Morphological Techniques for Face Localization

by

Bogdan Raducanu

B.S., University "Politehnica" of Bucharest (1995)

Submitted to the Department of Computer Science and Artificial Intelligence
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

UNIVERSITY OF THE BASQUE COUNTRY

July 2001

Signature of Author . . . .
Department of Computer Science and Artificial Intelligence
25 May 2001

Certified by. . . . .
Manuel Graña
Professor, Computer Science and Artificial Intelligence Department
Thesis Supervisor

Accepted by . . . .

Morphological Techniques for Face Localization

by

Bogdan Raducanu

Submitted to the Department of Computer Science and Artificial Intelligence
on 25 May 2001, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The scope of the research work reported in this thesis is the experimental evaluation of face localization methods based on mathematical morphology. Our aim is to obtain more robust detection against linear and non-linear transformations in both spatial and intensity image domains. The morphological approaches we tested are: Morphological Multiscale Fingerprints obtained from the Morphological Scale Space analysis of the images, Gray scale Hit-or-Miss Transform and Morphological Associative Memories. The experiments were performed on a custom image database and on the image database from the CMU.

Thesis Supervisor: Manuel Graña
Title: Professor, Computer Science and Artificial Intelligence Department

Contents

List of Figures

To my family

Acknowledgments

Many persons contributed to the success of this thesis, and I wish to use this opportunity in order to express my personal consideration for their continuous support.

On the personal level, I would like to address special thanks to my uncle Dan, for being the person who proposed to me to realize my Ph.D. thesis and the initiator of this project. Together with my mother, Gabriela and my aunt, Steluta they all offered me, in all these years, day after day, an unconditional support to make my dreams come true and for this reason I will remain grateful to them for the rest of my life.

On the professional level, the first person to whom I would like to express my gratitude is the Ph. D. thesis advisor, professor Manuel Graña. Without his permanent support, encouragements and enthusiasm for research, this thesis hasn't been possible. Thanks to him, I have reached a professional milestone that may entirely change my whole life ahead.

I would like to thank professor Leon Rothkrantz and my friend Maja Pantic, from TU Delft, The Netherlands, for receiving me as a visiting researcher in their department and offering all the conditions to realize my research activity.

I want to thank to my friends Iñaki Rañó, Iñaki Inza and Jose-Manuel Peña for the ideas aroused from our talks regarding artificial intelligence. To my lab colleagues Ana and Imanol for useful ideas and helping in implementing some algorithms in IDL and performing some experiments. Also thanks to my ex-colleagues, Virginia, Monica, Richar, Txema and Israel for their support and everyday encouragements.

Besides the persons mentioned above, there are many other people, members from my department and my faculty, worldwide friends that support me and encouraged me to realize my thesis. Because they are so many, I wouldn't like to give their names in order not to miss someone. I can guarantee them, they will remain forever in my heart for everything that they did for me.

And finally, support for this thesis has been provided, in part, by a predoctoral grant offered to me by the University of The Basque Country.

Realizing my Ph. D. thesis, it didn't mean only a professional breakthrough and the fulfillment of a wish, but represented also the dawns of an excitement journey of amazing self-discoveries.

Contributions

The relevant contributions of this thesis, can be summarized as follows:

- We have done a comprehensive review of the literature up to the present date regarding the state of the art of the face localization problem.

- Some authors have proposed an generalization of the definition of the Hit-or-Miss Transform to gray-level images. We have tested the application of this gray scale Hit-or-Miss Transform to the task of face localization. We have contributed a definition of a gray scale Hit-or-Miss Transform based on level sets, and we tested its application to face localization.

- The Scale Space image analysis is a well-known approach to robust image analysis. We have adopted the Morphological Scale Space defined by Jackway using multiscale erosions and dilations. Fingerprints, in this approach, are defined as the local extrema of the filtered images. We have tested the use of these fingerprints as features for face localization in gray scale images.

- In the field of Artificial Neural Networks, recent works have proposed a new architecture based on morphological operators: the Morphological Associative Memories. We have proposed an algorithm for face localization based on Morphological Heteroassociative Memories and Morphological Scale Space based on multiscale erosions and dilations.

- The construction of our own image database and a recollection of diverse face databases. Some of the collected databases are general face recognition databases useful for the general research on face recognition works of our group.

Publications

The works conducting to this thesis have been partially published and presented in the following conferences and journals:

- B. Raducanu and M. Graña. An Approach to Face Localization Based on Signature Analysis. *Proc. of Int'l. Workshop on Audio Visual Speech Processing, AVSP97*, pp. 109-112, Rhodes, Greece, September 1997

- B. Raducanu and M. Graña. Robust and Fast Face Localization on Image Sequences. *Proc. of EUROMEDIA98*, pp. 81-85, Leicester, United Kingdom, January 1998

- B. Raducanu, M. Graña, A. D'Anjou and F. X. Albizuri. ANN for Facial Information Processing: A Review of Recent Approaches. *Proc.of the European Symposium on Artificial Neural Networks, ESANN98*, pp. 369-375, Bruges, Belgium, April 1998

- M. Graña, A. I. Gonzalez, B. Raducanu and I. Echave. Fast Face Localization for Mobile Robots: Signature Analysis and Color Processing. *Proc. of Int'l Workshop on Intelligent Robots and Computer Vision, IRCV98*, pp. 387-398, Boston, USA, November 1998

- M. Pantic, B. Raducanu, L. J. M. Rothkrantz and M. Graña. Automatic Eyebrow Tracking Using Boundary Chain Code. *Proc. of ASCI99* , pp. 137-143, Heijen, Holland, June 1999

- B. Raducanu and M. Graña. Testing some Morphological Approaches to Face Localization. *Proc. of Int'l Symposium on Mathematical Morphology, ISMM2000*, pp. 415-424, Palo Alto, USA, June 2000

- B. Raducanu and M. Graña. Morphological Neural Networks for Robust Visual Processing in Mobile Robotics. *Proc. of Int'l. Joint Conf. on Neural Networks, IJCNN2000*, pp. 140-143, Como, Italy, July 2000
- B. Raducanu and M. Graña. Face Localization Based on the Morphological Multiscale Fingerprints. *Proc. of Int'l Conf. on Pattern Recognition, ICPR2000*, pp. 929-932, Barcelona, Spain, September 2000
- B. Raducanu and M. Graña. A Gray scale Hit-or-Miss Transform based on Level Sets. *Proc. of Int'l Conf. on Image Processing, ICIP2000*, pp- 931-933, Vancouver, Canada, September 2000
- B. Raducanu and M. Graña. Morphological Techniques for Human Face Localization. *To appear in proceedings of IMA-IP2000*, Leicester, United Kingdom, September 2000
- M. Graña and B. Raducanu. Adaptive Kernels for Morphological Heteroassociative Neural Networks. *Proc. of Electronic Imaging*, pp. 130-137, San Jose, USA, January 2001
- B. Raducanu, M. Graña, F. X. Albizuri and A. D'Anjou. Testing some Morphological Approaches to Face Localization. *Pattern Recognition Letters*, 22(3-4):359-371, 2001
- B. Raducanu and M. Graña. Visual Self-Localization with Morphological Neural Networks. *Proc. of European Symposium on Artificial Neural Networks, ESANN2001*, pp. 25-30, Bruges, Belgium, April 2001
- B. Raducanu, M. Graña and P. Sussner. Advances in Mobile Robots Self-Localization using Morphological Neural Networks. *To appear in Proc. of Int'l Workshop on Mobile Robot Technology,* Cheju Island, Korea, May 2001
- B. Raducanu, M. Graña and P. Sussner. Morphological Neural Networks for Vision-Based Self-Localization. *To appear in Proc. of Int'l Conf. on Robotics and Automation, ICRA2001*, Seoul, Korea, May 2001
- B. Raducanu and M. Graña. On the Application of Heteroassociative Morphological Memories to Face Localization. *To appear in Proc. of Int'l. Workshop on Artificial Neural Networks, IWANN2001*, Granada, Spain, June 2001

- M. Graña and B. Raducanu. Some Applications of Morphological Neural Networks. *To appear in Proc. of Int'l. Joint Conf. on Neural Networks, IJCNN2001*, Washington, USA, July 2001

- B. Raducanu and M. Graña. On the Application of Morphological Heteroassociative Neural Networks. *To appear in Proc. of Int'l. Conf. on Image Processing, ICIP2001*, Thessaloniki, Greece, October 2001

Outline and Motivation

The scope of this thesis is the experimental evaluation of face localization methods based on mathematical morphology. The expected benefit from this paradigm specialization is more robustness against linear and non-linear deformations in both spatial and intensity domains. Face recognition and face localization have been challenging computational problems since the beginning of Computer Vision and Artificial Intelligence. Recently, a great deal of effort has been devoted to them, which is reflected in the large number of recent publications that report research works applied to these problems. The introduction of techniques based on Principal Component Analysis gave a big a impulse to the face recognition research, approaching it to the industrial/commercial level. Face localization, however, remains a problem solved via heuristic and *ad hoc* methods. For this reason, we have focused our work in face localization over monochrome (gray scale) images.

The number of publications related to face recognition and other face processing problems is so large that an exhaustive review could be almost impossible. In our reviewing process, we have examined a number of publications that represent many or most of the approaches explored in the literature. We start with this review of the state of the art. A brief comment about the main face databases is also provided. Afterwards, we present three approaches we implemented and tested: gray scale Hit-or-Miss Transform, Morphological Multiscale Fingerprints and Associative Morphological Memories .

The complexities of the general formulation of face localization arise from its character as a two class classification problem, one the classes being the universal set of images minus the set of face-like images. This implies the need of sophisticated bootstrapping statistical methods for the construction of the classifier. The morphological methods tested in this thesis are robust in the sense of giving good results with naive constructions made up from casually selected face patterns.

The *Hit-or-Miss Transform* is a fundamental tool in mathematical morphology image analysis. It played a key role in its development. Besides its role as a template matching algorithm, it is the basic building block in the definition of more complex morphological operators. The traditional binary Hit-or-Miss has been extended by some authors to cope with uncertain matching and gray scale images. Here we propose a new gray scale Hit-or-Miss Transform based on the decomposition of a gray scale image on its building gray levels. The new transform is defined as the supremum of all binary Hit-or-Miss Transforms computed at each level set of both the image and the pattern. Face detection is realized by the detection of the local maxima of the gray scale Hit-or-Miss Transform, when the structural element is a face pattern.

Multiscale space analysis has been proposed in vision processes for robust matching in stereo and in early vision algorithms. Linear multiscale space is based on the convolution with Gaussian kernels. Recently, multiscale spaces based on morphological operators have been proposed. Whereas in the linear multiscale spaces border detection define the primal sketch for pattern matching, in morphological multiscale spaces, the primal sketch is given by the local extrema. In our approach we use *Morphological Multiscale Fingerprints* to characterize the pattern. Morphological Multiscale Fingerprints are plots, over scale, of the set of signal local extrema. Face detection becomes a subgraph isomorphism recognition problem, which is a task with high tradition in Computer Vision. A simple approach based on counting the number of fingerprints at each scale is enough to obtain good detection results.

Many recent approaches used in computer vision for pattern recognition are based on neural networks. Although the origins of the classical neural networks date back to the 50's, it's only recent (1990) when a new type of neural networks emerged:the *Morphological Associative Memories.* In Morphological Associative Memories the operations of multiplication and addition have been substituted with the operations of addition and maximum/minimum respectively. Due to this particularity, Morphological Associative Memories present some properties that are radical different from the classical ones'. We use the *Morphological Heteroassociative Memories* for face detection. Face detection is accomplished via the recognition performed by the response of the Morphological Heteroassociative Memories when storing a set of face patterns. To increase the robustness of the approach, it is combined with the multiscale erosion and dilation of the face patterns.

- **Chapter 1** is devoted to an introduction in a very general sense to facial information processing. The motivation from the side of the technically oriented scientific community is presented, as well as a psychological perspective.

- **Chapter 2** presents a review of the literature on face localization up to this date. Besides this, it also provides an overview of applications of visual processing of faces and facial features, like biometrics, multimodal interaction, wearable computing, etc. in which the task of face localization is of critical importance.

- **Chapter 3** presents and attempt to attack the face localization task from a first principles perspective. The chapter gives some introductory ideas on mathematical morphology and two generalizations of the Hit-or-Miss Transform to gray scale images. One proposed by Shaffer and the other is our own proposition. Experimental results of the application of these gray scale Hit or Miss Transforms are presented and discussed.

- **Chapter 4** introduces the concepts of Scale Space analysis. The Morphological Scale Space based on multiscale erosions and dilations is presented. Fingerprints in this Scale Space are introduced. The ideas of object recognition based on graph matching are reviewed. Object recognition based on matching the graphs defined by the morphological fingerprints is presented. Experimental results on the face localization task based on the fingerprint graph matching approach are presented and discussed.

- **Chapter 5** is devoted to the use of Morphological Associative Memories for the same problem. We introduce the relevant definitions and latter we present and approach to construct Morphological Heteroassociative Memories based on the Morphological Scale Spaces induced by the multiscale erosion and dilations. We present some experimental results based on this approach.

