle between each normal vector $\vec{w}_\xi$ obtained from the set Ξ of 200 clients from the XM2VTS database and the closest direction among the 200 most significant Principal Components from the same database (left eye expert; median: 77°).

could modify the performance of both the SVF and the LDA based techniques. In particular, such groupings as women vs men, wearing glasses vs not wearing glasses, etc. would be of particular interest, since they would provide preferential directions in feature space that would likely be independent of the specific identities and could be incorporated into a decision-tree based authentication system. However, finding a large amount of such "natural" groupings for use in a feature compression scheme like those considered in this Chapter would scarcely be feasible.

Experiments with a simple threshold-based classifier and SVMs suggest that the saturation of recognition/authentication performance with a low number of Principal Components or Support Vector Features heavily depends on the choice of the classifier. In particular, authentication of the 200 clients from the XM2VTS database using linear SVM classifiers is shown to require a 200-dimensional subspace. Support Vector Machine classifiers appear to rely on directions that are not necessarily associated to a large variance of the training data. Therefore, the classical criterion of discarding Principal Components associated with small Eigenvalues must be used with care in conjunction with SVM classifiers.

The problem of the dimensionality of the feature space necessary for reliable Face Authentication/Recognition remains open. However, the results presented in this Chapter suggest that safe bets should be situated in the range above one thousand dimensions. Considering the present-day capacity of storage devices such as chip cards, such a dimensionality might constitute a good compromise between the compactness of biometric data and their accuracy.

8 Conclusions

The present Thesis introduces a new vision paradigm, namely a concept of *Retinal Vision*. It consists of an attentional mechanism built around an artificial retina driven by saccades, that is able to locate complex structures in static images as well as in video streams.

The experiments reported indicate that a good integration between cognitive tasks and low-level visual processing can be achieved if the biological analogy that underlies retinotopic sampling and the Gabor decomposition is coherently pursued at higher stages of the chain that goes from the visual input to symbolic representations.

The main idea is that of providing a simple view-based mechanism that allows building models of the objects to be located and possibly authenticated. Situated as an intermediate layer between raw data conditioning and complex cognitive processes, such a mechanism is intended to allow the latter to be more perspicuous and effective in their use of visual information. At the same time, this intermediate layer must be flexible enough not to limit the nature of the cognitive tasks, that should see their capability to access meaningful visual information augmented.

We believe that the model of *Saccadic Search* presented in this Thesis attempts to implement such a concept by providing a flexible *general purpose* attentional mechanism. This is confirmed by the experimental results obtained in Real-Time Head Detection and Tracking, Facial Features Detection and Face Authentication. The low-level visual processing tools involved, i.e. retinotopic sampling and the Gabor decomposition, have already been widely applied in Computer Vision in general as well as in the specific domain of Face Authentication, which makes the specific contribution of the more organic *Retinal Vision* approach plausible.

In the following, we briefly review the main contributions of the present work and suggest directions for further developments.

8.1 Saccadic Search

The *Saccadic Search* strategy introduced within the *Retinal Vision* paradigm represents a powerful general-purpose attentional mechanism. Although computational models of saccades relying on orientation selective multi-scale filters have already been proposed, our model exploits *a-priori* knowledge about the patterns of interest in a context of *partial information* about the image, which makes object localisation possible without the need for an exhaustive search.

Appearance-based models of the search targets are encoded by means of Support Vector Machines. The use of such classifiers in the Gabor feature space also represents an element of novelty.

Finally, the application of Log-polar sampling to the Gabor decomposition represents an original attempt at merging two established mathematical models of the early stages of biological vision, describing respectively the retino-cortical mapping and the responses of the simple cells in the primary visual cortex.

The robust performance achieved in Facial Features Detection over a total of more than 2700 images from two different Databases confirms the feasibility of our approach.

8.2 Real-time performance

The *Saccadic Search* strategy makes such a powerful Computer Vision tool as the Gabor decomposition affordable for active-vision applications. This is achieved by focusing the computational power on relevant regions of the image. The extraction of Gabor filter responses is only performed at a small fraction of the image points; image domain filtering therefore offers a straightforward and convenient solution.

The Real-time Head Detection and Tracking application presented demonstrates the suitability of the *Saccadic Search* strategy in an active-vision framework. We believe the system described to constitute the first example of the application of the Gabor decomposition to this type of tasks.

8.3 Face Authentication

Very low Equal Error Rates have been achieved in Face Authentication experiments over the two largest verification oriented databases available.

Besides constituting a considerable advance in the reliability of Face Authentication algorithms, these results confirm the descriptiveness and discriminative power of the appearance based features employed in the *Retinal Vision* paradigm. It is remarkable, in this sense, that optimal authentication performance can be achieved using linear classifiers.

A detailed study of the dimensionality of the Face Authentication problem has been presented. Several classification techniques and feature compression schemes have been considered, ranging from widely used statistical tools such as PCA and Linear Discriminant Analysis to a novel compression algorithm that we named Support Vector Features. Such a study contributes to filling a gap in the existing literature on Face Authentication, which presents a variety of widely disagreeing estimates concerning the number of features required for reliable authentication.

8.4 Possible extensions

Much of the research effort for the Saccadic Search strategy was spent in the development of robust features and flexible models of the search targets that afford reliable pattern recognition in a context of partial information, with a dramatic reduction of the computational load. While this objective has been achieved, the computational efficiency could be further increased by using more sophisticated optimisation algorithms. In the second stage of Facial Feature Detection, after the saccadic pattern has converged to a candidate feature, the retinotopic grid is displaced in a pixel-by-pixel fashion in order to maximise the output of the extended models. Although it only involves a small image region, this procedure is computationally demanding and could be improved.

The use of more advanced optimisation techniques in the planning of saccades would only lead to a marginal efficiency gain in the case of static images; however, time integration would prove useful in enhancing real-time performance. A detailed comparison of the saccadic patterns generated by the *Saccadic Search* strategy with the actual eye movements of human subjects performing the same tasks over identical images or video sequences might provide a deeper insight into the role of motion extrapolation and of target representation in the planning of saccades.

Although the databases used for Facial Features Detection and Face Authentication included scale changes up to 10% and severe pose disturbances in the supposedly frontal images and the results were good, extended scale and pose invariance should yet be included into the Retinal Vision approach. Ideally, those issues should be addressed by introducing generalisation ability with respect to scale and pose at the level of the saccadic planning, thus providing the cognitive level with a more powerful interface to the world. The Log-polar transformation naturally provides a correspondence between scaling and rotations of the image and translations in the transformed space. This fact could be used to compute scale and rotation invariant features starting from the Gabor responses on the points of the retinotopic sampling grid. This might eventually close the loop on neurological studies by providing stronger computational evidence for the possible role played by the retino-cortical mapping in achieving scale invariance in biological vision.

Acknowledgements

I wish to thank Prof. Josef Bigün, who first employed me at EPFL, for investing two years of his time in introducing me to Computer Vision research and techniques. Without his continued supervision and advice the present work would not have been possible.

Prof. Wulfram Gerstner took over the supervision of my Thesis and allowed me to continue this research as a member of his group. I am indebted to him for his great flexibility and availability and for providing me with constant, broad-angle stimulation.

Special thanks to Nathalie Capdevielle, who was my co-worker during one year on the VIRSBS Project and helped me out of many a local minimum. Yousri Abdeljaoued, Benoît Duc and Stefan Fischer all contributed to the present Thesis with their valuable suggestions and their friendship.

Olivier Carmona and Sébastien Menot, the developers of the Eyeball-mounted camera (EPFL Vision Sphere), made available both their help and their hardware for the Head Detection and Tracking system. Marcel Bachmann implemented part of the software of the Head Detection and Tracking system while he was a student at EPFL.

This work would have been computationally unfeasible without the help of Yuri López de Meneses Novosilzov and Michel Pahud, who together with Olivier Carmona kept the computers up and running (almost) to the end.

Monique Dubois and Marie-Jo Pellaud saw me through all the bureaucracy and never failed to make sure that I was feeling at ease and enjoying my work.