- **Chapter 6** presents our conclusions and lines of research that may be considered as open research tracks.

- **Appendix A** gives a short abstract presentation of Scale-Spaces that unifies linear and morphological scale-spaces, extracted from the works of Heijmanns.

Chapter 1

Artificial Intelligence, Computer Vision and Face Image Processing

The definition of Artificial Intelligence (AI) is the source of a lot of debates among the members of this community. Nowadays, there are many definitions of AI (personal views) widely accepted.. From the multitude of definitions, we choose the one offered by John McCarthy (see [93]) in an interview published in Internet: AI is *the science and engineering of making intelligent machines, especially intelligent computer programs.* For this reason, it deals with the study of the computational aspects of human intelligence, i.e. how can it be modelled into a computer program. Of course, AI doesn't claim a complete reproduction of the human mind in an artificial brain. It only simulates parts of it and the ultimate goal of the scientists is to bring it as close as possible to the human level. An exact reproduction of the human perception and information processing mechanisms is impossible. For instance, there is a very complex interplay between the methabolic processes and the functionality of the nervous system [107], that can not be modelled by an electronic device. However, there exist artificial intelligent systems able to learn, to generalize the learned knowledge and to adapt themselves (the so-called *autonomous systems*).

The more recent trends in AI have meet a frontier, whose roots can be found long ago in philosophy, that states the interdependence between the mind and the body. This is known as the *mind/body problem*. In other words, the human brain needs to extract its processing information from the real world. It heavily relies on the body, this collection of senses and activators, that continuously interrupt the mental processes with a huge amount of information that must be processed *instantaneously*. This interdependence is very plastically described by Kelly in [68], when it says that: "The body is the anchor of the mind [...]. A mind cannot possibly consider anything beyond what it can measure or calculate; without a body it can only consider itself." With these statements, it is obvious that an artificial intelligent system must be made the same way, by embodying an artificial brain into a physical entity. A system which is completely isolated from the real-world is senseless. Therefore, AI systems need to be endowed with artificial perception systems and these new frontiers has made apparent the complexities inherent to the sensory systems developed by natural evolution. It is being realized that the simple task of sensing the world is computationally more demanding than the simulation of most abstract thought processes, like chess-playing or theorem proving, already developed to a great extent in the early times of AI.

For a better understanding of how AI aroused, let's take first an insight at the word *intelligence*. In our daily life, we are used to give a single, common-sense acceptance, to the term intelligence. Nevertheless, it is a much broader concept. In [48], the psychologist Howard Gardner demonstrates the existence of up to eight different types of intelligence, instead of the two, logical and linguistically, that were considered previously. Initially, he identified seven types of intelligence: *visual/spatial, musical, linguistic, logical/mathematical, interpersonal, intrapersonal and bodily/kinesthetic*. Recently, he added the *naturalist* to expand his model to eight different forms of intelligence. One very important observation is that the *interpersonal* and *intrapersonal* intelligence form the so-called *emotional* intelligence, i.e. our ability to handle emotions. A very good reference of how this type of intelligence can be understood and developed is [52]. In this reference, based on brain and behavioral research, Goleman argues that our vision about human success based on the IQ factor is too narrow. Instead, he claims that "emotional intelligence" is the most responsible for human success. People who possess high emotional intelligence are the people who truly succeed in work and in building flourishing careers, as well as lasting relationships.

Let us examine briefly how each of the forms of intelligence enumerated above is being realized in AI systems. In the dawns of AI, the focus was on developing systems with *logi-*

cal/mathematical intelligence, by performing symbolic or connectionist computations. Further developments in the fields of signal processing, acoustics and music theory allowed *musical* intelligence to be developed in AI systems through audio signal processing techniques. The *linguistic* intelligence is nowadays present in the form of systems able of natural language processing. Much research is addressed to the understanding of the emotional expression to add flexibility and adaptability of the AI systems to the human emotional states, among them the facial expression of emotions. Also the synthesis of emotional expressions is sought for improved interaction. And finally, the *visual/spatial* intelligence is implemented through artificial sensing systems, among them vision systems, whose tasks are to analyze, recognize and build semantic relations between image objects in a 3D scene, but are also able to generate synthetic, realistic 3D environments.

If we can talk about an IQ test to measure the level of human intelligence, what can we say about the Artificial Intelligence? How can we evaluate it? In which situation can we decide that a system possesses Artificial Intelligence or not? The oldest test proposed to answer these questions is due to Alan Turing, appeared in [128] in 1950, and is known as the *Turing test for intelligence*. Briefly, the Turing test consists in the following: an interrogator is connected to one person and one machine via a terminal, therefore can't see his counterparts. His task is to find out which one of the two candidates is the machine, and which one is the human only by asking them questions. If the machine can *fool* the interrogator into believing it to be human, it is declared intelligent. This test has been subject to different kinds of criticism and has been in the center of many discussions in AI, philosophy and cognitive science. The Turing test is a one-sided test. A machine that passes the test should certainly be considered intelligent, but a machine could still be considered intelligent without knowing enough about humans to imitate a human. In [32], Dennett presents an excellent discussion of the Turing test and the various partial Turing tests that have been implemented, i.e. with restrictions on the observer's knowledge of AI and the subject matter of questioning. It turns out that some people are easily led into believing that a rather dumb program is intelligent.

Basically, an AI systems may be viewed as being formed by three parts: perception, cognition and action. Perception is the part of the system in charge of the capture of the incoming information from the real world (throughout sensors, classical or advanced input devices, video camera, etc.) and to translate it into symbols and numbers. Cognition deals with the interpretation of the data throughout symbolic and numerical paradigms. The symbols are clustered and matched against some predefined or learned templates that model the system knowledge. The action is about the response of the system as a result of the input received.

In humans, the visual system is by far the most complex and important perceiving mechanism. It's relevance is again very plastically evoked by Kelly in [68]: "The eye is most important because being half brain itself it floods the mind with an impossibly rich feed of half-digested data, critical decisions, hints for future steps, clues of hidden things, evocative moments and beauty. The mind grinds under the load, and behaves. Cut loose from its eyes suddenly, the mind will rear up, spin, retreat". It seems needless to further emphasize the importance of the vision system and, at the same time, becomes obvious the interest and the struggle of researchers from AI to enable the machines *to see* things. Their work gave rise to the Computer and Machine Vision fields. This way, AI embarked on a new realm of unlimited possibilities and challenging promises. Although in many works, computer vision and machine vision are taken with the same meaning, there is though an important difference between these two concepts. *Computer Vision* is mainly concerned with the study of the scientific basis of human and animal vision and development of algorithms for image processing. On the other hand, *Machine Vision* deals with Systems Engineering and the solution of practical problems, such as guiding robots, automatic product inspection and process monitoring. However, from now on, we will refer at both concepts in an interchangeable manner.

In the design of artificial vision systems, the researchers have taken inspiration from the biological vision systems. The biological vision can be understood as consisting of two levels. Two levels have been identified in biological vision systems. Low level vision processes are characterized by the concepts of *fixation* and *attention.* Fixation implies a mechanical movement of the eye, whereas attention is the focus on the current region of interest. High level vision processes deal with object representations for cognition and interpretation.

At the low (physical) level, the model proposed was aimed to recreate an active vision mechanism. It is known that the density of the photoreceptors in the retina is greater in the central part (fovea), while it is coarser at the periphery. As a result, the resolution of the captured image decreases with the distance from the center. Thus, only the image captured by the fovea (the center of retina) is used for the recognition. The image formed at the periphery of the retina is used as a hint for the next perception process and it acts like a control mechanism for the eye movement.

At the higher (cognitive) level, an artificial vision system should be able to recognize *familiar* objects. Thus, the challenging problem is: how we can represent the objects for recognition? There are some proposals to represent the object in term of edges or junctions [21]. Other approaches [22] make use of the 3D information. At this point, there are two possible representations: one is fully 3D viewpoint invariant (i.e. requires that a similar 3D representation of the input be recovered from the image before it is matched to like-representations in memory), whilst the other one is viewpoint dependent (i.e. requires that 3D representations of the input be normalized by an appropriate spatial transformation from the viewpoint of the image to the viewpoint of a view-specific representation in memory). Besides these approaches, there is the so-called behaviorist paradigm, that claims that an object becomes known to the system when it is able to predict the object behavior to its actions. This is exploited in [116]. According to this paradigm, the internal representation of the object contains chains of alternating traces in eye movement and perceptual information. Each of these chains reflects an alternating sequence of motor actions and sensory signals that are expected to arrive in response to each action (verification of the image fragments). This behavior forms the basis for object recognition.

Summarizing, we can say that the during visual perception and recognition, the eyes move and fixate on the most informative parts of the scene which, afterwards, is processed with the highest possible resolution. In other words, there is a strong connection between the changes in sensory information and cognition processes. That's why, a possible strategy to build a complete artificial vision system (that integrate both physical and cognitive levels) would consists of the following stages: find some *structure* to attend to, fixate on it, find the next attention point and decide if a refixing is needed, decide if two features should be connected, relate the structure from one fixation to another.

Recently, there are some critical voices about the definition of Computer Vision (CV) as part of AI. These claims were sustained by the evidence that general AI conferences show a decline in both the number and the quality of vision papers, whereas there is a great growth and specialization of CV conferences. A closer look at many CV conference and journal papers reveals that there is a clear inclination towards more complex mathematics (PDE models, functional analysis, geometric invariance), physics (color and illumination properties, motion analysis), statistics and general non-symbolic image processing techniques. In their majority, they even don't mention the name of AI nor it's methods. On the other hand, AI is more concerned with models of logical and non-logical reasoning, heuristic and uncertain reasoning, knowledge representation and learning. Despite of these apparent *divergences*, we do consider that CV is still part of AI. We agree with the argumentation made in [46]. We will only review only the main aspects.

First, AI and CV share methodological approaches. The main strategies for problem solving in AI are: pattern classification, neural-networks, symbolic computation and behaviorist methods. CV also manifests some of these strategies. And furthermore, it developed its own methodologies. The most ones are: digital image processing (where the algorithms are based on the geometrical or physical structure of the image) and active vision (the ensemble of optical and mechanical techniques that allow a system to develop perceptual processes).

Second, both AI and CV share domain assumptions. Knowledge representation is basically an AI domain. On its turn, CV makes also use of knowledge representation: shape and apparent structure of objects. And it also makes use of geometrical models for object representation and rules for objects recognition. Reasoning processes in AI are often based on heuristics, due to the lack of information or its inaccuracy. Despite recent advances in this field, CV still makes use, on a large scale, of heuristics for object recognition. Many times, these heuristics are based on a common-sense information of the object to be recognized. For instance, many object recognition algorithms based on edge detectors use heuristics, assuming that the output of the algorithm are the object' edges and discarding light, shadows and reflectance. In AI no single theory can explain all behaviors. In CV also, many types of theories are being developed. Low-level image processing is more data driven (at pixel level), whilst high-level vision deals with more symbolic computation (object recognition, 3D scene understanding).

Third, AI and CV share common goals. Throughout the goals of AI we can mention: characterize intelligence and intelligent behavior, understand human abilities and computational processes, develop tools for the advancement of computer science and engineering, extend the philosophical view of mind/body problem to artificial brain in physical entities. The goals of CV are: understand and reproduce the human or other biological vision systems, provide machines that extend human perceiving abilities and identify and extract key information for spatial description of real-world environments.

And finally, AI and CV share a common intellectual context throughout three elements. One of them is the philosophical support. Both AI and CV give strong emphasis to the study of relationship between a physical reality and our perceived understanding of the world. Another element is represented by biological and psychological support. The vision system is not only a transducer from light into nervous stimuli.. Neurophysiologists showed that in visual cortex we develop different regions that appear to extract shape, motion and color from the perceived image. Furthermore, the output of visual processing is used to maintain balance, track moving objects, etc. The point is there are many biological processes that contribute to create our intelligent behavior. The third and last element is about computational methodology. A non-exhaustive list of common tools and techniques used by both AI and CV would include: probabilistic and uncertain reasoning, symbolic and neural network techniques, hierarchical representations and reasoning methods, search tree exploration, etc.

Having seen the embedment of CV in AI, a general overview about CV will not be complete until a brief evocation of two decisive moments that officially marked the beginning of this field, as reported in [59].

In 1961, Larry Roberts created a program that was intended to *see* a block-based structure, analyze its components and build a scene map. The strategy of the program was to find the changes in gray scale, which were supposed to be the objects' boundaries. Afterwards, the algorithm was able to build the lines throughout detected points. At the next level, the corners, the facets were identified and their combination was used to decide which of them belong to the same block.

The second attempt in the CV history, to create an artificial vision system, is attributed to Marvin Minsky. In the summer of 1966, he asked one of its students, Gerald Sussman, to solve one of the most challenging problems in AI. The student had to connect a TV camera to a computer and to write a program in order to *describe* the scene view recorded with the camera. At that time, Minsky was unaware of the difficulty and complexity of this task. Even if this project proved to be finally unsuccessful, it served to show the general underestimation of the difficulties involved in simulating the sensorial system.