This research was partly funded by the European Community IT–LTR Project VIRSBS. Finally, I would like to thank the University of Halmstad for providing me with financial and logistical support for my numerous visits.

Appendices

A Gabor Filters

A.1 Introduction

Gabor filters [23, 37, 59, 65, 74] represent a classical tool in Computer Vision. However, due to the limited space to which scientific publications are often confined, little documentation exists on the technical aspects of their implementation. In particular, no guidelines are readily available for such practical problems as the choice of a range of suitable frequencies and of the other parameters that specify a family of Gabor filters.

The present Appendix has the purpose of providing hands-on information on filter design and implementation. No claim is made at completeness or at a rigorous formal treatment of the subject. This material is provided with the hope to convey the practical knowledge and expertise acquired by the author while experimenting with the Gabor decomposition.

The definition of one-dimensional and two-dimensional filters is given in Sections A.2 and A.4 respectively. Section A.5 reports guidelines for the choice of the filter parameters. The different ways of computing the filters are briefly discussed in Section A.6.

The rest of the Appendix focuses on the set of modified Gabor filters used in the present Thesis as well as in [8, 9, 51]. Their definition is recalled in Section A.7. These filters are characterised by a mathematically elegant and practically convenient design, the details of which are treated in Section A.7.1.

A.2 One-dimensional Gabor filters

A family of one-dimensional Gabor filters (or more precisely Gabor wavelets, see Section 2.3) can be obtained starting from the 1D Gabor function

$$G(x|\sigma, \omega_0) = e^{-\frac{x^2}{2\sigma^2}} e^{i\omega_0 x} \tag{A.1}$$

This represents a complex sinusoidal wave windowed by a Gaussian. The standard deviation of the window is σ, while ω_0 is the angular frequency of the wave. This function is then linearly scaled by the parameter s, giving the family of Gabor wavelets $G_s(x|\sigma, \omega_0) = G(x|\sigma s, \omega_0/s)$. In practice only a discrete set of values is used for s. These values are most often chosen according to a geometric progression.

A.3 Frequency selectivity

We note that the width of the Gaussian envelope σ and the wavelength of the sinusoidal wave $\lambda_0 = 2\pi/\omega_0$ are scaled by the same factor, so that the number of cycles of the sinusoidal wave under that part of the envelope which is significantly different from zero is always the same. For this reason J. Morlet, who first applied these filters to Signal Analysis, described them as "wavelets of constant shape" [65]. Therefore, what essentially matters in defining the shape of the filters is the ratio $\lambda_0/\sigma = 2\pi/(\sigma\omega_0)$ rather than the individual values of these parameters. The meaning of this quantity is elucidated by the Fourier domain representation of Equation (A.1):

$$\hat{G}(\omega|\sigma,\omega_0) = Ne^{-\frac{(\omega-\omega_0)^2}{2(1/\sigma)^2}}, \tag{A.2}$$

where N is a normalisation factor. As can be seen, the ratio between the width $(1/\sigma)$ of the filters in the Fourier domain and the frequency ω_0 at which they are tuned is $1/(\sigma\omega_0)$. The product $\sigma\omega_0$ therefore measures the (relative) frequency selectivity of the filters. It follows that filter sets having a low λ_0/σ ratio, i.e. those showing a large number of cycles of the sinusoidal wave under the Gaussian window, have highly selective frequency responses, and vice-versa.

A.4 The two-dimensional case

The two dimensional Gabor function can be defined as in Equation (2.6):

$$G(x, y|\sigma_x, \sigma_y, \omega_0) = e^{-\frac{x^2}{2\sigma_x^2}} e^{-\frac{y^2}{2\sigma_y^2}} e^{i\omega_0 x}. \tag{A.3}$$

This filter is a Gaussian blob centred at the origin, modulated by a plane sinusoidal wave with wave vector parallel to the x axis. It will therefore be sensitive to vertical structures in the image, such as vertical edges.

The Fourier domain representation of G is a Gaussian with horizontal frequency width $\sigma_{\omega_x} = 1/\sigma_x$ and vertical frequency width $\sigma_{\omega_y} = 1/\sigma_y$. The centre of the Gaussian is translated along the horizontal frequency axis ω_x by the modulation frequency ω_0 (see Equation 2.7).

Since vertical structures in the image (those that are described by an intensity pattern of the form $I(x, y) = f(x)$) map to the ω_x axis in the Fourier domain, we can see that:

- as said above, this filter will mostly respond to vertical structures;
- the parameter σ_y controls the orientation selectivity of the filter. As we make σ_y *larger*, the filter will shrink *closer* to the ω_x axis in the Fourier domain and will therefore become *more* orientation selective.

In order to generate a family $G_{s\phi}$ of two-dimensional Gabor filters, one operates two kinds of transformations on function G: scaling by parameter s, just as in the one-dimensional case, and rotation by an angle ϕ. Of course, the rotation parameter ϕ must also be discretized — if there are no specific orientations that are relevant to the image, fewer than ten uniformly spaced values between 0 and π are normally sufficient. The following choices have been made in this Thesis:

- Real-time Head Detection and Tracking: 3 frequency channels and 8 uniformly distributed orientation channels;
- Facial Feature Detection and Face Authentication: 5 frequency channels and 6 uniformly distributed orientation channels.

A.5 Guidelines for the choice of the parameters

No strict rules exist that define the parameters of the optimal set of Gabor filters for a given application. However, some general guidelines can be obtained by considering the filtering process both in the image domain and in the frequency domain. This analysis somehow boils down to a number of "fuzzy constraints" on the parameters, but the answer still largely depends on the problem at hand and on personal choice. This of course applies only to Pattern Recognition applications; coding and reconstruction impose tighter constraints on the sampling of the local frequency spectrum.

The tentative "check list" presented below is meant to help determining most of the filter parameters based on heuristic reasoning in the image domain. Considerations involving the Fourier domain representation are presented in a separate paragraph.

A.5.1 Filter design check list

- Choose the number of orientation and frequency channels. If there are preferential orientations in the images, these should be represented in the filter set; if not, choose equally spaced orientation channels, say six. The total number of channels will often be limited by the time that is affordable for feature extraction in the specific application.
- Fix the frequency ω_h of the highest frequency channel. This is best done by looking at the associated wavelength λ_h in pixels, since $\lambda_h = 2\pi/\omega_h$. In a way, λ_h limits the positional accuracy that can be expected from a pattern recognition algorithm based on the filter set. This is the case especially for systems that only use the magnitude of the response, since this quantity will practically not vary on a scale that is small compared to λ_h. The value of λ_h can reasonably be matched to the linear dimensions of the smallest meaningful structure one is interested to detect in the images. A reasonable choice for an eye detection application, for instance, would be to set λ_h equal to the diameter of the pupil.

- Fix the frequency selectivity by choosing a value of σ_x for the highest frequency filter; the values of σ_x and ω_0 for all the other filters are then determined by the scaling and rotation transformations. Setting $\sigma_x = 3\lambda_h$ probably is a good guess; however, the final choice represents a trade-off between the locality of the filters and their frequency selectivity.

- Fix the frequency of lowest frequency channel. One way to do this is to match its wavelength with the dimensions of the largest structure of interest in the images. For an eye detection application, for instance, λ_l would be of the order of the width of the eye. The discretization of the scale parameter s only depends, at this point, on the desired number of frequency channels.

- Consider the actual feature extraction procedure. If you are computing the filters on the points of a square grid, you may not want the widest (low frequency) filters to overlap too much, or their responses will no longer be independent. The overlap can be tuned acting on two parameters: the wavelength λ_l of the low frequency filters and the x axis width of the Gaussian envelope σ_x.

A.5.2 Fourier domain considerations

The above procedure allows determining all of the filter parameters but for σ_y. This last quantity can be fixed by considering the Fourier domain representation of the filter bank, which also gives a deeper understanding of the role of the other parameters.

As a general rule in feature extraction for Pattern Recognition, it is important that the Fourier domain be sampled in a rather uniform way. Once a frequency range has been fixed by choosing λ_h and λ_l, a uniform coverage can be obtained by tuning the width of the filters, i.e. σ_x and σ_y, according to the number of frequency and orientation channels. In particular, σ_y should be chosen so as to obtain adjacent orientation channels that do not overlap too much.