We stop here with the early years of CV and AI and we focus now on another challenging problem. In the late 70s, the interest of researchers in CV, AI and other fields was drawn toward the study of human faces. This is a very complex theme, that was the starting point of a whole new chapter in CV field and, in the last years, become a quasi-omnipresent topic in the most prestigious conferences and journals. This study involves several issues: face localization, face recognition, facial feature extraction and facial expression recognition.

But why are the faces so important? They represent the main cue we use for person identification. In [45], the authors argue that the new-born children come to the world pre-wired to be attracted by faces. It seems that, in general, they prefer to look at moving stimuli that resemble face-like patterns. In the same reference, it is showed that the humans are able to remember faces easier than other objects when presented in an upright orientation.

Faces are also used as the main gateway to express our feelings, because facial expressions act like a mirror of our internal emotional states. Thus, they are amongst the most important *means* of interaction between humans. For this reason, the human visual system developed an specialized ability to recognize and interpret facial images. This ability is even more surprising, taking into account the fact that despite of the uniformity of face patterns, we are able to distinguish among a huge variety of faces. With only one snapshot of a human face, we can recognize that person later on under different head poses and facial expressions, having occluded parts of the face due to beard or glasses, independent to illumination conditions and even after a long period of time (months, years). By contrast, nowadays, the automatic face recognition systems need to be trained with hundreds (or thousands) of images, under different head poses, facial expressions, scales and illumination conditions. From this evidence, it becomes obvious that what is a very easy task for humans, is still a big challenge for machine vision. That's why, to build a robust computer vision system able to perform accurately the face recognition task, we should better understand how we perform the generalization from single views. Automatic face recognition systems may profit from the work that psychological researchers have carried out in order to identify the factors that make for us the face recognition task a *quasi-instantaneous* and *natural* process. In [127] for instance, the authors address the problem of face recognition under different poses, investigating the role of texture and shape in human subjects.

In clinical neuropsychology, the interest of how humans are able to recognize faces is justified by the existence of a disability called *prosopagnosia* or *face-blindness*: the loss of the ability to recognize familiar faces, while the ability to recognize objects remains intact (see [111]). Due to some damages in the certain regions of the brain, the link between the memorized faces and the names associated with them is broken. This observation implies that names and faces are two distinctly separate entities, and propopagnosia is not synonym to the difficulty of remembering names. The existence of prosopagnosia suggests that the face recognition is a very complex task that develops in different areas of the brain, involving more specialized subprocesses or modules (if this pathway is genetically predetermined or emerges under the social activity, is still unclear, as stated in [87]). Bruce and Young [17], propose a model of face recognition based on the interaction between different modules within face processing pathway. Processing of facial identity is the responsibility of face recognition units (FRUs), which are nodes that respond to structural description of known faces. On the other hand, the *generic* information regarding a certain person (age, occupation, etc.) is stored in the so-called Personal Identity Nodes (PINs). And finally, the names associated with that person, are stored separately, in the name generation module (NGM). The persons affected by prosopagnosia, developed other cues (gender discrimination, age range, voice recognition) to counteract the drawback of visual identification. Several attempts have been reported to reproduce these alternative cues of person identification in computer systems like the sonar based proposed in [36] for person identification. Other researches were aimed towards an integration between audio (voice) and visual scores for more accurate person identification results. In this case, the audio channel is used to reinforce the visual one. Some results of these researches are reported in [18] and [19].

From the point of pattern recognition, the faces can be considered as geometric templates with a rough feature configuration, arranged in an identical spatial relationship, but showing slight deformations. For this reason, the results obtained in the study of face recognition can be used for the study of other classes of objects that share similar properties. Perhaps the oldest reference related with human face recognition by computers is [67]. Face recognition has been since then a test field of many diverse approaches to knowledge representation, pattern recognition and image processing techniques. Approaches tested and tried encompass linear and non-linear, dense and sparse, feature extraction and all kinds of classifier systems (see for instance the collection of works in [136]). . Some relative success under some circumstances (good illumination, up-right frontal view) have pushed face recognition from the realm of science into that of industrial solutions.

To finish this introduction, we remark the strong relationship between the natural intelligence and the artificial intelligence and how the development of the latter heavily relies upon the understanding of the first. On a broader view, the advance of AI is the history of the interconnection between the biology (the world of the *born)* and mechanics (the world of the *made*). With time passing, and the progress in various scientific domains, this interconnection will become stronger and stronger. It is very poetically expressed again by Kelly in [68]: "When the union of the born and the made is complete, our fabrications will learn, adapt, heal themselves, and evolve. ... The world of the made will soon be like the world of the born: autonomous, adaptable and creative, but, consequently, out of our control".

Chapter 2

Face Localization: State of the Art

Face recognition has been a landmark problem in AI for so long that an exhaustive review is almost impossible. Face localization and detection, although assumed as a minor problem by most researchers, is always part of the face recognition systems and near the same variety of approaches have been tried and tested for face localization as for face recognition. In this chapter we give a review as comprehensive as possible of the approaches found in the literature. We start by defining the face localization and detection problem, detaching it from the face recognition problem, and giving it recognition of its own difficulties. Then we review techniques and applications of face localization *per se*.

2.1 Statement of the Problem

Researchers from very different areas (computer science, AI , psychology) are devoted to the study of face related issues. We enclose these studies under the name of facial information processing, which includes, but it is not limited to, the following topics: face detection and localization, face recognition, facial expression recognition/synthesis, facial feature extraction, etc. At first sight, these topics can be considered as independent. There are evidences published in the psychology literature that suggest that specialized neural systems are developed for diverse facial information processes. For instance, in [125], it is suggested that face recognition and recognition of facial expressions are done in parallel.

The diversity of applications and computational and sensor configurations give rise to many combinations of the basic facial information processes. For example face detection and facial feature extraction can be combined in two basic ways. The process can first detect the face using a holistic technique (PCA, Neural Networks) and afterwards it can start the localization of facial features (following a *top-down* approach [82], [117] and [134]). On the other hand, we can first detect features like mouth, nose, nostrils, eyes and then, checking the spatial relationship between them, try to decide the presence of a face or not (following a *bottom-up* approach [147], [65] and [143]). These mixed approaches can be considered as result of the psychological researches that suggest [24] that both holistic and feature information are crucial for perception and recognition of faces. Another fact is that not all facial features (hair, face shape, mouth, eyes, nose, chin) have the same significance. For the problem of face recognition from frontal views, the nose play an irrelevant role. Nevertheless, it is very important for face recognition from profile views [122] and [148]. Furthermore, it has been established that the upper part of the face is more useful for recognition the lower part.

From the point of view of statistical pattern recognition [47] and [56], both face localization and face recognition are classification problems. In the case of face localization, we have to decide if a certain region in the image contains a face or not. In other words, we have to decide if a region belongs to the face class or to the non-face class, thus it is a two-class classification problem. On the other hand, face recognition can be stated as matching a face against a database consisting of hundreds or thousands of face images, thus it is a N-class classification problem. Let us give a more precise set of definitions:

- Face detection: it is the problem of deciding if there is a face or a set of faces in an image.
- Face localization: it is the problem of finding the image coordinates of the face.
- Face verification: it is the problem of a user identity authentication based on the subimages corresponding to faces.
- Face recognition: it is the problem of discovering the user identity based on the face image, it involves the search over a whole database of faces.

Even if the main focus of this work is to realize an overview of face detection and localization techniques and applications, we will often relate this problem to face recognition and facial feature extraction. As we will see throughout this chapter, these three problems are in a very strong interdependence and an isolated analysis is senseless. At the same time, as the literature on these topics has grown exponentially in the last years, we don't claim an exhaustive review of all existing works, but rather a state of the art of the most relevant researches.

Although for humans, face detection seems a *natural* task even in cluttered scenes, for machine vision applications it is still a very challenging problem. Face detection asks for robustness against changes in illumination conditions, changes in scale, rotation and translation transformations, deformations due to pose variations and emotional expressions and occlusions due to beards, glasses, etc.

The face detection and localization problem depends largely on the sensor configuration, for example it is relatively trivial in infrared images where the contrast between the human being and the background is strong. In color images, the face color can be used to produce fast initial guesses of face position, and also in video sequences it is possible to perform very simple motion analysis to segment the human shape [141] and [139].

In the literature, we have found mainly two paradigms to address the problem of face localization in still gray-level images. The first one is a holistic process of the image. Within this paradigm, we can mention techniques like: Principal Component Analysis (PCA), Artificial Neural Networks, color or shape based segmentation and probabilistic approaches. The other paradigm is based on feature extraction and grouping, checking some geometric constraints imposed to a parameterized face model or through an energy function minimization (as we will see in the case of graph-matching based techniques).

Following a statistical pattern recognition formulation the face detection problem can be stated as follows: Let x be a feature vector extracted from the image (or a subimage block). The values of this feature are characterized by probability distribution $p(x)$ which can be expressed in terms of the face F and non face $\overline{F}$ classes conditional distribution:

$$p(x) = p(x\,|F)\,p(F) + p\left(x\,|\overline{F}\right)p\left(\overline{F}\right), \tag{2.1}$$

the detection problem is therefore a decision problem that can be solved using the Bayesian Maximum A Posteriori (MAP) rule: choose w such that $w = \arg\max_{w=\left\{F,\overline{F}\right\}} \{p(w\,|x)\}$. The A Posteriori probabilities of the face and non face classes are of the form:

$$\begin{aligned} p\left(F\left|x\right.\right) &= \frac{p\left(x\left|F\right.\right)p\left(F\right)}{p\left(x\right)}, \\ p\left(\overline{F}\left|x\right.\right) &= \frac{p\left(x\left|\overline{F}\right.\right)p\left(\overline{F}\right)}{p\left(x\right)}. \end{aligned} \tag{2.2}$$

The MAP decision can be expressed in terms of the following likelihood ratio:

$$L\left(x\right) = \frac{p\left(x\left|F\right.\right)}{p\left(x\left|\overline{F}\right.\right)}, \tag{2.3}$$

once evaluated, $L\left(x\right) > 1$ means that the subimage contains a face and $L\left(x\right) < 1$ means that it does not contain a face. The solution of the face detection problem involves the construction of models for both distributions and their parameter estimation from a set of face instances. A non trivial difficulty in face detection is the estimation of the non-face conditional distribution, that involves the modelling of the almost universal set of images (all the images minus the face images). The diverse approaches deal with this problem in different ways. Some approaches give a distance value $d\left(x\right)$ measuring the similarity of the input features to a kind of "face space", so that the face detection decision is translated into the determination of a threshold on this distance.

Multiple face detection (and translation invariant detection) is accomplished convolving the face detection algorithm over the image. Recognition over a set of scales involves the processing of the multiscale pyramid generated by sampling the image at several resolutions, a costly process. Face detection algorithms, that show some robustness against scale variations reduce this computational load. Rotation invariance is sometimes achieved by the detection of the potential angle with a specialized subsystem and recovering the upright pose before application of the face detection algorithm. In other works, both rotation and pose invariant detection is achieved through the separate detection as isolated classification problems of potential rotations and deformations due to pose changes.

2.2 Technical approaches to face detection

In this section we will review the main technical approaches to face localization. The literature on this issue is large and is continuously growing, as can be appreciated by the recent dates of some of the references. Most of the early literature did consider the problem of face localization as a secondary one, a step previous to the main task of face recognition. Therefore, most of the early references assume that a simplified version of the face recognition approach would suffice for the face localization task. However, the practical application of face recognition and other face processing applications has produced a growing realization of the difficulties involved in robust face localization.

2.2.1 Principal Component Analysis

Principal Component Analysis (PCA) is a well-known technique [47] used for dimensionality reduction of data-spaces. Images are considered as vectors in a very high dimension space, and PCA is an optimal linear transformation into a lower dimension space. PCA (also known as Karhunen-Loève Transform) was first introduced in [70] as a face representation tool. It was applied in [129] for face recognition, but the authors also showed how it can be used for face localization and tracking in a video sequence.

We consider an initial set of M face images, each one of size $N \times N$ pixels. Let them be Γ_1, Γ_2,.., Γ_M. The average face of the set is denoted by Ψ. Each face in the set differs from the average by the quantity $\Phi_i = \Gamma_i - \Psi$, with $i = 1, .., M$. We consider the covariance matrix of the original set of mean centered face images: $C = \frac{1}{M-1} \sum_{i=1}^{M} \Phi_i \Phi_i^T$, where C is a matrix of size $(N \times N) \times (N \times N)$. Usually, the available data is scarce, $M << N^4$. Therefore, only the M most significant eigenvectors may be computed from C. We denote them $\{u_l; l = 1, .., M\}$. These eigenvectors are referred as eigenfaces when visualized as images.

To test the *faceness* of an image Γ (i.e. how it resembles a face or not), we project it on the subspace defined by the eigenfaces: $\omega_k = u_k^T (\Gamma - \Psi)$, $k = 1, .., M$. The reconstruction from this projection is given by $\Phi_f = \sum_{i=1}^{M} \omega_i u_i$. The reconstruction error $\varepsilon^2 = \|\Phi - \Phi_f\|^2$, where $\Phi = \Gamma - \Psi$, can be used as a *faceness criterion* because it gives the distance to the face subspace spawned by the eigenfaces. In order to decide if at the current location in the image there is a face present or not, a threshold is applied to the distance ε. This threshold must be set experimentally. In order to set it optimally a ROC analysis on the relation between false and true detections depending on the threshold can be performed.

In this approach does not exist an explicit characterization of the conditional distribution of the non-face class. Non-face space is the complementary subspace, orthogonal to the face space. It has been extended in [95] to a multiscale search for scale invariant detection using a single set of eigentemplates (at a fixed scale) and linearly remapping the input image through a given range of scales and computing a separate distance map at each scale. The estimate of the position and scale is obtained by identifying the best global minimum among all scale-indexed distance maps.

A mixed approach that employs morphological image segmentation and PCA for face localization is presented in [131]. From the initial watershed segmentation of the color image, a region adjacency graph (RAG) is created. Regions are merged following the tree structure of the RAG, to obtain more coarse segmentation. The PCA based face detection is applied at the nodes of the RAG that represent aggregate regions that can be faces.

Local Feature Analysis (LFA), introduced in [104], is derived from the eigenface method but overcomes some of its problems by not being sensitive to deformations in the face and changes in poses and lighting. LFA considers individual features instead of relying on only a global representation of the face. The system selects a series of blocks that best define an individual face. These features are the building blocks from which all facial images can be constructed. Applying LFA, the system selects the subset of building blocks, or features, in each face that differ most from other faces. Any given face can be identified with as few as 32 to 50 of those blocks. The most characteristic points are the nose, eyebrows, mouth and the areas where the curvature of the bones changes. The patterns have to be elastic to describe possible movements or changes of expression. The computer knows that those points can move slightly across the face in combination with the others without losing the basic structure that defines that face.

The formal description of LFA can be summarized in the following way. Let us consider a set of face images (represented in a vector form) $\left\{\phi^t\left(x\right), t = 1..T\right\}$ where by T we denote the number of images. Applying the PCA scheme over this set, we obtain the orthonormal set of eigenvectors ψ_r, $r = 1..T$ and their respective eigenvalues λ_r sorted in decreasing order of their magnitude. Now, we can construct the following two functions:

$$K(x,y) = \sum\nolimits_{r=1}^{T} \psi_r(x) \frac{1}{\sqrt{\lambda_r}} \psi_r(y) \tag{2.4}$$

$$P(x,y) = \sum\nolimits_{r=1}^{T} \psi_r(x) \psi_r(y) \tag{2.5}$$

where $K(x,y)$ is the kernel of the representation and $P(x,y)$ is the residual correlation of the outputs. The output of the LFA is given by:

$$O(x) = \int \sum\nolimits_{r=1}^{T} \psi_r(x) \frac{1}{\sqrt{\lambda_r}} \psi_r(y) \phi(y) \tag{2.6}$$

To reconstruct the original image from the output $\{O(x)\}$ we have the following relation:

$$\phi^{rec}(x) = \sum\nolimits_{r=1}^{T} \int \sqrt{\lambda_r} \psi_r(y) O(y) \psi_r(x) = \int K^{-1}(x,y) O(y) \tag{2.7}$$

where the inverse kernel is given by:

$$K^{-1}(x,y) = \sum\nolimits_{r=1}^{T} \psi_r(x) \sqrt{\lambda_r} \psi_r(y) \tag{2.8}$$

Exactly as in the case of PCA, the reconstruction error is $\varepsilon^2 = \|\varphi - \varphi^{rec}\|^2$. This means that the topographic LFA representation has the same best reconstruction, generalization and object constancy properties as the global non-topographic PCA one.

As a final comment on PCA based detection, it has been of relative practical utility because the linear transformation can be realized in real-time for small images, given that a hint on the face localization has been obtained by other means (color processing, motion). However PCA are very sensitive to rotations and deformations due to pose changes.

2.2.2 Neural Networks

Among the most influential works which address the problem of face localization is the application of artificial neural networks in [114]. Their database has become a benchmark for face detection algorithms. They developed a system able to detect upright frontal faces with slight rotations (up to 10^o both clockwise and counterclockwise). The system has two phases: a Neural Network filter and a Merging and Arbitration scheme (that also may be partially implemented by a neural network).

The input image is explored extracting overlapping 20×20 pixel image blocks, which are fed into a feedforward neural network, a Multilayer Perceptron (MLP), after a preprocessing step consisting in the equalization of pixel intensities. In order to obtain scale invariant detection, the input image is iteratively down sampled.. A multiresolution pyramid is obtained and the MLP is applied at each resolution level. The hidden layer of the MLP is formed by three types of units whose receptive fields are designed so that they can be trained to look for particular features in the window: mouth, pair of eyes, nose, corners of the mouth. The output of the network is a single-valued unit, which indicates if the input window contains a face or not.

The neural network is intended to model both the face and non-face conditional distributions, and to perform the MAP decision described before. The difficulty here consists of choosing the non-face patterns, so that MLP learning can build up a model of the conditional distribution of the non-face class. This problem is solved by a *boostrapping* strategy. The neural network is initially trained upon a set of face images, extracted from a face database, and 1000 random nonface images. The system is applied to a set of images that do not contain faces. False detections are included in the training set as nonface instances.

Merging reduces the number of false positive detection by merging close detections and eliminating partially overlapping detection windows. Arbitration consists in the combination of the detection performed by different neural networks, obtained after different training processes, when applied to the same image.

In [115], they extended the system in order to deal also with large rotations. Previous to the face detector module, they feed the input image into a *router network*, a neural network whose output gives the orientation (in degrees) of the face. In the case that the input image doesn't contain a face, then a meaningless result is generated. Once the orientation of the face has been established, the image is *derotated* and afterwards presented as input for the face detector module.

Another work that combine both the neural networks and PCA is reported in [124]. In this paper, they built, using 19×19 pixel patterns, six face-like and six non-face-like clusters characterized by their centroids (also called *prototypes*) and their covariance matrices. Dimension reduction is performed independently for each cluster to a 75 dimensions sub-space defined by the most significative eigenvectors of its covariance matrix. For a given test pattern, two distances are considered relative to each cluster. The first is the normalized Mahalanobis distance between the PCA projection of the test pattern and the cluster center, inside the 75 dimension principal subspace associated to the cluster. The second is the Euclidean distance between the test pattern and its reconstruction after projection in the principal subspace of the cluster. The classifier is a MLP whose input is given by the 12 pairs of distances to the face and non-face clusters. The role of the hidden layer is to combine the two-valued distance pair into a single distance value. The output of the network is formed by a single unit, whose value is 1 for an input vector corresponding to a face pattern and 0 otherwise. To train the neural network, a *bootstrapping* strategy, like the one reported in [114], was used.

Other neural network approach is reported in [83]. They introduce a probabilistic neural network for face recognition. The system has a hierarchical structure, consisting of three modules: face detector, eye localizer and face recognition. These three modules are chained in this precise order, in the sense that the output of the current module is the input for the next one. We will comment only on the face detector module. This module consists of a so-called Probabilistic Decision-Based Neural Network (PDBNN). The PDBNN learning rules have two properties. First, they are *decision-based*, this means that the teacher only tells the correctness of the result and not the exact desired result. Second, the learning rules are a mixture of *locally unsupervised* (LU) and *globally supervised* (GS) learning. The LU phase can adopt one of the unsupervised learning schemes: vector quantization (VQ), k-mean, expectation maximization (EM), etc. On the other hand, in the GS phase (after the LU phase convergence), learning follows the decision-based learning rule. When a training pattern is presented, the reinforced (2.9) or antireinforced learning (2.10) rules are applied depending on the response of the network:

$$w^{j+1} = w^{j} + \eta \nabla \phi \left(x, \omega \right) \tag{2.9}$$

$$w^{j+1} = w^{j} - \eta \nabla \phi \left(x, \omega \right) \tag{2.10}$$

where by w we denote the weight matrix, η is the learning rate parameter and $\phi\ (x, \omega)$ is the discriminant function that depends on the training pattern x and the face class ω. As for the discriminant function ϕ, it models the log-likelihood function:

$$\phi(x,\omega) = \log p(x,\omega)$$

where by $p(x,\omega)$ we denote the face class likelihood function. The authors of [83] use an Elliptic Basis Function (EBF) to model the class likelihood.

In order to ensure sufficient diversity of real face images, artificial training patterns have been generated. Up to 200 artificial patterns were created from each original pattern by applying several affine transformations (rotation, scaling, shifting, etc.). But not all the artificial patterns should be considered as valid face patterns. Some are slightly perturbed patterns can still be considered as belonging to the *positive* training patterns, but if the perturbation exceeds a certain threshold, then the pattern in cause is included in the *negative* training pattern set. The system has been trained with the set of artificial training patterns generated from 92 images and tested upon 473. The images are of size 320x240 pixels and the face is about 140x100 pixels. The orientation of the head varies up to 15^o in any of the four directions (up, down, left, right). All the images were taken in indoor scenes with cluttered background, under normal lightning conditions.

Finally, in [65], they propose a hybrid system for face localization in complex images with more than one person per image. In the first stage, a high-pass filter is used for edge detection. In the second stage, a hierarchical neural network is used. More precisely, it is a collection of four fully connected back-propagation neural networks. Three of them are devoted to detect the eyes, the nose and the mouth (in this phase, the input edge enhanced image is convolved with each of the three neural networks and new images are formed). The fourth network was trained to recognize the presence of a face if the outputs of the three previous networks indicate the presence of the corresponding features (in this phase, the three images obtained in the previous step are convolved with the fourth neural network that decides if a face is present or not).

Due to the basic nature of the problem as a two-class problem, modelling the non-face class via learning algorithms in the neural networks paradigm imposes the need of rather extensive pattern sets, the generation of artificial patterns to simulate deformations and bootstrapping methods. In the final analysis, there is little guarantee of the generalization of the detection outside the training data.

2.2.3 Deformable Template Matching

Deformable template matching techniques are based on graph matching approaches in Computer Vision. Feature points are extracted from the image. Their spatial relations are modelled by a graph whose nodes are the feature points and the edges embody the neighboring topology between feature points. Training patterns are used to estimate the model graph parameters which usually are assumed as being Gaussian distributed. Recognition is based on the measure of the disparity between featured graphs. Graph matching algorithms are inherently invariant against rotation and scale transformation and robust against deformations. However, they are computationally expensive and the robust detection of feature points is not a trivial task.

A seminal contribution in this line of work for face recognition is the Dynamic Link Architecture (DLA) by [76]. Although this method was proposed mainly for face recognition, it has been proved that it can be also used previously for face localization. A rectangular grid is applied over the object in the image. The nodes of the grid are labeled with collections of features that describe the gray-level distribution. These features are computed by convolution of the image with a family of Gabor-type wavelets of different frequencies and orientations. Grid edges are labeled with metric information on the relative position of the vertices. The matching process is based on the minimization of a cost function that measures how much the test graph resembles the model. Formally, the matching process can be described in the following way.

The convolution of the image I with a family of kernels ψ_k at a certain location $s_0 = s_0\,(x, y)$ is given by:

$$(\psi_k \star I)\,(s_0) = \int \psi_k\,(s_0 - s)\, I\,(s)\, d^2 s \tag{2.11}$$

where the kernels ψ_k are of the form:

$$\psi_k\,(x) = \frac{k^2}{\sigma^2} \exp\left(-\frac{k^2 s^2}{2\sigma^2}\right)\left[\exp\,(iks) - \exp\left(-\sigma^2/2\right)\right] \tag{2.12}$$

The family of kernels ψ_k is self-similar under the application of the group of translations, rotation and scalings. This family is the well-known *Gabor wavelets* base. The family index k

is a function of frequency υ and orientation μ, i.e. $k = k(\nu, \mu)$ and its analytical expression is given by the formula:

$$k(\nu, \mu) = k_\nu \exp\left(i\varphi_\mu\right) \tag{2.13}$$

The magnitudes of the convolutions $\left(\psi_{k(\nu,\mu)} \star I\right)$ form a feature vector located at s_0. This vector receives the name of a *jet*:

$$J_{\nu\mu}(s_0) = \left|\left(\psi_{k(\nu,\mu)} \star I\right)(s_0)\right| \tag{2.14}$$

As part of graph matching process, the similarity between the jets of the test and the model patterns should be evaluated. In [76], the normed dot product of jets is considered:

$$D(J_I, J_T) = \frac{J_I J_T}{\|J_I J_T\|} \tag{2.15}$$

where by J_I we denote the jet corresponding to the model and by J_T the jet corresponding to the test image respectively. The matching process performs small displacements of the grid nodes placed over the test pattern trying to minimize the disparity between jets of corresponding nodes. The matching cost function balances the similarity of the jets and the distortion of the grid incurred in the minimization process. To perform face localization, the undistorted grid with only the smaller frequency jets is shifted along the test image. At any image location, a cost function is estimated and the new grid position is accepted if it reduces the cost function. The cost function for face localization is given by the relation:

$$Cost = \sum\nolimits_{i \in V} D_i(J_I, J_T) \tag{2.16}$$

where V denotes the set of graph nodes.