An excessive overlap of the filters in the frequency or image domain will result in the features becoming correlated due respectively to a lack of frequency/orientation selectivity or spatial locality. This effect might be undesirable, depending on the application, and one should generally be aware of it. It is also good to remember that the shape of the filters in the frequency and in the image domain is, in a way, inversely correlated, meaning that filters with a narrower pass-band will be larger in the image domain and vice-versa.

A.5.3 Summing up

There is no straightforward way to choose the parameters of a filter set. A good piece of advice probably is: design a reasonable filter set, display it and iterate until all the above requirements are more or less satisfied. Again, it does help to actually look at the filters both in the frequency and in the image domain.

A.6 Computing the filters

Gabor filters can be computed in several ways. Which way to go depends on personal preference and on whether one wants to implement the filtering in the image domain or in the Fourier domain. Apart from the horizontal and vertical orientations, Gabor filters are non-separable. Fourier domain filtering is therefore convenient if filter responses along generic orientations must be computed on large portions of the image.

In this section, we present a quick overview of the different possibilities. Details on the construction of the modified filters employed in this Thesis are given in Sections A.7 and A.7.1.

In a straightforward approach, a scaled version $G(x, y|s\sigma_x, s\sigma_y, \omega_0/s)$ of the Gabor function defined in Equation (A.3) can be computed directly for suitable values of s to give the filters tuned to the horizontal direction. The resulting bitmaps can then be rotated to obtain the other orientation channels. Alternatively, one could rewrite Equation (A.3) in a rotated and scaled coordinate frame, which would probably lead to reduced numerical errors. In this case, of course, σ_x and σ_y no longer refer to directions parallel to the image borders.

Another option consists in computing the filters in the Fourier domain. This is quite practical since Gabor filters are real-valued Gaussians in the frequency plane. The case with $\sigma_x = \sigma_y = \sigma$ is particularly simple. Different orientation channels for the same frequency correspond to identical Gaussians centred on equidistant points on a circle around the origin. The radius of the circle determines the frequency of the filters. The other frequency channels are obtained by scaling. As an example, let us consider a set of filters comprising N_f logarithmically separated frequency channels and N_o orientation channels. Let ω_l (ω_h) be the angular frequency of the lowest (highest) frequency channel. The ratio δ_ω of the progression of the frequencies is computed as follows:

$$\delta_\omega = \left(\frac{\omega_h}{\omega_l}\right)^{\frac{1}{N_f - 1}}. \tag{A.4}$$

Octave-separated filters, for instance, correspond to $\delta_\omega = 2$. The Fourier domain coordinates of the centres of the filters are then given by

$$\Omega_{fo,x} = \omega_l (\delta_\omega)^f \cos\left(\frac{2\pi}{N_o} o\right) \tag{A.5}$$

and

$$\Omega_{fo,y} = \omega_l (\delta_\omega)^f \sin\left(\frac{2\pi}{N_o} o\right), \tag{A.6}$$

where $f = 0, 1, \ldots, N_f$ and $o = 0, 1, \ldots, N_o$ are the frequency and orientation indexes. Finally, the Fourier domain representation of the filters has the form

$$\hat{G}_{fo}(\omega_x, \omega_y) = C e^{-\frac{(\omega_x - \Omega_{fo,x})^2}{2(\delta_\omega)^{2f}/\sigma^2}} e^{-\frac{(\omega_y - \Omega_{fo,y})^2}{2(\delta_\omega)^{2f}/\sigma^2}} \tag{A.7}$$

where C is a normalisation factor.

Each filter must of course be designed on a separate image. Filters obtained in this way can be directly employed for filtering in the Fourier domain, or anti-transformed for convolution filtering in the image domain.

The Fourier domain construction becomes tricky in the case of rectangular images, since the frequency resolution is different along the two axes. If one does not account for the different scaling, the filters become distorted in the image domain (see Section A.8.2).

A.7 Modified Gabor filters

The Head Detection and Tracking, Facial Feature Detection and Face Authentication applications presented in this Thesis actually make use of modified Gabor filters [9, 51], that have already been successfully applied in Texture Analysis [8]. These filters are represented by Gaussians in the Log-polar frequency plane rather than in the Fourier domain itself. That is, a filter tuned to orientation ϕ_0 and angular frequency $\omega_0 = \exp(\xi_0)$ has the form specified by Equation (2.8):

$$\hat{G}'(\xi, \phi) = Ae^{-\frac{(\xi-\xi_0)^2}{2\sigma_\xi^2}} e^{-\frac{(\phi-\phi_0)^2}{2\sigma_\phi^2}} \tag{A.8}$$

where A is a normalisation constant and (ξ, ϕ) are the coordinates on the Log-polar frequency plane:

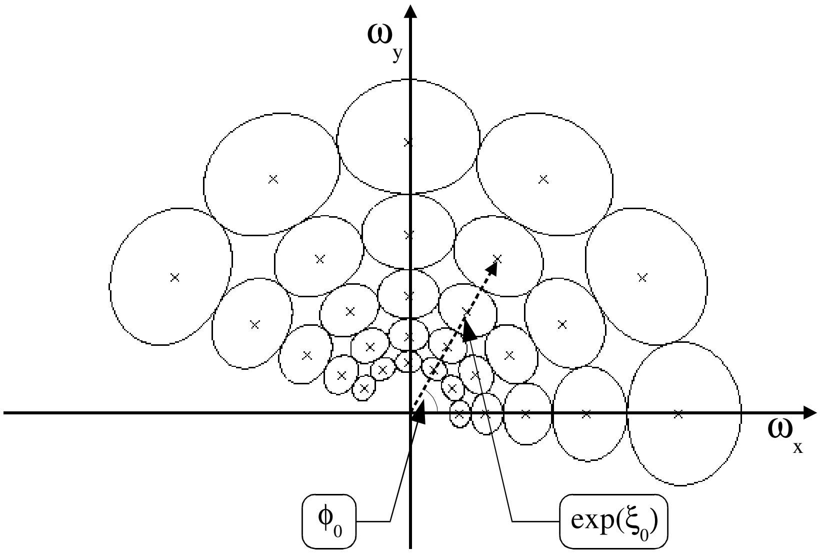
$$(\xi, \phi) = \big(\log \|(\omega_x, \omega_y)\|, \tan^{-1}(\omega_y/\omega_x)\big) \tag{A.9}$$

Although these filters involve the use of Log-polar mapping of the frequency plane their construction, as will be detailed in the following Section, is rather elegant and straightforward. It essentially involves drawing a set of *identical* Gaussians in the Log-polar frequency plane. The Gaussians are arranged on a rectangular grid (Figure A.1). This eliminates the need to rotate the Gaussians while computing different orientation channels and also allows choosing the amount of overlap in a convenient way. Furthermore, Gaussians belonging to different frequency channels have the same width in the Log-polar domain, since scaling is directly provided by the Log-polar mapping.

Besides the increased simplicity of design, another advantage of using the complex logarithm is that it guarantees a more uniform coverage of the frequency plane. The mapping compensates the tendency of standard Gabor filters to overlap too much towards the (densely sampled) origin of the frequency plane while leaving the (loosely sampled) high frequency regions poorly covered (see also Section 2.4.2 and Figure 2.3).