Works reported in [73] and [74] propose an alternative to the Gabor-type wavelets. They implement the dynamic link architecture using morphological multiscale dilation-erosion, i.e. jets are build up from erosions/dilations of the image with structuring objects of increasing scale. Besides the dimensionality of the grid node feature vectors is reduced using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA).

A different approach based on flexible appearance models is presented in [77]. The flexible models are sketches of face structure with a set of control points. The model is fitted upon the face based on the edge detection. Face recognition is performed based on a statistical discrimination of the fitted flexible models. In this case, face localization is achieved through the displacement of the flexible model upon the image until a overlapping match of the image edges and the model is obtained.

Approaches on graph-matching and deformable templates suffer from the following shortcomings. First, the robust computation of control points is a delicate task. Second, the minimization of the cost function may be very expensive computationally preventing real-time realizations. For images with several faces, there is the added difficulty of the discrimination between local minima of the matching cost function that may correspond to non-face subimages.

2.2.4 Probabilistic Models of Local Feature Spatial Relations

An special case of the structural models, some authors propose [79], [23] an algorithm for quasi-frontal face localization by coupling a set of local feature detectors with a statistical model of the mutual distances between facial features. This approach is claimed to be invariant with respect to translation, rotation and scale and that can handle partial occlusions of the face. In the initial step of the algorithm, a set of local feature detectors is applied to the image to locate the candidates for facial features. The arrangement of the facial features can be viewed as a random graph in which the nodes correspond to the features and the lengths of the graph arcs correspond to the distances between the features. Since different people have different distances between their features, the arc lengths are modeled as a random vector drawn from a joint probability distribution. Therefore, the face localization task corresponds to the problem of random graph matching.

In order to detect the candidate points for facial features, they convolve the input image with a family of Gaussian derivative filters at different scales and orientations. Considering the vector of filter responses at a particular location in the image as $R(x,y)$, then the matching score $Q_i(x,y)$ between $R(x,y)$ and model P_i of the ith facial feature is given by the normalized dot product, subject to a magnitude similarity test:

$$l_i(x,y) = \frac{|(|P_i| - |R(x,y)|)|}{|P_i|} \tag{2.17}$$

$$Q_i\left(x,y\right)=\begin{cases}\frac{P_i^T R(x,y)}{|P_i||R(x,y)|}, & if \quad l_i\left(x,y\right)<\tau_0\\ -1, & otherwise\end{cases} \tag{2.18}$$

They use for each facial feature (left eye, right eye, nostrils, the junction between the tip of the nose and the upper lip) a dedicated detector. The brute-force approach for grouping the feature candidate points seems unreliable because of the big number of positive detections. Instead, a controlled selective method is used. The process of grouping starts with selecting two features, for which $Q_i\left(x,y\right)>\tau$. After that, the remaining three features are detected based on the search restricted to position prediction ellipses of the features. The group formed by five feature points is referred to as a *constellation*. The next step is to decide if a candidate constellation can be a face or not. As a preprocess, the scale parameter λ is estimated, based on the assumption that the distances between facial features are jointly Gaussian-distributed with some mean and covariance provided that faces are normalized for scale.

Let us consider the vector X of mutual distances between constellation points. It is not yet normalized by scale, but it will be, if we divide it by $\lambda\left(X\right)$, i.e. the new normalized vector of mutual distances is given by:

$$L=\frac{X}{\lambda\left(X\right)} \tag{2.19}$$

Based on the above assumption, L is approximately Gaussian-distributed. This means, that L has a mean and a covariance matrix. The density of L is given by:

$$f_L\left(l\right)=N(l;\overline{L},\Sigma)=\frac{1}{\left(2\pi\right)^{\frac{n}{2}}\left|\det\left(\Sigma\right)\right|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\left(1-\overline{L}\right)^T\Sigma^{-1}\left(1-\overline{L}\right)\right) \tag{2.20}$$

where is n is the number of components in L. For N features, n is $N\left(N-1\right)/2$. The statistics $\overline{L}$ and Σ can be estimated from the training data. A *maximum likelihood* scale estimator is built as follows. Conditioned upon λ, the density of X is given by:

$$f_X\left(x|\lambda\right)=N(l;\lambda\overline{L},\lambda^2\Sigma)=\frac{1}{\left(2\pi\right)^{\frac{n}{2}}\left|\det\left(\Sigma\right)\right|^{\frac{1}{2}}\lambda^n}\exp\left(-\frac{1}{2}\left(\frac{x}{\lambda}-\overline{L}\right)^T\Sigma^{-1}\left(\frac{x}{\lambda}-\overline{L}\right)\right) \tag{2.21}$$

The maximum likelihood for the scale is the value of λ, which is the solution of the equation:

$\frac{\partial}{\partial\lambda} \log f_X(x|\lambda) = 0$, i.e.:

$$\lambda = \left(\frac{x^T\Sigma^{-1}x}{x^T\Sigma^{-1}\mu}\right)\frac{2}{1+\sqrt{1+\frac{4n(x^T\Sigma^{-1}x)}{(x^T\Sigma^{-1}\mu)^2}}} \tag{2.22}$$

The final decision step is to select between two potential constellations, the one which is the most face-like. This is done using the likelihood ratio given by equation (2.3) between the face and non-face models.

The approach taking in [147] to feature localization is to convolve the original image with a pair of Gaussian derivative filters (DoG) in quadrature phase. Peaks in the total energy of the filtered image are the candidates to be facial feature points. The steerable basis given by the DoG filters are well-suited for detection at diver scales and orientations. Once the initial estimates are detected, they are grouped into four PFGs (partial face groups), corresponding to the four regions of the face: top, bottom, left and right. This is useful in the case of different viewpoints, when only partial information of the face is available. The potential candidates for facial features are selected among those that satisfy some geometrical restrictions (in terms of angles and distances) imposed by the four PFGs.

To further reduce the false positives, a refinement step is used. For this step, they use a belief network represented as a tree, whose root is the face and its four children are the four PFGs detected in the previous stage. The initial probabilities for each PFG node are obtained by summing the normalized filter response and the inverse of the normalized geometric errors in each PFG in the grouping phase of feature candidates:

$$P = \frac{\sum R_i}{k_1}\left(1-\frac{\sum \epsilon_j}{k_2}\right) \tag{2.23}$$

where R_i is the filter response for the i-th feature, ϵ_j is the j-th geometric error in PFG and k_1 and k_2 are some normalized constants. Once the network is built, the belief values are updated by evidence propagation through the network. Different instances of detected faces are compared using their belief values and improbable face candidates discarded.

The main disadvantage of the local feature approaches is the lack of reliable detection of the local features. The false positive rate is very high, and the complexity of the evaluation of the face structural hypothesis grows consequently and also its uncertainty.

2.2.5 Information Theory Criteria

Another probabilistic approach to face detection are the techniques that make use of Information Theory [26], [80]. In these papers produced independently, the authors propose a learning technique that maximizes the discrimination between positive and negative examples in a training set. A family of discrete Markov processes is used to model the face and background patterns and estimate the probability models using the data statistics. An optimization step is implemented in order to select the Markov process that optimizes the information-based discrimination between the two classes. The face detection is realized by computing the likelihood ratio (2.3) using the probability models built from the training data.

As distance measure between two probability distributions (face and background), the Kullback divergence (also called *cross entropy*) was used , as a criteria to optimize the discrimination between two classes. Let X be a random process and P_X and M_X be two probability functions for X. The divergence of P with respect to M is given by the following relation:

$$H_{P||M} = \sum_X P_X \ln \frac{P_X}{M_X} \tag{2.24}$$

Let X^n be a random process and $S = \{s_i \in [1, n], i = 1, 2..., n\}$ be a list of indices such that $s_i \neq s_j$ for $i \neq j$. The k-th order Markov process $X^n(S)$ from X^n is constructed by reordering its random variables such that

$$P\left(X_{s_n} \left| X_{s_1} ... X_{s_{n-1}}\right.\right) = P\left(X_{s_n} \left| X_{s_{n-k}} ... X_{s_{n-1}}\right.\right) \tag{2.25}$$

If P_{X^n} and M_{X^n} are two probability functions for $X^n(S)$, the divergence of P with respect to M is obtained as:

$$H_{P||M}\left(X^n(S)\right) = \sum\nolimits_{i=1}^{k} H_{P||M}\left(X_{s_i} || X_{s_1} ... X_{s_{i-1}}\right) + \sum\nolimits_{i=k+1}^{n} H_{P||M}\left(X_{s_i} || X_{s_{i-k}} ... X_{s_{i-1}}\right) \tag{2.26}$$

We are interested to find S^* such that

$$H_{P||M}\left(X^n(S^*)\right) \geq H_{P||M}\left(X^n(S)\right) \qquad \forall S \tag{2.27}$$

This is equivalent to search for the optimal set of feature points. Once S^* is found, the likelihood ratio is computed as below:

$$L\left(X^n\left(S^*\right)\right)=\frac{P\left(X^n\left(S^*\right)\right)}{M\left(X^n\left(S^*\right)\right)} \tag{2.28}$$

and this response is the one that optimizes the divergence for the data belonging to the training set.

2.2.6 Approaches Based on the Face Contour

In [135], a method for face localization based on shape information is proposed. This is an efficient method for images with simple background. First, the image is enhanced using the histogram equalization technique. Afterwards, the edges are detected using a multiple-scale filter, for instance Difference of Exponential (DOE), whose formula is given by:

$$DOE=K_e\left(e^{-|x|/t_1}-e^{-|x|/t_2}\right) \tag{2.29}$$

The extracted edges are then merged based on the estimation of an energy function:

$$H=2h_0+\frac{1}{2}\int k_s\left(\nabla l_s\right)^2 ds+\frac{1}{2}\int b_s\left(\nabla^2 l_s\right)^2 ds \tag{2.30}$$

where l_s is the line contour corresponding to the arc s, k_s and b_s are elastic coefficients and h_0 is a constant to represent the base of the energy function.

The face contour is finally detected by a templated matching technique. In this case, the face template is modeled by an ellipse based on edge direction. The difference in direction between a template and a test pattern is computed based on the inner product of two vectors.

Another work that assumes the ellipse-like shape of human face is [88]. The method exhaustively searches for all ellipses associate to the edge map information, detected using the Canny's algorithm. The set of all possible ellipses is ordered according to a measure of fitness. Then, the best-fit ellipse is formed from this set, by averaging the existing ellipses. Beside this, another ellipse is calculated from this set in order to define a cost function that measures the quality of the detection process. The cost function is called *the similitude trace* T_r and is computed based on measurements of the differences between the two ellipses. Basically, T_r is the Euclidean distance between the intersection points of the two ellipses.

In [72], they built a system for face localization under non-uniform illumination conditions. The system consists of the following steps. First, they apply the Haar wavelet transform, followed by a facial edge detector. Afterthat, they detect the symmetry axis and decide the location of the face. This result is further passed to a face recognizer module.

By applying the Haar wavelet transform, the original image f of 256x240 pixels is divided in four subimages of size 128x120 pixels. The four subimages are: a low-pass filtered image f_{LL}, a horizontally low-pass filtered and vertically high-pass filtered f_{LH}, a horizontally high-pass filtered and vertically low-pass filtered f_{HL}, and a high-pass filtered image f_{HH}. f_{LL} is used for face detection, while both f_{LH} and f_{HL} are used for facial edge detection and symmetry axis. f_{HH} is not used, as it doesn't contain any useful information.

The next step consists of extracting facial edges. After the thresholding of the images f_{LH} and f_{HL}, the resulting binary images are merged using the OR operator. The resulting image is analyzed using blocks of 8x8 pixels in order to extract the facial edges by examining the connections between neighboring blocks.

In order to detect the symmetry axis of the face, they assume that the right half of the face is brighter than the left half. Thus, instead of image gradient or intensity, they use gradient orientation to detect the symmetry axis. The gradient orientation is defined as:

$$\theta\left(x,y\right) = \arctan\left(\frac{\partial f\left(x,y\right)}{\partial y} / \frac{\partial f\left(x,y\right)}{\partial x}\right) \tag{2.31}$$

The above image gradients $\frac{\partial f(x,y)}{\partial y}$ and $\frac{\partial f(x,y)}{\partial x}$ can be approximated using the subimages f_{LH} and f_{HL}. Thus, the above relation can be rewritten in the following form:

$$\theta\left(x,y\right) = \arctan\left(\frac{f_{LH}\left(x,y\right)}{f_{HL}\left(x,y\right)}\right) \tag{2.32}$$

To estimate the accuracy of the detection process, a template matching is performed between the input image $f\left(x,y\right)$ and a face template $w\left(x,y\right)$. Namely, a cross-correlation is performed:

$$\gamma_{xy} = \frac{\sum_{j=y}^{Y-1}\sum_{i=x}^{X-1}\left[f(i,j)-\overline{f}\right]\left[w(i-x,j-y)-\overline{w}\right]}{\sqrt{\sum_{j=y}^{Y-1}\sum_{i=x}^{X-1}\left[f(i,j)-\overline{f}\right]^{2}\sum_{j=y}^{Y-1}\sum_{i=x}^{X-1}\left[w(i-x,j-y)-\overline{w}\right]^{2}}} \tag{2.33}$$

where $\overline{f}$ is the average of $f(x,y)$ over the overlapped region with $w(x,y)$ and $\overline{w}$ is the average of $w(x,y)$. The value of γ_{xy} is normalized in the range $[-1,1]$. The face template is generated by averaging among 15 subimages f_{LL}.