A.7.1 Constructing modified Gabor filters

Modified Gabor filters are easily constructed in the Fourier plane. They can then be inverse-transformed to the image domain if convolution filtering is preferred. It is convenient to design the filters on a square image, since the frequency resolution is in this



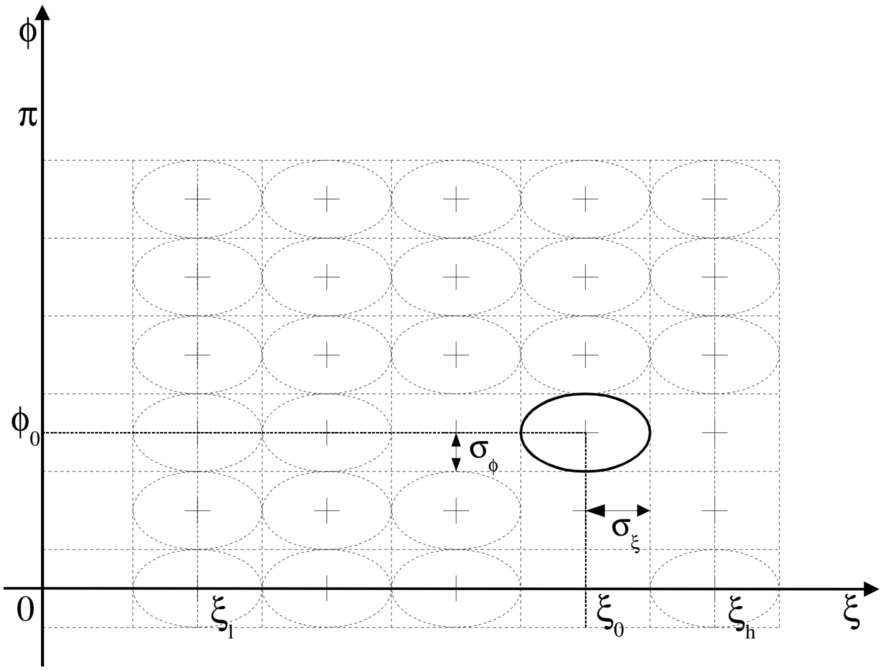


Figure A.1: A set of modified Gabor filters is displayed in the standard Fourier domain (top) and in the Log-polar frequency plane (bottom), in which the filters form a rectangular lattice of identical Gaussians. The contours mark the level curves of the Gaussians at the cell borders; the crosses mark the positions of the maxima. We note that, in going from the (ξ, ϕ) plane to the (ω_x, ω_y) plane, the complex exponential mapping introduces an asymmetric distortion that displaces the inner part of the level curves towards the maximum.

case the same along the two axes. In the following we will assume we are working in the Fourier domain of a square gray level image, which is initially zeroed. The choice of the dimensions of the image is not critical, since one is allowed to specify the wavelength of the filters in pixels directly, independent of the size of the frame. However, the image should be large enough to accommodate the widest (lowest frequency) filter. For our applications, We have found a 512×512 image to represent a convenient choice. The frame can at the end be cropped to match the size of the filters.

A.7.2 Tunable parameters

A set of modified Gabor filters constructed according to this procedure is determined by the following parameters:

- the number of frequency channels N_f,
- the number of orientation channels N_o,
- the angular frequency of the lowest frequency channel ω_l and
- the angular frequency of the highest frequency channel ω_h.

It might in practice be easier to estimate the range of filter wavelengths that are suitable for a given application (see Section A.5). If this is the case, the angular frequencies can be obtained from the wavelengths expressed in pixels using the relation $\omega = 2\pi/\lambda$.

A.7.3 Computing a modified filter

Given the above parameters, let us suppose we want to calculate the filter with orientation o in frequency channel f (the orientations and the frequency channels are numbered starting from zero). We proceed in the following way: first, we find the Log-polar coordinates of the centre of the Gaussian. Then, we map each pixel of our square image to the Log-polar domain and compute the corresponding value of the filter as given by Equation (A.8).

Finding the centre of the Gaussian

The N_o orientation channels cover the range from 0 to π. Therefore, the centre of the filter on the orientation axis can be computed as

$$\phi_0 = \frac{\pi}{N_o} o. \tag{A.10}$$

In the same way, in order to find the centre of the Gaussian on the logarithmic frequency axis ξ_o, we consider the range of angular frequencies defined by ω_l and ω_h. This translates, in Log-polar coordinates, to $\xi_l = \log(\omega_l)$ and $\xi_h = \log(\omega_h)$. The separation between the centres of the Gaussians on the ξ axis is therefore $\delta_\xi = (\xi_h - \xi_l)/(N_f - 1)$, so that we finally get

$$\xi_0 = \xi_l + f\delta_\xi \tag{A.11}$$

Computing the width of the Gaussian

We still have to determine the standard deviation of the Gaussian filter in the Log-polar frequency plane. This can be done once and for all, since all the filters are translated copies of the same Gaussian. It is convenient to specify the width of the filters by fixing its ratio α to the separation between the centres of adjacent filters along both axes. We therefore set $\sigma_\xi = \alpha\delta_\xi$ and $\sigma_\phi = \alpha\pi/N_o$. A value of 0.5 for α seems to be a reasonable choice. In this way, the amplitude of each filter is down to about 10% of its maximum value at the location of the maxima of the neighbouring filters.

Drawing the filter

We are now ready to actually draw the Fourier representation of the filter in an $N \times N$ pixel frame. It is customary to think of the Fourier domain as a "transformed" domain, into which data normally acquired in the direct, or image, domain are mapped via the Fourier transform. We would therefore like to stress once more that we are here entering data directly in the Fourier domain of an image, and that the pixels of the $N \times N$ frame (which we shall call Fourier frame) are to be interpreted as representing spatial frequencies rather than image points.

It is useful to think of the filter as "existing" as a Gaussian in the Log-polar frequency plane. In this way, designing the filter becomes a simple matter of finding, for each pixel in the Fourier frame, the corresponding location in the Log-polar frequency plane, where the value of the filter can be "looked up".

We assume that the pixels of the Fourier frame can be addressed via the pair of pixel coordinates (j, k), where $(0, 0)$ represents the centre of the frame. The orientation of the axes should be consistent with the conventions of one's preferred implementation of the Fast Fourier Transform, so that the (zeroed) Fourier frame can be thought of as the *output* of the FFT of a zeroed image. Specifically, we will think of the j axis as the transform of the x axis in the image and of the k axis as corresponding to the y axis. The Fourier frame pixel (j, k) can then be labelled with the corresponding angular frequencies (ω_x, ω_y), which according to the definition of the Discrete Fourier Transform are

$$(\omega_x, \omega_y) = \frac{2\pi}{N}(j, k). \tag{A.12}$$

In order to map the pixel to the Log-polar frequency plane, it is now sufficient to apply the coordinate transformation in Equation (A.9). This yields

$$\xi = \log \|(\omega_x, \omega_y)\| = \log\left(\frac{2\pi}{N}\sqrt{j^2 + k^2}\right) \tag{A.13}$$

and

$$\phi = \tan^{-1}(\omega_y/\omega_x) = \tan^{-1}(k/j). \tag{A.14}$$

"Looking up" the corresponding value of the filter now simply amounts to substituting the Log-polar coordinates (ξ, ϕ) into Equation (A.8). By iterating this procedure over all the pixels (j, k) in the Fourier frame, the Fourier domain representation of the filter is constructed. The typical appearance of a modified Gabor filter in the Fourier domain is shown in Figure A.2 (a). The Fourier frame can then be inverse-transformed to obtain the image domain representation of the filter; an example is reported in Figure A.2 (b–d).

A.7.4 Debugging tips

It is advisable to display all the filters both in the frequency and in the image domain in order to check that they have the expected appearance. This is probably the best way to find errors in the code. One should then try filtering images with simple geometrical structures, such as a star-like pattern centred in a dark frame, and qualitatively compare the filter responses with the expectations. Most of the bugs are typically due to the infamous phase wraparound problem.