In [145], they propose a method for face localization imposing no constraints neither to the background complexity nor to the size, orientation and additional features (like glasses, beards, etc.) of the face. In a first step, the edges are extracted from the input image. Afterwards, the best fit ellipse is searched using Genetic Algorithms (GA). The final decision is taken upon the matching of the detected region against some predefined templates.

In the first step, the image is denoised using a smoothing filter and after that edges are detected based on the Sobel operator. The resulting image is a binary one, consisting of edge pixels and background. This image is further *blurred*, i.e. the binary image is converted to a L -level quantized image. The face shape is approximated with an ellipse and several ellipses are formed from the detected edges. In terms of GA, an ellipse (a chromosome of the population) is characterized by 5 parameters: the coordinates of the ellipse's center (a,b), the orientation angle θ and the radii corresponding to each axis, r_x and r_y , respectively. Starting from an initial population, GA generates a new one, using the operators of *cross-over*, *mutation* and *selection*. The *fitness* of a chromosome is given by the following function:

$$Fitness = \frac{\sum_{j=1}^{n} f(x_j,y_j)}{L \times n} \tag{2.34}$$

where $f(x_j,y_j)$ is the quantized image value at (x_j,y_j) and n the number of pixels.

Once an ellipse is detected using GA, this doesn't mean automatically that it corresponds to a face. An additional decision phase is considered. If the following three criteria are fulfilled, then a face is declared: a symmetry line of the face can be found, a pattern matching against some predefined templates is passed and the presence of the hair can be established.

Although appealing because of its intuitive soundness and clear formulation, approaches based on the detection of the face contour have not been widely adopted in the real life systems because of the unreliable detection of the face edges and the extensive casuistic that face edge images show.

2.2.7 Color and Motion-based Segmentation

Color-based segmentation is the problem of classifying image pixels based on the partition of the color space. Clustering applied in the color space is one of the most frequent methods for color-based segmentation. The issue of the color space is usually solved using the Hue-Saturation (H, S) space or the chromatic (r, g) space. In [84], a comparison between neural and statistical approaches for color image segmentation is presented. A novel technique based on the application of morphological operators in the color space is introduced in [100] as a means to obtain more defined clusters in color space.

The motion-based segmentation is achieved basically through the difference between two consecutive frames from a video sequence. The static parts of the scene becomes black, and the moving objects in the image are highlighted in real-time. More precise detection may be based on optical flow, but at a greater computation cost.

A method based on color and motion information for face detection in complex images is described in [78]. In the first stage, they use the motion information as the primary cue for a preliminary face segmentation. By thresholding the velocity field vector, only the area that shows a clear motion is preserved. In this case, it is assumed that the face is the only possible moving object in the input sequence. In the next phase, the thresholding of hue space is used for a more precise face region localization and furthermore, knowledge-based information is used to detect facial features like eyes, eyebrows and mouth. The velocity field of the input image is computed using a modified line clustering algorithm. The face region is detected by simply taking those positions that have the velocity value between the maximum histogram velocity and a certain velocity level. As this region shows an irregular aspect, it is fit into an ellipse for a better segmentation.

Several color spaces have been tested and only the one where the segmentation process is the most effective was selected. More concrete, four different color coordinate systems have been considered: RGB, HSI, L*u*v and Karhunen-Loève transformation. Finally, it has been concluded that the HSI is the most suitable color space for face segmentation. In this space, the skin cluster is positioned in the low level of the hue value and can be segmented with a threshold in the hue axis. To find the threshold, the skin histogram in the hue axis is smoothed first, convolving it with a Gaussian kernel.

Another reference that address the problem of face segmentation using color information based on the HSI color space is [146].

In [6], they propose a method for face localization and tracking in video sequences. For face localization, both color and background information are used. Before starting the procedure, a reference background image is taken (in the current context, the user is referred as foreground and all the other objects are referred as background). The background image is updated constantly throughout the tracking process following the rule:

$$ref\_image_t\,(i,j) = ref\_image_{t-1}\,(i,j) + \lambda\,[image_t\,(i,j) - ref\_image_{t-1}\,(i,j)] \tag{2.35}$$

where by $ref\_image_t\,(i,j)$ we denote the (i,j) pixel from the reference image at the time t, $image_t\,(i,j)$ is the (i,j) position in the current image and λ controls the adaptation speed.

For face segmentation, they used the normalized r and g components of the RGB system. This color representation removes the brightness dependencies while preserving its color and has a lower dimension. In order to distinguish colors between faces and other objects, the distribution of the skin color should be known a priori within the normalized color space. Samples of users' faces are extracted and thresholded in order to define the distribution that has the greatest probability to correspond to a face. In order to speed-up the segmentation process, a LUT (lookup-table) that maps (r,g) into (R,G,B) has been implemented.

Color information alone is not able alone to an accurate face detection. Thus, it is combined with motion-based segmentation for better results. Regions recognized as both skin color and foreground are labeled as facial regions. The segmentation is further improved by applying morphological filters (like erosion/dilation) in order to remove small regions and to fill gaps between large connected regions. Other work that uses the (r,g) normalized color space for face localization is [134].

In [117], they describe an algorithm for face detection based on color and shape information. First, a supervised pixel-based color classifier is employed to find all the pixels that belong to skin space (this skin space has been built previously from a couple of skin patches). Afterwards, the previous detected regions are smoothed using GRF (Gibbs Random Field) filters to merge the contiguous areas. And finally, the detected region is fit into an ellipse. To measure the correctness of this approximation, the Haussdorf metric has been used.

As color space, they chosed the YES space. This has been proposed by SMPTE (Society of Motion Picture and Television Engineers) and can be obtained by a linear transformation from the RGB space. Y states for luminance channel and E and S are the chrominance components. Each pixel of the image (i, j) is represented as a two-valued vector whose elements corresponds to the chrominance components: $w_{ij} = [E_{ij}, S_{ij}]^T$. The probability that a given pixel x_{ij} belongs to the skin class is modelled by a two-dimensional Gaussian. A binary hypothesis test with an image-adaptive threshold is afterwards employed to decide if each pixel of the image belongs to the skin class or not.

The resulting image is further smoothed using GRF. This is achieved by maximizing the a posteriori probability of the segmentation classes **x** given the observed chrominance **w**, i.e.:

$$p(\mathbf{x}|\mathbf{w}) = \frac{p(\mathbf{w}|\mathbf{x})\, p(\mathbf{x})}{p(\mathbf{w})} \tag{2.36}$$

where $p(\mathbf{x})$ is modeled by a Gibbs distribution.

The final step of the face detection algorithm consists in discarding all those regions that are *dissimilar* to a face template, that has been modeled in this case with an ellipse. The accuracy of the detection process between the found face and the template is estimating by the Hausdorff distance. Given two set of points $A = \{a_1, a_2, ..., a_n\}$ and $B = \{b_1, b_2, ..., b_m\}$ then the Hausdorff distance is defined as:

$$H(A, B) = \max(h(A, b), h(B, A)) \tag{2.37}$$

where $h(A, B) = \max_{a \in A} \min_{b \in B} \|\mathbf{a} - \mathbf{b}\|$.

Most real-time approaches to face detection use the motion and color features. However, they are very sensitive to illumination conditions and the radiance interactions of the scene. Usually, color models are too simple and do not generalize well. Nevertheless, under controlled conditions, this is the best approach for industrial solutions.

2.3 Applications of Face Localization

In the previous section, we saw some of the techniques proposed for face localization. The face localization is a preprocessing step for face processing applications. The next *natural* step, after a face has been localized, is to pass it to a face recognition module. This can be part of a security system, whose aim is to realize a person identification/authentication to access restricted areas. On the other hand, once a face has been localized, a further process of facial feature identification and extraction could take place. The extraction of facial features (like the iris, for instance) could be also useful for person identification. The human-computer interaction (HCI) systems make also use of the whole face (for facial expression recognition) or only certain features, like eyes (for gaze tracking) or mouth (for lipreading).

In the following sections, we will review some practical applications of the facial information processing: biometric identification, multimodal interaction, wearable computing and multimedia communications. A good state-of the art review of the applications that imply human-computer interaction is found in [43].

2.3.1 Biometric Identification

Biometrics is the field dedicated to person identification/authentication based on the physiological or behavioral characteristics of the user. The systems able to perform biometric identification make use of features like face images, facial thermogram, fingerprint, hand geometry, iris, retinal scan, handwritten signature and voice patterns.

A biometric identification system should meet the following requirements: universality (each person should have the characteristic), uniqueness (no two persons should be the same in terms of the characteristic), permanence (the characteristic should not degrade in time) and collectability (the characteristic must be measurable).

A biometric system is a pattern recognition system that consists basically of two units [99]: the *enrollment* module and the identification module. The enrollment module is responsible for system training. During this phase, the biometric measurements of the persons are digitized and converted to a more compact and expressive representation, called *feature code* and stored in a database.

In the identification phase, the biometric information acquired is compared to that stored in the database and on a matching basis, the identity of the person is established. Identification can either be an identity authentication (checking if indeed a person is who he/she claims to be) or a recognition (checking if the person exists in the database).

The biometric systems based on face or iris recognition are examples of non-intrusive techniques for person identification, i.e. they don't violate person's intimacy for authentication. For instance, the advantage of a biometric system based on facial or iris recognition becomes obvious because it is no longer necessary to remember numerical codes to have private access in restrained areas (airports, a computer network) or to access certain services. Nowadays, the most common service where the iris pattern can be used instead of a PIN is cash dispensing machines.

There are some face-based commercial biometric systems. Examples are *TrueFace* (from *eTrue* company [39]) and *FaceIt* (from *Visionics* [132]). Both systems are based on the concept of Local Feature Analysis (LFA) that is introduced in [104].

But even the most reliable face recognition systems has their limitations. The performance of such systems is strongly eroded by changes in illumination conditions, modification in the face appearence due to presence/absence of glasses, beard, etc. On the other hand, face aspect varies over long period of times.

For these reasons, iris patterns became a very interesting alternative for visual recognition. Several research studies revealed that the iris is relatively insensitive to angle of illumination and changes in viewing angle cause only affine transformations [29]. The pattern of iris is invariant in time and its variability among different persons is enormous, from this point of view being similar with fingerprints. Face detection and robust face feature localization play an important role in the construction of less intrusive systems for iris recognition.

IrisScan, from *Iridian Technologies* [61], is an example of a system that performs person identification based on iris recognition. The scientific basis of this system is given in [30].

Experimental iris recognition systems are tested for a wide-range of applications. Only to name some of them we could mention the fact that in United Kingdom, The Nationwide Building Society introduced iris recognition within its cash dispensing machines (in lieu of PIN numbers) as of 1998. A new development at some airports (Charlotte/Douglas International

Airport in North Carolina, USA) is ticketless air travel [120], allowing passenger and baggage check-in and other security procedures based on the traveller's iris patterns. The iris recognition technology has been also featured throughout 2000 at Millennium Dome (London, England) and at EXPO2000 in Hannover, Germany.

Combination of biometric cues is a promising approach to enhance recognition verification accuracy [31]. An example where two biometrics have been combined for person identification is [60]. More concrete, the authors integrated faces and fingerprints. The motivation of this mix is offered by the evidence that face recognition is fast but not very reliable, meanwhile fingerprint verification is reliable but inefficient in database retrieval. By the combination of these two features, the prototype performances are better than those of the face recognition module considered alone.

2.3.2 Multimodal Interaction

Even if speech is the most common way of communication between humans, we very often make use of our facial expression and gesture in order to reinforce the idea expressed through our verbal communication. Nowadays, despite technologies advances, the interaction between humans and computers remains widely dominated by *classical interfaces* represented by such devices like keyboards, mice, touch-screens or *data-gloves* (for Virtual Reality applications). There is a stream of sustained efforts in order to make the interaction between user and machine more humans-like. There are some progress regarding the verbal interaction, in order to make computers able to understand operator's commands. On the other hand, there is much interest to enhance the interaction between humans and computers to make use also of non-verbal communication, like facial expression and gesture recognition. This can be achieved by providing the system with a video-camera and use the input to understand the user action. Thus, by combining the audio and visual cues, we can talk about *multimodal interaction* between users and machines. Examples of multimodal interaction are: lip-reading, gaze-tracking, facial expression and hand gesture recognition. Many of these multimodal interaction means rely on robust face localization.

Automatic *lip-reading* is a primary application resulted from the combination of audio and video channels is. Lip movement is a visual information strongly related with the speech process, thus it can be viewed as an integer part of the speech recognition task. By adding the visual information consisting of lip movement, the aim is to reinforce the audio speech recognition, not to substitute it. Understanding commands in very noisy environments, or focusing the attention to one speaker in a multitude of people (*cocktail party effect* [4]*)*, are practical applications of lip-reading.