The algorithm described in Section A.7.3 above allows one to specify the frequency behaviour of a set of filters by giving the number of desired frequency channels along with the longest and shortest wavelengths. The width σ_ξ of the filters is determined on the basis of these parameters. However, one should take care that the filters do not come out so large that they don't fit in the image. This can be done by verifying the following two inequalities. On the low frequency side, one should make sure that the centres of the filters in the lowest frequency channel are well above the lowest representable frequency $||(\omega_x, \omega_y)|| = 2\pi/N$. The inequality

$$\xi_l - 3\sigma_\xi > \log(2\pi/N) \tag{A.15}$$

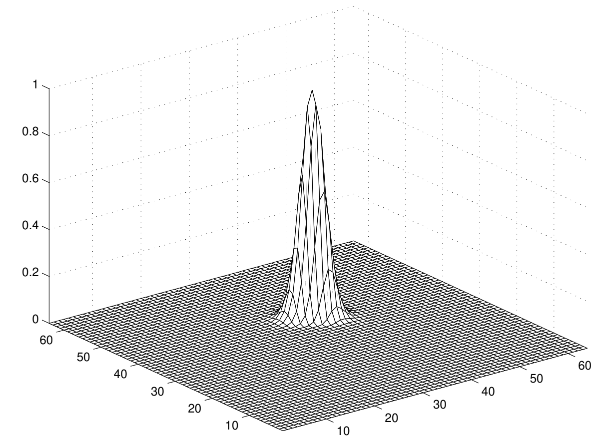
represents a good rule of thumb. On the other hand, the highest frequency channel should not be too close to the maximum representable angular frequency $||(\omega_x, \omega_y)|| = \pi$, which is to say

$$\xi_h + 3\sigma_\xi < \log(\pi). \tag{A.16}$$

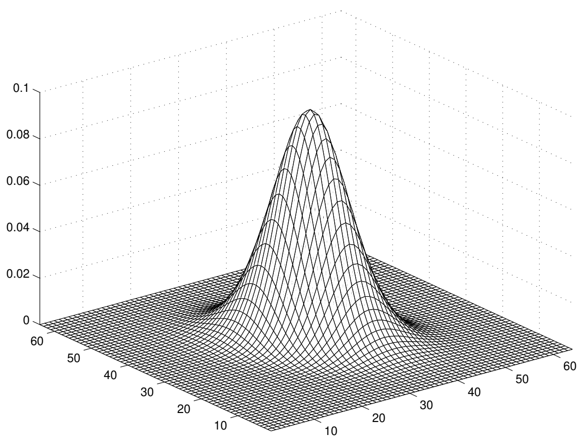
A.8 Final remarks

A.8.1 Normalisation of the filters

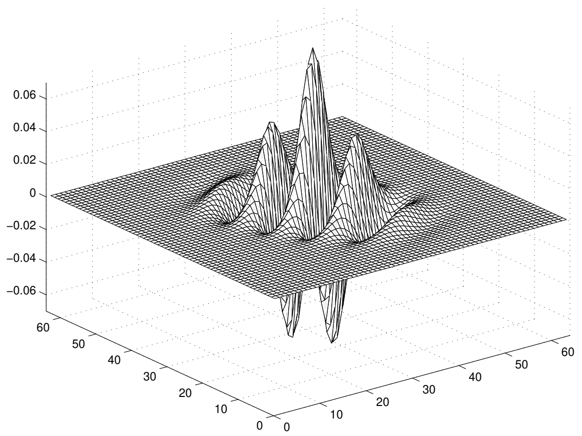
Normalisation of the filters depends both on the application and on the intended use (in the frequency or in the image domain). The filters employed in this Thesis were



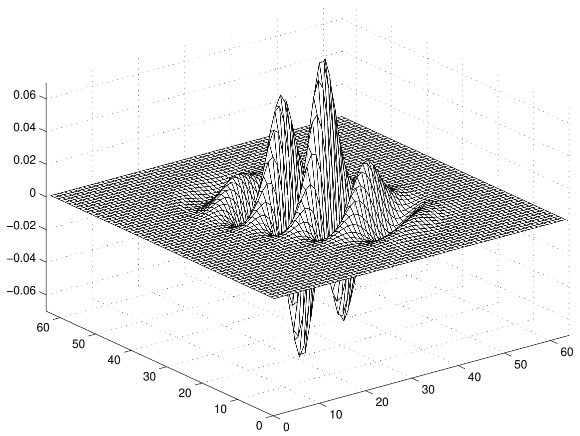
(a) Fourier domain representation of the filter.



(b) Filter magnitude (image domain).



(c) Real part of the filter (image domain).



(d) Imaginary part of the filter (image domain).

Figure A.2: Example of a modified Gabor filter drawn on a 64×64 square frame. All graphs are in pixel coordinates. The Fourier domain representation (a) is real-valued; for the image domain representation the filter magnitude (b), the real part (c) and the imaginary part (d) are shown.

constructed as indicated in Section A.7.3, then anti-transformed to the image domain and normalised to unity according to the L^2 norm. Another option consists in choosing a different normalisation for each frequency channel in order to compensate for the decay towards high frequencies in the spectrum of natural images [31].

If the images must be Fourier transformed for filtering, then the normalisation of the filters will in general have to be consistent with the definition of the Fourier transform.

A.8.2 Square and rectangular images

If the filters are to be used on square images for frequency domain filtering, it is convenient to set the size N of the Fourier frame equal to the size of the images. The filters designed according to the algorithm in Section A.7.3 can then be employed directly without further modifications.

Fourier domain filtering on a rectangular image is slightly trickier, since the frequency resolution along the two axes in the frequency plane is not the same. An easy way of overcoming this difficulty is as follows. Draw the filters in the Fourier domain as explained, using a large square Fourier frame; anti-transform them to the image domain and crop them so that they are centred in a rectangular image of the desired size. At this point, simply Fourier transform this image to get the correct frequency domain representation of the filters in the rectangular Fourier frame.

If image domain filtering on either a square or a rectangular image is desired (as was the case in the present Thesis), one can simply draw the filters on a (large) square Fourier frame, anti-transform them to the image domain, crop them and normalise them as required.

B Databases

In this Thesis we presented Facial Feature Detection and Face Authentication results on the subsets of frontal images from both the M2VTS and the Extended M2VTS (XM2VTS) databases.

These large multi-modal databases have been collected by the M2VTS Consortium specifically for identity verification purposes. They feature audio and video sequences of a number of volunteers, namely 37 persons for the M2VTS database and 295 for the XM2VTS database. Each subject has been acquired on four different occasions separated by a significant time interval. The four sessions in the XM2VTS database were collected over a period of four months.

The presence of multiple sessions of the same person constitutes the specificity and the major advantage of the M2VTS databases with respect to Face Authentication. Testing a Face Authentication algorithm also involves assessing the robustness of the system with respect to variations of a subject's appearance over time, including changes in tan, makeup, hair style, beard/moustaches, expression, the presence or absence of eyeglasses, etc. This variability is well represented in both the M2VTS databases (Figure B.1), which allows drawing significant False Rejection curves. Many other databases, including FERET, are not representative of such within-class client spread, a fact that reduces the reliability of False Rejection curves.

The M2VTS Consortium has also established standardised test protocols for their databases. These protocols, to which we adhered in the Face Authentication experiments reported in Sections 6.4 and 6.5 and throughout Chapter 7, allow an immediate comparison of the results obtained by different research groups. For these reasons, the M2VTS databases are becoming a reference for a significant part of the Face Authentication community.

The set of frontal images from the M2VTS database we employed (the database also contains non-frontal frames) contains a total of 551 images of the 37 subjects, that is 3–4 portraits of each person in each of the four sessions. The frames, originally in colour, were converted to gray level and down-sampled to a resolution of 174×143 pixels, corresponding to an average inter-eye distance of 29 pixels.

A set of frontal images from the XM2VTS database is directly released by the M2VTS Consortium as a CD-ROM distribution. It includes 2 image frames per session for each of its 295 subjects, yielding a total of 2360 frames. Again we converted these images to gray level and we down-sampled them to a resolution of 180×144 pixels, which resulted in an average inter-eye distance of 27 pixels.

Figure B.1: Subjects display significant changes in appearance across the four sessions of the XM2VTS database. Variations include, for the same subject, changes in hair style, makeup, expression, beard/moustaches, eyeglasses.

Bibliography

[1] Y. Abdeljaoued, J. Bigun, B. Duc, and S. Fischer. Node significance determination for face matching by Bayesian statistics. *Accepted to the journal: Image and Vision Computing*, 1998.

[2] A. N. Akansu and R. A. Addad. *Multiresolution Signal Decomposition.* Academic Press, 1992. ISBN 0-12-047140-X.

[3] J.-P. Antoine, P. Carrette, R. Murenzi, and B. Piette. Image analysis with two-dimensional continuous wavelet transform. *Signal Processing*, 31:241–272, 1993.

[4] H. Bårman, L. Haglund, H. Knutsson, and G. H. Granlund. Estimation of velocity, acceleration and disparity in time sequences. In *Proceedings of IEEE Workshop on Visual Motion*, pages 44–51, Princeton, NJ, USA, October 1991.

[5] M. J. Bastiaan. A sampling theorem for the complex spectrogram, and Gabor's expansion of a signal in Gaussian elementary signals. *Optical engineering*, 20(4):594–598, 1981.

[6] D. E. Benn, M. S. Nixon, and J. N. Carter. Robust eye centre extraction using the Hough transform. In *Proceedings of the 1st international conference on Audio- and Video-based Biometric Person Authentication (AVBPA'97), Crans-Montana, Switzerland*, pages 3–9, March 1997.

[7] D. E. Benn, M. S. Nixon, and J. N. Carter. Extending concentricity analysis by deformable templates for improved eye extraction. In *Proceedings of the 2nd international conference on Audio- and Video-based Biometric Person Authentication (AVBPA'99), Washington DC, U.S.A.*, pages 1–6, March 1999.

[8] J. Bigun. Gabor phase in boundary tracking and region segregation. In *proceedings of DSP & CAES conf. Nicosia, Cyprus*, pages pp. 229–237. Univ. of Nicosia, July 14-16 1993.

[9] J. Bigun. Speed, frequency, and orientation tuned 3-D Gabor filter banks and their design. In *Proceedings of International Conference on Pattern Recognition, ICPR, Jerusalem*, pages C–184–187. IEEE Computer Society, 1994.

[10] J. Bigun and J. M. H. du Buf. N-folded symmetries by complex moments in Gabor space. *IEEE-PAMI*, 16(1):80–87, 1994.

[11] J. Bigun, B. Duc, S. Fischer, A. Makarov, and F. Smeraldi. Multi modal person authentication. In H. Wechsler et al., editor, *Nato-Asi advanced study on face recogniton*, volume F 163, pages 26–50. Springer, 1998.

[12] J. Bigun, G. H. Granlund, and J. Wiklund. Multidimensional orientation estimation with applications to texture analysis and optical flow. *IEEE-PAMI*, 13(8):775–790, August 1991.

[13] C.M. Bishop. *Neural networks for pattern recognition.* Oxford University Press, Oxford, 1995.

[14] A. C. Bovik, M. Clark, and W. S. Geisler. Multichannel texture analysis using localized spatial filters. *IEEE-PAMI*, 12(1), 1990.

[15] C. J. C. Burges. A tutorial on Support Vector Machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):1–47, 1998.

[16] P. J. Burt. *"Smart sensing" in machine vision*, chapter in Machine Vision: algorithms, architectures, and systems (H. Freeman Ed.), pages 1–30. Academic Press, 1988.

[17] R. H. S. Carpenter, editor. *Eye movements*, volume 8 of *Vision and visual dysfunction.* Macmillan Press, 1991.

[18] R. Chellappa, C. L. Wilson, and S. Sirohey. Human and machine recognition of faces: a survey. *Proceedings of the IEEE*, 83(5):705–740, 1995.

[19] C. Cortes and V. Vapnik. Support–Vector Networks. *Machine Learning*, 20:273–297, 1995.

[20] I. Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, 36(5):961–1005, September 1990.

[21] I. Daubechies. *Ten lectures in wavelets.* Society for industrial and applied mathematics, publisher, 3600 University city science center, Philadelphia, PA, 19104-2688 USA, ISBN 0-89871-274-2, 1992.

[22] J. G. Daugman. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, 20:847–856, 1980.

[23] J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Am. A*, 2(7):1160–1169, July 1985.

[24] J. G. Daugman. Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE ASSP*, 36:1169–1179, 1988.

[25] I. Debusschere, E. Bronckaers, C. Claeys, G. Kreider, J. Van der Spiegel, P. Bellutti, G. Soncini, P. Dario, F. Fantini, and G. Sandini. A 2D retinal CCD sensor for fast 2D shape recognition and tracking. In *Proceedings of the 5th International Conference on Solid-State Sensors and Transducers, Montreux, Switzerland*, June 1989.

[26] P. A. Devijver and J. Kittler. *Pattern Recognition: a Statistical Approach.* Prentice–Hall International, London, 1982.

[27] J. Dias, I. Fonseca, and J. Lobo. New perspectives on mobile robot navigation with visual and inertial information. In A. T. Almeida and H. Araujo, editors, *Proceedings of the 5th International Workshop on Advanced Motion Control, Coimbra (Portugal)*, pages 261–266. IEEE Cat. Num. 98TH8354, IEEE Service Center, 445 Hoes Lane, P.O. Box 1331 Piscataway, NJ 08855-1331, USA, 1998.

[28] B. Duc, S. Fischer, and J. Bigün. Face authentication with Gabor information on deformable graphs. *IEEE Transactions on Image Processing*, 8(4):504–516, April 1999.

[29] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis.* John Wiley and Sons, 1973.

[30] B. Efron and R. J. Tibshirani. *An introduction to the Bootstrap.* Chapman & Hall, New York, 1993.

[31] D.J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of Optical Society of America*, A4:2379–2394, 1987.

[32] S. Fischer. *Multi-scale analysis for pattern recognition.* PhD thesis, Swiss Federal Institute of Technology – Lausanne, Switzerland, No. 1744 (1997).

[33] S. Fischer and J. Bigun. Texture boundary tracking with Gabor phase. In G. Borgefors, editor, *Theory and Applications of Image Analysis II - Selected papers from the 9th Scandinavian Conf. on Image Analysis*, pages 101–112, Singapore, 1995. World Scientific.

[34] S. Fischer and B. Duc. Shape normalisation for face recognition. In *Proceedings of the 1st international conference on Audio- and Video-based Biometric Person Authentication (AVBPA'97), Crans-Montana, Switzerland*, pages 21–26, March 1997.

[35] D. Fleet, A. Jepson, and M. Jenkin. Phase-based disparity measurement. *Computer Vision, Graphics and Image Processing*, 53(2):198–210, 1991.

[36] D. J. Fleet and A. D. Jepson. Computation of component image velocity from local phase information. *International Journal of Computer Vision*, 5(1):77–104, 1990.

[37] D. Gabor. Theory of communication. *Journal of the IEE*, 93:429–457, 1946.

[38] P. Goupillaud, A. Grossmann, and J. Morlet. Cycle-octave and related transforms in seismic signal analysis. *Geoexploration*, 23(1984/85):85–102, 1984.

[39] A. Grossmann and J. Morlet. Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM J. Math. Anal.*, 15(4):723–736, July 1984.

[40] J. Huang, C. Liu, and H. Wechsler. Eye detection and face recognition using evolutionary computation. In H. Wechsler et al., editor, *Nato-Asi advanced study on face recogniton*, volume F 163. Springer, 1998.

[41] D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *J. Physiology (London)*, 148, 1959.

[42] D. H. Hubel and T. N. Wiesel. Brain mechanisms of vision. *Scientific American*, 241(3):150–162, September 1979.

[43] D.H. Hubel. *Eye, brain and vision.* Scientific American Library, 1988.

[44] A. J. Jain and S. Battacharjee. Text segmentation using Gabor filters for automatic document processing. *Machine vision and applications*, 5:169–184, 1992.

[45] A. K. Jain and F. Farrokhnia. Unsupervised texture segmentation using Gabor filters. *Pattern Recognition*, 24(12):1167–1186, 1991.

[46] R. Jain, S. L. Bartlett, and N. O'Brien. Motion stereo using ego-motion complex logaritmic mapping. *"IEEE-PAMI Transactions on Pattern Analysis and Machine Intelligence"*, 9(3):356–369, May 1987.

[47] T. Joachims. *Making Large-scale SVM Learning Practical*, chapter 11 of Advances in Kernel Methods - Support Vector Learning. MIT Press, 1998. Eds. B. Schölkopf, C. J. C. Burges, A. J. Smola.

[48] K. Jonsson, J. Matas, and J. Kittler. Saliency-based robust correlation for real-time face registration and verification. In M. S. Nixon, editor, *Proc British Machine Vision Conference BMVC98*, pages 44–53, 1998.