Lip-reading systems must solve three problems: the recognition of speech based on audio signals, the recognition of speech based on visual signals and the fusion of both recognition processes. Audio processing follows conventional speech recognition techniques based on Hidden Markov Models (HMM) or Time-Delay Neural Networks (TDNN). The same approaches have been tried for the visual recognition, when extended to deal with images. In [133], they used a Multi-State Time Delay Neural Network (MS-TDNN) to perform this recognition. Other methods proposed for lip-reading through lip movement tracking are based on the computation of optical flow [90], building and training a 3D lip model [9] and Hidden Markov Model (HMM) [109].

Another type of multimodal interaction between the user and machine is through eye or gaze tracking. The function of a gaze tracking system can be either active or passive. For example, a system can identify user's message target by monitoring the user's gaze, or the user could use his gaze to directly control an application or launch actions. Current approaches to gaze tracking use active sensing to measure the the orientation of the subject's eyes. The eye is illuminated with infrared light and the gaze direction is estimated based on the physiological properties of the eyes [56]. Nevertheless, this is an intrusive method for gaze tracking and despite its performance, the main drawback is represented by the system calibration. The system is very sensible to very small movement of subject's head, thus the person should remain perfectly still.

As an alternative to this method, there is a passive, vision-based àpproach that proved to be tolerant with large head movements and able to follow a person's gaze at some distance using little or no calibration. In [49], they propose the gaze detection and tracking method through the identification of head pose from a single, monocular view of face. For this purpose, they create a geometric model of the face, defined by the outer corners of the eyes and mouth. The plane defined by these points is called *facial plane*. The facial models comprise a single ratio $R = L_e/L_f$, where L_e is the distance between the eyes' corners and L_f is the distance, on the symmetry axis of the face, between the center of the mouth and the center of the line that connects the eyes.

Another work exploits the problem of gaze detection and tracking using self-organization maps [10]. First, an eye is detected in a face image and the gaze direction is estimated by computing the position of the pupil with respect to the center of the eye. The system is based on unsupervised learning. It creates a map of self-organized gray scale image units that collectively describe the eye outline.

A practical application of gaze tracking and head pose estimation is reported in [7], where they used these visual cues to control the cursor movement on the screen. The same idea of controlling a cursor has been pursued in [126]. Here, in order to estimate the head pose, the tip of the nose has been taken as reference. The vision system robustly tracks 3D face position and orientation in real time using a framework called Incremental Focus of Attention (IFA). IFA integrates tracking based on multiple cues (including color, intensity templates, and dark point features) which cooperate to track under adverse visual conditions. The pose recovered from tracking is then used to compute the intersection between the plane of the monitor screen and an imaginary ray extending forward from the user's nose.

Another research issue of great interest in multimodal interfaces is the understanding of facial expression. With facial expression recognition, a new concept emerged in the human-computer interaction known as *Affective Computing* [108], i.e. the ability of the systems to detect and adapt themselves to the user's emotional states. Related with this aspect, is also the intent to enable the virtual actors (*talking heads*) the ability of showing emotional states through changes in facial expression. The virtual news speaker Ananova [3] is such an example. An overview of several aspects related with affective computing, referring to intentions in language, facial expressions and gestures, can be found in [94].

The psychological framework of the facial expression recognition was established as early as 1978 by Ekman et al. in [40]. They designed a system for describing all visually distinguishable facial movements called the *Facial Action Coding System* (FACS). In this system, each expression can be represented in terms of action units. Thus, the automatic recognition of facial expression can be achieved by categorizing a set of predetermined facial motions. Ekman considered in [41] that there are six fundamental facial expressions: anger, happiness, sadness, fear, disgust and surprise.

One of the first attempts for automatic facial expression recognition was due to Mase and reported in [91]. The work was based on the computation of the optical flow over a sequence of consecutive frames. This approach has been extended in [142] by detecting motion in six predefined and hand-picked rectangular regions on a face and then correlate this motion with simplified versions of the FACS rules for the six expression recognition.

An approach based on the analysis of whole-face dynamics has been proposed by Essa and Pentland in [44]. They have analyzed video data of facial expressions and then probabilistically characterized the facial muscle activation associated with each expression. This characterization is achieved using a detailed, physically based dynamic model of the skin and muscles coupled with optimal estimates of optical flow in a feedback-controlled framework. They used 2D spatiotemporal templates to represent facial expressions. The main drawback of their system is that it is not suitable for real-time applications. Further developments, reported in [28] present a method for facial tracking and interactive animation of faces that runs in real-time. The idea was to realize a fine-grained analysis of a subject's expression and then store the spatiotemporal representation of this expression on a generic model of a face. Then simple visual measurements can be used to establish the relationship between an image and the dynamic motion parameters of the model. These measurements are coupled with the parameters of the physical-based model using an interpolation process, resulting in a real-time facial tracking and animation system. In addition, Hidden Markov Models (HMM[1]) can be used for recognition of expressions based on a similar set of visual measurements.

Even if it seems a topic collateral to the current purpose of this thesis, but for the sake of completeness, we will refer very briefly to hand and gesture recognition, because they play also a major role in human-computer interaction, as a non-verbal communication cue. The motivation to enhance a computer with the ability to understand human gestures is determined by the evidence that for humans, the gestures facilitate the understanding of ideas expressed through verbal communication. Furthermore, the hand and gesture recognition become the main cue of communication for the persons affected by certain classes of diseases that make the verbal communication impossible. For instance, some attempts are reported in order to enable

[1] Please not confuse with the Heteroassociative Morphological Memories (HMM) of chapter 5.

machines to recognize the *alphabet* of the American Sign Language [144].

As the hands have the same color as faces, it is straightforward to assume that color-based approaches applied for face detection can be successfully applied for hand detection and tracking. Besides this, other methods like optical flow analysis. [110], Hidden Markov Model [97] and probabilistic learning [96] has been also applied. For a more comprehensive overview of techniques and applications of hand gestures recognition please refer to [101].

The hand gesture recognition has been implemented for robotic applications, too. For instance, in [85] a visual system is used to control a robot arm for object manipulation through hand gestures recognition. Several neural networks are integrated to a single system that visually recognizes human hand pointing gestures from stereo pairs of color video images. The output of the hand recognition stage is further processed by a set of color-sensitive neural networks to determine the Cartesian location of the target object that is referenced by the pointing gesture. Finally, this information is used to guide a robot to pick the target object and place it at another location that can be specified by a second pointing gesture. This work is an example of *learning-by-watching* paradigm [5] and [75].

The several multimodal cues presented above can be integrated in different devices or home appliances in order to make easier and user-friendly the interaction between humans and machines. This integration can be resumed in the term of *Perceptual User Interfaces* (PUI), i.e. interfaces that are characterized by interaction techniques that combine an understanding of human capabilities (emotion and intention expression through speech, gesture and facial expression recognition) with computer's I/O devices and machine perception and reasoning [130]. PUI seek to make the user interface with machines more natural, taking advantage the way humans interact each other. Devices and sensors should be transparent and non-intrusive and machines should perceive relevant human communication channels as well as generate responses that are easily understood. An in-depth analysis of how PUIs are embedded in machines (robots) that offer assistance to human user can be find in [98].

Recently, PUIs are redirected toward a sector that present perhaps the most dynamic evolution rate: intelligent toys. Two years ago, SONY announced the production of robot-dog called AIBO [1]. Sensing consists of a bunch of sensors, microphones and video-cameras spreaded all-over its body. It is able to express the six basic emotions and can become friendly or furious..

A voice recognition system understands the *AIBO tonal language* and furthermore, learns its given name. Its design allows to adaptability learn new skills (play with a ball for instance) from direct interaction with the user.

2.3.3 Wearable Computing and Smart Environments

Once computers are enabled with PUIs, they are able to detect, track and identify people, as a primary task, and, at a higher level, to interpret human behavior.. The principle behind this latter task is represented by *wearable devices* and *smart environments*. Some authors call it the *fourth generation* of computers [105]. These embedded computing devices will be everywhere: clothes, home, car and offices. Their economic impact and cultural significance are expected to shadow the previous computing generations. It is a broad and challenging field of research for innovative computational techniques.

Such systems must know enough about the people in their environment so they can act appropriately with the minimum of explicit instruction. In other words, the focus is shifted to the *context sensing* problem and the detection of user's intentionality. Several channels of communication form the domain information: facial, hand, whole-body and voice. A time-scale analysis on each of these information sources seems appropriate to get valuable information in order to foresee user's intentions. At the longest time scale are semipermanent physical attributes like facial shape and appearance, vocal characteristics, body shape and basic movements (like walk, for instance). At a shorter time scale, we have goal-directed behaviors which usually have durations raging from several minutes to several hours (like speaking to phone, talking in a conference, walking to reach a certain destination). In their turn, behaviors can be decomposed in a multimodal sequence of individual actions, such as pointing, grasping, etc. At the lowest level of the hierarchy, these actions consists of physical observations such as image analysis (for instance,.face localization or shape recognition), which in turn it are not motivated by any intentionality at all. Thus, user's intentionality can be predicted by following the ascending path in a such hierarchy.

From a point of view, intentionality can be determined by the observations of *direct behaviors*. These class of behaviors are in general related with robotics and their purpose is to influence the surrounding physical environment (for instance by moving certain objects). On the other hand, we can talk also about *communicative behaviors* (voice-based interaction, facial or hand gestures, gaze direction, etc.). In this case, the intention is not determined by an explicit action, but it is reflected in the influence that such a behavior has on other agents and it is often referred as higher-order intentionality. To interpret communicative behaviors, the system must know more about the context and the goal of interaction. For instance, we can figure the gesture of pointed with a finger with an extended arm. This can be interpreted either as the intention to activate a button (direct behavior) or just a pointing gesture (communicative behavior).

The following examples about wearable computing are under development by several research groups at the Media Laboratory, from MIT. First of them is a wearable video camera [57]. StartleCam is a wearable video camera, computer, and sensing system, which enables the camera to be controlled via both conscious and preconscious events involving the wearer. Traditionally, a wearer consciously hits record on the video camera, or runs a computer script to trigger the camera according to some pre-specified frequency. The system described here offers an additional option: images are saved by the system when it detects certain events of supposed interest to the wearer. The implementation described here aims to capture events that are likely to get the user's attention and to be remembered. Attention and memory are highly correlated with what psychologists call arousal level, and the latter is often signaled by skin conductivity changes; consequently, StartleCam monitors the wearer's skin conductivity. StartleCam looks for patterns indicative of a "startle response" in the skin conductivity signal. When this response is detected, a buffer of digital images, recently captured by the wearer's digital camera, is downloaded and optionally transmitted wirelessly to a webserver. This selective storage of digital images creates a "flashbulb" memory archive for the wearable which aims to mimic the wearer's own selective memory response. Using a startle detection filter, the StartleCam system has been demonstrated to work on several wearers in both indoor and outdoor changing environments.

The second example is an orchestral conductor's jacket, in fact a wearable physiological monitoring system that has been designed to provide a test bed for the study of emotional expression as it relates to musical performance [89]. Their conclusions indicate that several forms of expressive communication can be measured and detected in physiological signals. These include the use of handedness to emphasize musical changes, the signaling of upcoming events with sudden changes in effort, the difference between information-bearing and non-information-bearing gestures, the indication of intensity and loudness with changes in muscular force, and the use of breathing to express phrasing in the music.

Besides wearable computing, another application for PUIs are *smart environments*, that is closed areas where the people is tracked, their actions understood and eventually assisted. For instance, imagine a room where can be *aware* of how many people are present, what they are doing, etc. Or a car that senses when you are tired and warns you to pull over. Or smart office that has been personalized thus once you step in, the lights are turning on, the air conditioning is starting, coffee maker is preparing the coffee or even smart desks that knows your preferences of how the items has to be arranged [106].

A practical example is offered by *Kidsroom* [12], a project developed at the Media MIT Laboratory. Kidsroom is a fully automated and interactive narrative playspace for children. Built to explore the design of perceptually based interactive interfaces, the Kidsroom uses computer vision action recognition simultaneously with computerized control of images, video, light, music, sound and narration to guide children through a storybook adventure. The Kidsroom does not require that users wear any kind of wearable computing. The system was designed to use computational perception to keep most interaction in the real, physical space, even if the participants interact with virtual characters.

Let's now take a look at the room architecture. Two walls resemble the real walls in a child's room, complete with real furniture, posters and windows. The other two walls are large, back-projected video screens used to transform the appearance of the room environment. Three video cameras overlooking the space provide input to computer vision people-tracking and action recognition algorithms (up to four children can be monitored by the system).

During the story, children interact with objects in the room as well as with virtual characters projected onto the screen-walls. Perceptual recognition makes it possible for the room to respond to the physical actions of the children by appropriately moving the story forward thereby creating a compelling interactive narrative experience. The vision systems use the context established by the story for robust and realistic performances. Although the context of the story is linear, the room continuously reacts to the children's actions, giving the environment an interactive feel. For instance, after a short *walk*, the children reach the river world, and the narrator informs that the bed magically transformed in a boat. Thus, the children are asked to *paddle* to make it move, which is represented by images of the river flowing by on the screens.

2.3.4 Multimedia Communications

Other area where the techniques of face detection, tracking and recognition can be put in practice, resulting in very useful applications, is represented by multimedia communication systems. This is a very broad field and that is why we will try to give our own view of it. It contains, but it is not limited to, the following aspects: video-telephony, video-conferencing, content-based multimedia database retrieving, virtual actors (talking heads), etc.