[49] S. L. Marple Jr. *Digital spectral analysis with applications.* Prentice-Hall, 1987. ISBN 0-13-214149-3.

[50] C. F. Keating and E. G. Keating. Monkeys and mug shots: cues used by Rhesus monkeys (Macaca mulatta) to recognize a human face. *Journal of Comparative Psychology*, 107(2):131–139, 1993.

[51] H. Knutsson. *Filtering and reconstruction in image processing.* Number 88 in Linköpings Studies in Science and Technology: Dissertations. Linköping University, 1982. ISBN 91-7372-595-1.

[52] J. Kopecz, C. Engels, and H. Janssen. Flexibility through a neural architecture for visual orientation in a natural environment. *Physica D*, 75, 1994.

[53] C. Kotropoulos and I. Pitas. Rule–based face detection in frontal views. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'97)*, volume IV, pages 2537–2540, April 1997.

[54] C. Kotropoulos, A. Tefas, I. Pitas, C. Fernandez, and F. Fernández. Performance assessment of morphological dynamic link architecture under optimal and real operating conditions. In *Proceedings of the IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP'99), Antalya (Turkey)*, pages 355–359, June 1999.

[55] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Hurtz, and W. Konen. Distortion invariant object recognition in the dynamic link architectures. *IEEE Trans. on Computers*, 42(3):300–311, March 1993.

[56] L. S. Lasdon. *Optimization theory for Large Systems.* Operations Resarch Series. MacMillan, 1970.

[57] C. Liu and H. Wechsler. A unified Bayesian framework for face recognition. In *International Conference on Image Processing, ICIP-98, Chicago, October 4-7*, volume 1, pages 151–155, 1998.

[58] D. G. Luenberger. *Optimization by vector space methods.* John Wiley and Sons, 1969.

[59] B. S. Manjunath, C. Shekhar, and R. Chellappa. A new approach to image feature detection with applications. *Pattern Recognition*, 31:627–640, 1996.

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

[61] J. Matas, K. Jonsson, and J. Kittler. Fast face localisation and verification. In A. F. Clark, editor, *Proceedings of the British Machine Vision Conference (BMVC97)*, 1997.

[62] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: the Extended M2VTS database. In *Proceedings of the 2nd international conference on Audio- and Video-based Biometric Person Authentication (AVBPA'99), Washington DC, U.S.A.*, pages 72–77, 1999.

[63] S. Mika, G. Rätsch J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Neural Networks for Signal Processing IX*, pages 41–48. IEEE, 1999.

[64] B. Moghaddam and A. Pentland. Beyond linear eigenspaces: Bayesian matching for face recognition. In H. Wechsler et al., editor, *Nato-Asi advanced study on face recogniton*, volume F 163, pages 230–243. Springer, 1998.

[65] J. Morlet. Sampling theory and wave propagation. In C. H. Chen, editor, *Issues in acoustic signal – image processing and recognition : (proceedings of the NATO advanced research workshop; San Miniato, August 5-9, 1982)*, volume 1 of *series F*, pages 233–261. NATO-ASI, Springer, Berlin, 1983.

[66] D. W. Murray, K. J. Bradshaw, P. F. Mclauchlan, I. D. Reid, and P. M. Sharkey. Driving saccade to pursuit using image motion. *International Journal of Computer Vision*, 16:205–228, 1995.

[67] A. Nikolaidis, C. Kotropoulos, and I. Pitas. Facial feature extraction using adaptive Hough transform, template matching and active contour models. In *Proceedings of the 13th International Conference on Digital Signal Processing (DSP'97), Santorini (Greece)*, volume 2, pages 865–868, July 1997.

[68] K. Okada, J. Steffens, T. Maurer, H. Hong, E. Elagin, H. Neven, and C. v. d. Malsburg. The Bochum/USC face recognition system and how it fared in the FERET phase III test. In H. Wechsler et al., editor, *Nato-Asi advanced study on face recogniton*, volume F 163, pages 186–205. Springer, 1998.

[69] G. A. Orban. *Neuronal operations in the visual cortex.* studies of brain functions. Springer, 1984.

[70] E. Osuna, R. Freund, and F. Girosi. Improved training algorithm for Support Vector Machines. In *Proceedings of IEEE NNSP'97*, pages 276–285, September 1997.

[71] C. P. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Proceedings of the International Conference on Computer Vision (ICCV'98), Bombay, India*, pages 555–562, January 1998.

[72] J. B. Pelz. *Visual representations in a natural visuo-motor task.* PhD thesis, Carlson Center for Imaging Science, Rochester Institute of Technology, 1995.

[73] M. Pontil and A. Verri. Properties of Support Vector Machines. *Neural Computation*, 10:977–996, 1998.

[74] M. Porat and Y. Y. Zeevi. The generalized Gabor scheme of image representation in biological and machine vision. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 10(4):452–468, July 1988.

[75] R. P. N. Rao and D. H. Ballard. An active vision architeture based on iconic representations. *Artificial Intelligence Journal*, 78:461–505, 1995.

[76] R. P. N. Rao, G. J. Zelinsky, M. M. Hayhoe, and D. H. Ballard. Modelling saccade targeting in visual search. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing System*, volume 8 (NIPS'95), pages 830–836. MIT Press, 1996.

[77] R. P. N. Rao, G. J. Zelinsky, M. M. Hayhoe, and D. H. Ballard. Eye movements in visual cognition: a computational study. Technical Report 97.1, National Resource Laboratory for the Study of Brain and Behavior, Department of Computer Science, University of Rochester, 1997.

[78] M. U. Ramos Sánchez, J. Matas, and J. Kittler. Statistical chromaticity models for lip tracking with B-splines. In *Proceedings of the 1st international conference on Audio- and Video-based Biometric Person Authentication (AVBPA'97), Crans-Montana, Switzerland*, pages 69–76, March 1997.

[79] G. Sandini and V. Tagliasco. An antropomorphic retina-like structure for scene analysis. *Computer Graphics and Image Processing*, 14:365–372, 1980.

[80] G. Sandini and M. Tistarelli. *Vision and space-variant sensing*, volume 1, chapter II.11. Academic Press, 1992.

[81] B. Scholköpf. *Support vector learning.* R. Oldenbourg Verlag, Munich, 1997.

[82] B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors. *Advances in Kernel Methods - Support Vector Learning.* MIT Press, 1998. ISBN 0-262-19416-3.

[83] B. Schölkopf, A. J. Smola, and K.-R. Müller. *Kernel Principal Component Analysis*, chapter 20 of Advances in Kernel Methods - Support Vector Learning. MIT Press, 1998. Eds. B. Schölkopf, C. J. C. Burges, A. J. Smola.

[84] E. L. Schwartz. Computational anatomy and functional architecture of striate cortex: A spatial mapping approach to perceptual coding. *Visual Research*, 20:645–669, 1980.

[85] A. Shashua. On the relationship between the Support Vector Machine for classification and sparsified Fisher's Linear Discriminants. *Neural Processing Letters*, 9:129–139, 1999.

[86] L. Sirovich and M. Kirby. Application of the Karhunen-Loève procedure for the characterization of human faces. *IEEE-PAMI Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.

[87] F. Smeraldi and J. Bigün. Retinal vision applied to facial features detection and face authentication. *Submitted to the IEEE-PAMI Transactions on Pattern Analysis and Machine Intelligence*, 1999.

[88] F. Smeraldi, J. Bigün, and W. Gerstner. Face space dimensionality determination in the context of face authentication. *Submitted to the IEEE-PAMI Transactions on Pattern Analysis and Machine Intelligence*, 1999.

[89] F. Smeraldi, J. Bigün, and W. Gerstner. On the role of dimensionality in face authentication. In *Proceedings of the Symposium of the Swedish Society for Automated Image Analysis, Halmstad (Sweden)*, pages 87–91. Halmstad University, P.O. Box 823, S-301 18 Halmstad (Sweden), March 2000.

[90] F. Smeraldi, N. Capdevielle, and J. Bigün. Face authentication by retinotopic sampling of the Gabor decomposition and Support Vector Machines. In *Audio and Video based Biometric Person Authentication – AVBPA99*, pages 125–129, 1999.

[91] F. Smeraldi, N. Capdevielle, and J. Bigün. Facial features detection by saccadic exploration of the Gabor decomposition and support vector machines. In *Proceedings of the 11th Scandinavian Conference on Image Analysis – SCIA 99, Kangerlussuaq, Greenland*, volume I, pages 39–44, June 1999.

[92] F. Smeraldi, O. Carmona, and J. Bigün. Real–time head tracking by saccadic exploration and Gabor decomposition. In A. T. Almeida and H. Araujo, editors, *Proceedings of the 5th International Workshop on Advanced Motion Control, Coimbra (Portugal)*, pages 684–687. IEEE Cat. Num. 98TH8354, IEEE Service Center, 445 Hoes Lane, P.O. Box 1331 Piscataway, NJ 08855-1331, USA, 1998.

[93] F. Smeraldi, O. Carmona, and J. Bigün. Saccadic search with Gabor features applied to eye detection and real-time head tracking. *Image and Vision Computing*, 18(4):323–329, 2000.

[94] K. Sobottka and I. Pitas. Extraction of facial regions and features using color and shape information. In *Proceedings of the International Conference on Pattern Recognition (ICPR'96), Vienna (Austria)*, volume III, pages C421–C425, August 1996.

[95] B. Takács and H. Wechsler. Face localization using a dynamic model of retinal feature extraction. In M. Bichsel, editor, *Proceedings of the International workshop on automatic face and gesture recognition, Zurich*, pages 243–247. Multimedia Laboratory; Univ. of Zurich; Winterthurers. 190 CH-8057 Zurich, June 26-28 1995.

[96] B. Takács and H. Wechsler. A saccadic vision system for landmark detection and face recognition. In H. Wechsler et al., editor, *Nato-Asi advanced study on face recogniton*, volume F 163, pages 627–636. Springer, 1998.

[97] A. Tefas, C. Kotropoulos, and I. Pitas. Variants of dynamic link architecture based on mathematical morphology for frontal face authentication. In *Proceedings of the IEEE Computer Society conference on Computer Vision and Pattern Recognition (CVPR 98), Santa Barbara (CA), U.S.A.*, pages 814–819, June 1998.

[98] A. Tefas, C. Kotropoulos, and I. Pitas. Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication. *Submitted to the IEEE-PAMI Transactions on Pattern Analysis and Machine Intelligence*, May 1999.

[99] M. Tistarelli and E. Grosso. Active vision-based face recognition: issues, applications and techniques. In H. Wechsler et al., editor, *Nato-Asi advanced study on face recogniton*, volume F 163, pages 262–286. Springer, 1998.

[100] M. Tistarelli and G. Sandini. On the advantages of log-polar mapping for direct estimation of time–to–impact from optical flow. *IEEE Transactions on Pattern Analysis and Machine Recognition*, 15(4):401–410, 1993.

[101] F. Tong and Z. N. Li. The Reciprocal-Wedge Transform for space-variant sensing. In *Proceedings of the 4th International Conference on Computer Vision, Berlin, Germany*, pages 330–334, May 1993.

[102] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience (Winter)*, 3(1):71–86, 1991.

[103] V. N. Vapnik. *The nature of statistical learning theory.* Springer–Verlag, 1995.

[104] Web Site. American National Institute of Standards and Technology (NIST). http://www.nist.gov/speech/.

[105] A. L. Yarbus. *Eye movements.* Plenum, New York, 1967.

[106] W. Zhao, A. Krishnaswami, R. Chellappa, D. L. Swets, and J. Weng. Discriminant analysis of principal components for face recognition. In H. Wechsler et al., editor, *Nato-Asi advanced study on face recogniton*, volume F 163, pages 73–85. Springer, 1998.

Curriculum Vitae

Fabrizio Smeraldi was born in Chiavari (Italy) on April 5th, 1972.

In 1991 he earned his General Certificate of Education ("Maturità Scientifica") with full marks (60/60) at Liceo Scientifico Statale "G. Marconi", Chiavari, Italy. In the same year he gained admission to a two–week tutorial on management and communication strategies held at Università Commerciale Luigi Bocconi (Milan, Italy).

In 1994, after a national selection, he attended a one–week seminar on information technologies organised in Spoleto (Italy) by the IBM Italia Foundation.

In 1996 he completed his MSc in Physics with full marks and honours (Laurea in Fisica, 110 e Lode) at the University of Genoa (Italy), with a thesis on Quantum Computation and Cryptography.

In 1997 he joined as Research Assistant the Computer Vision Group led by Dr. J. Bigün at the Microprocessor and Interface Laboratory of the Swiss Federal Institute of Technology (EPFL), Lausanne. He took active part in the development of novel face authentication algorithms for the IT–LTR European Project VIRSBS (Visual Intelligent Recognition for Secure Banking Services). Meanwhile, he registered as PhD student at EPFL.

In 1999 he moved to the Centre for Neuro-Mimetic Systems (MANTRA) led by Prof. W. Gerstner at EPFL. Besides continuing his research activity in Computer Vision and supervising diploma projects at EPFL, he is also involved in a Swiss CTI (Commission for Technology and Industry) Project concerning the simulation of the dynamics of complex mechanical systems by means of artificial neural networks.

Publications

Articles in Journals

F. Smeraldi, J. Bigün and W. Gerstner: *Face space dimensionality determination in the context of face authentication*, IEEE Transactions on Pattern Analysis and Machine Intelligence, submitted, October 1999

F. Smeraldi and J. Bigün: *Retinal vision applied to facial features detection and face authentication*, IEEE Transactions on Pattern Analysis and Machine Intelligence, submitted, July 1999

F. Smeraldi, O. Carmona and J. Bigün: *Saccadic search with Gabor features applied to eye detection and real-time head tracking*, Image and Vision Computing, volume 18, number 4, pages 323–329, 2000

Articles in Conference Proceedings

F. Smeraldi, J. Bigün and W. Gerstner: *On the role of dimensionality in face authentication*, in Proceedings of the Symposium of the Swedish Society for Automated Image Analysis, Halmstad (Sweden), pages 87–91, published by: Halmstad University, P.O. Box 823, S-301 18 Halmstad (Sweden), March 2000.

F. Smeraldi, N. Capdevielle and J. Bigün: *Facial features detection by saccadic exploration of the Gabor decomposition and Support Vector Machines*, in Proceedings of the 11th Scandinavian Conference on Image Analysis, Kangerlussuaq (Greenland), volume I, pages 39–44, June 1999

F. Smeraldi, N. Capdevielle and J. Bigün: *Face authentication by retinotopic sampling of the Gabor decomposition and Support Vector Machines*, in Proceedings of the 2nd International Conference on Audio and Video Based Biometric Person Authentication (AVBPA'99), Washington DC (USA), pages 125–129, March 1999

F. Smeraldi, N. Capdevielle and J. Bigün: *Gabor decomposition driven saccades applied to facial features detection for visual biometrics applications*, in Proceedings of the 1999 Symposium of the Swedish Society for Automated Image Analysis, Göteborg (Sweden), pages 85–88, published by: School of Signals and Systems, Chalmers University of Technology, Göteborg, March 1999.

F. Smeraldi and J. Bigün: *Facial features detection by saccadic exploration of the Gabor decomposition*, in Proceedings of the 1998 International Conference on Image Processing (ICIP'98), Chicago (USA), volume III, pages 163–167, 1998

F. Smeraldi, O. Carmona and J. Bigün: *Real-time head tracking by saccadic exploration and Gabor decomposition*, in Proceedings of the 5th International Workshop on Advanced Motion Control (AMC'98), Coimbra (Portugal), pages 684–687, IEEE Cat. Num. 98TH8354, IEEE Service Center, 445 Hoes Lane, P.O. Box 1331 Piscataway, NJ 08855-1331, USA, 1998

J. Bigün, B. Duc, S. Fischer, A. Makarov and F. Smeraldi: *Multi modal person authentication*, in Proceedings of the Nato-Asi advanced study on face recognition, H. Wechsler et al. editors, pages 26–50, volume F 163, Springer, 1998