In a video-telephone application an active video system can be used to keep the face centered in the image, and with a reasonable lightning. Keeping the face at the same place, same scale and same intensity levels could represent a dramatically decreasing of the amount of information of the information to be transmitted. One possible coding for this purpose can be obtained by using a PCA representation of the last few frames (taken over the last 10-15 seconds, for instance). Once the eigenvectors are transmitted, subsequent images can be transmitted as a much smaller vector of coefficients resulted from the projection of the image onto the space spawned by the eigenvectors. Other image codings can also be accelerated if the face image is normalized in position, size, gray scale and held in focus. On the other hand, video-conference applications aims to follow the same scenario.

This is a very promising technique to achieve very large bit-rate reductions for moving images. Model-based coding [102] represents a scheme for video compression based on the representation of the moving objects present in the sequence, in terms of attributes such as size, location, quantity of motion. This is useful in broadcasting digital video over Internet, where the transmission band has a limited capacity. In order to implement it, a combination of computer graphics and computer vision techniques are needed. The information regarding object attributes is used to synthesize, by computer graphics methods, a model of each object. Tracking techniques are used to make the model mimic the movements of the object it represents. The parameters needed to animate the model are coded and transmitted to the receiver, which reconstructs the model.

The interest in the model-based coding for video transmission has been materialized in standards like MPEG-4 or MPEG-7. One of the first objects submitted to this coding procedures were the human faces. From a point of view, faces are rather easy to be modelled due to their relative invariant geometric shape, many times being approximated by an ellipse. But on the other hand, the difficulty in building an accurate face model arises from its internal flexible structure, represented by local deformations due to muscular activity.

As both analysis and synthesis techniques have been improved recently, realistic 3D models of faces, showing facial animation, for low-bit rate enconding, are highly attainable now. However, further research is looking to extend the work realized for face model-based coding to other classes of objects.

In [123], a system that uses text captions for image understanding is presented with a particular application for human face identification in newspaper photographs. The broader concept is to identify relevant information in the text commentary, extract and represent this information and finally, using it in a computer vision system in the task of image understanding. The information contained in both pictures and captions enhances overall understanding of the accompanying text, and often contributes additional information not contained specifically in the text. This information can be further incorporated into an integrated database as part of a larger information retrieval system which permits content-based retrieval, using for instance some keywords. PICTION, a multi-stage system, parses a natural-language caption of a newspaper photograph and afterwards, it is able to locate, label and give information about objects which are relevant. As an integrated language/vision system, it's essential task is to extract visual information from text. It uses four types of knowledge databases (KD): (i) a lexical KD which models the syntax and semantics of words as well as their interconnections, (ii) a visual KD which contains object schemes (declarative and procedural model of the objects), (iii) a world KD that contains information about people, events and general domain constraints and (iv) a picture specific KD which contains facts specific to previously processed pictures and captions. A key component of this system is the face localization module. Besides a vague reference about wavelets and active contour based techniques, little details are given.

In [20], another system, called *SpotIt,* is developed for browsing large mugshots databases and creating identikits of photographic quality. The two functions are interrelated: the available database provides feedback to the user building the identikit and the identikit itself can be used as an access key to the image database. The system provides a virtually unlimited set of alternative features that can be browsed efficiently in the appropriate context, interactive holistic feature modification coupled to syntactic access to a feature database and quantitative, automatic computation of face similarities, providing real-time feedback of the system which constantly shows the most promising matches to the identikit being built.

The initial gray scale images are annonated (taking the person's identity, race, age) and normalized(in size and orientation). The system allows to work at both pixel and geometrical level. The faces are decomposed in a set of features: eyes, nose, chin, etc.. In order to limit the data dimensionality, for each of these features a PCA technique is used, i.e. a feature is represented at the database level only by the most significant eigenvectors.

When used as identikit, the system enables the user to easily set and modify each facial feature. The similarity between the identikit image X and each of the images stored in the database I_i is given by:

$$d(X, I_i) = \sum_{F=1}^{Q} a_F \omega_F \left| c_F - c_{F_i} \right|^2$$

where Q is the number of available characteristics, F is the feature label, $a_F = 1$ if a certain feature is present in the user's model and 0 otherwise. The weights ω_F are normalizing factors for the different features and c_{F_i} are coefficients normalized in the range $[0, 1]$ that are associated with feature coordinates. As result, the whole database is sorted by increasing values of $d(X, I_i)$.

When the system is used as a database browser, the identikit can be used as a key to access the stored items: each feature can be considered as a user-selectable fields.. The result of the query is a list of images sorted by the similarity with the access key. Another way of browsing the database is moving along a tree structure computed by hierarchical clustering of the available images. The database is then recursively partitioned into four clusters, until the number of images within a cluster is less than four. Each cluster is then represented by the element nearest to its center. The branches of the resulting tree lead to faces which are more and more similar to one another. The system also shows for each of the representatives its distance from the available identikit image, thereby guiding the user exploration.

Chapter 3

Face Localization with the Gray Scale Hit-or-Miss Transform

This chapter is devoted to a first principles attempt to apply a morphological operator to face localization. It starts with an introduction of morphology ideas for image processing, then reviews the binary morphology definitions and the gray scale generalization of the Hit-or-Miss Transform which is the relevant operator to the this task. Finally, we report results of experiments comparing two instances of gray scale Hit-or-Miss Transform definitions.

3.1 Introduction

The interest in mathematical morphology in computer vision stems from its emphasis in shape information. Morphological operators are basically shape operators and their composition allow the natural manipulation of shapes for the identification and the composition of objects and object features.

The Morphological Hit-or-Miss Transform (HMT) (its definition can be found in any of the following references: [53], [55], [118], [50] and [121]) is the localization operator in mathematical morphology. It finds occurrences of an object and its nominal surroundings in a set or image. It is a natural operation to select out pixels that have certain geometric properties, such as corner points, isolated points or border points and that performs template matching. This transform is accomplished by using intersection of erosions. HMT has been used as a building block for complex morphological operators like thinning, thickening, skeletonization and image restoration or denoising filters.

For instance, in [33], Dougherty uses the HMT to build up optimal filters for gray scale signals. This is accomplished by applying a union of HMT to a signal's umbra and then taking the surface of the filtered umbra as the estimate of the ideal signal. The HMT is built up in order to provide the optimal mean-absolute-error restoration for both the ideal signal and its umbra. The method is developed for thinning HMT filters, but can be extended also for thickening filters. The same theory can be adapted to Hit-or-Miss filtering of gray scale images through the binary HMT on 3D sets.

The HMT has some direct application in the area of optical character recognition (OCR). In [34], a shape is viewed as a random process satisfying various model constraints and the recognition process is analyzed relative to the process. Expectations of various types of recognition errors are analyzed. The paradigm used here refers to measuring recognition efficiency and drawing a criterion of optimality from the probability model. Thus, optimal recognition results from finding a class of hit-or-miss structuring element pairs that results in minimal error, as measured by expectation relative to the shape processprocess model. This work follows the track of [149] that introduce conditions that allow the use of object boundary as the structural element for robust shape recognition in binary images via the HMT.

Although it has been proven that HMT is sensible to the types of noise found in scanned images, a robust extension has been introduced in [11] , called the *Blur Hit-or-Miss Transform*. The robustness of the new operator derives from its ability to treat both types of noise together and to remove them by appropriate dilations. For images with a random component of noise, the Blur HMT (BHMT) results to be sensitive only to the size of the noise. In the implementation of the Blur HMT, the foreground and background images are dilated with structuring elements that depend on image noise and pattern variability and afterwards the results are eroded with templates extracted from the patterns to be matched. A metric based on the BHMT is defined so that a map of distances to the patterns can be computed for a given image.

These HMT-based algorithms work upon binary images or sets that are the umbras of signals. We are interested in gray scale images for which the umbra approach implies the processing of 3D sets. We are interested in performing pattern recognition with gray scale

HMT and more specifically in its application to face localization.

3.2 Binary Hit-or-Miss Transform

The language of mathematical morphology is that of set theory [55] , [118], [50] and [121]. Sets in mathematical morphology represent the shapes that are manifested in binary or gray scale images. Sets in the Euclidean 2D space denote foreground regions in binary images. Sets in the Euclidean 3D space may denote time-varying binary image sequences, or static gray scale images. Mathematical morphology transformations apply to sets of any dimensions like Euclidean N-space or the set of N-tuples of integer, $\mathbb{Z}^N$. Those points in a set being morphologically transformed are considered to be the selected set of points, the foreground in binary images. The points in the background are those in the compliment of the foreground.

The primary morphological operators are dilation and erosion. The composition of dilation and erosion gives the opening, closing, hit-or-miss or other morphological operators. Dilation is the morphological transformation that combines two sets by using vector addition of set elements and corresponds to the Minkovsky addition.

Definition 1 *Given A and B sets in arbitrary space (E^N), the following are equivalent definitions of dilation:*

$$A \oplus B = \left\{ c \in E^N \left| c = a + b, a \in A, b \in B \right. \right\}. \tag{3.1}$$

$$A \oplus B = \bigcup_{b \in B} A_b. \tag{3.2}$$

where A_b is the translation of set A by point b. Formally, $A_b = \left\{ c \in E^N \left| c = a + b, a \in A \right. \right\}$.

Erosion is the morphological dual of dilation. It is the morphological transformation that combines two sets by using inclusion. It is a close relative to Minkowsky substraction [1].

Definition 2 *Given A and B sets in arbitrary space (E^N), the following are equivalent definitions of erosion:*

$$A \ominus B = \left\{ c \in E^N \left| c - b \in A, b \in B \right. \right\} = \left\{ c \in E^N \left| B_c \subseteq A \right. \right\}. \tag{3.3}$$

[1] A nice dicussion on the history of the morphological definitions and its relation with the set theory history is found in [63].

$$A \ominus B = \bigcap_{b \in B} A_{-b}. \tag{3.4}$$

For both dilation and erosion, the set B is called the structuring element, because its shape conditions the result of the operation and specifies the shape primitives considered in the analysis of the images. Proofs of a number of properties (duality, increasing, extensivity, anti-extensivity) can be found in the literature.

Although the conventional presentation of mathematical morphology for image analysis starts with the presentation of the erosion and dilation operators, the original works of Serra [118] started proposing the Hit-or-Miss Transform (HMT) defined as the point by point transformation of a set X that works as follows: choose a structuring element B, i.e. composed of two sets B^1 and B^2. A point x belongs to the HMT if and only if B^1_x is included in X and B^2_x is included in the complement of X, X^c:

$$X \otimes B = \left\{ x \left| B^1_x \subseteq X; B^2_x \subseteq X^c \right. \right\}. \tag{3.5}$$

Erosion is then presented as a special case of the HMT when B^2 is the empty set. The conventional definition of the HMT is an intersection of erosions

Definition 3 *Given A and B sets in arbitrary space (E^N), the Hit-or-Miss Transform is defined as follows:*

$$X \otimes \left(B^1, B^2\right) = \left(X \ominus B^1\right) \cap \left(X^c \ominus B^2\right). \tag{3.6}$$

It is common practice to assume a fixed window W of dimension $(2M+1) \times (2M+1)$ as the domain of definition of the structural element, so that $X \otimes B$ is equivalent to $X \otimes (B, W - B)$. The works of Serra are based on the definition of a topology based on the HMT. However, the mainstream of the literature reduces the role of the HMT to a specialized building block for shape extraction and filtering operations like skeletonization, thinning and thickening.

The optimal design of gray scale signal filters based on the HMT [33] looks after the optimal construction of a kernel of structural elements. Following the results of [8], any filter is translation invariant if it can be constructed as a union of hit-or-miss transformations:

$$\Phi\left(S\right) = \cup \left\{ S \otimes B \left| B \in \mathbf{B} \right. \right\}, \tag{3.7}$$

where $\mathbf{B}$ is a collection of hit-or-miss templates. The optimality of the selection follows from the computation of probabilities of templates B being observed in the umbra of the original uncorrupted signal: $P\left(S\left(x\right)=1\left|B_x\right.\right)$. If this quantity is greater than 0.5, B is included in the kernel. The question answered in [33] is how to ensure that the filters constructed in this way are *umbra* preserving given that the HMT is not umbra preserving in general.

We are interested in the shape recognition where B^1 is identified with the shape to be recognized and B^2 is a windowed complement of B^1. The point set result of the HMT indicates the occurrence of the shape in the image. In summary: the shape A in window W occurs in the image I at, and only at, the locations represented by set L, where $L = I \otimes (A, W - A)$.

In many situation, the shapes of the objects in the image, may not be exactly the same as the ideal original shapes used for generating the structuring elements. Then, a pattern is taken as a family of sets $\{A\left(\gamma\right), \gamma \in \Gamma\}$ along with the convention that the pattern $\{A\left(\gamma\right), \gamma \in \Gamma\}$ occurs in the image at z if, and only if, $A_z\left(\gamma\right)$ occurs in the image for some $\gamma \in \Gamma$. This concept is formalized Crimmins and Brown in [27]: the pattern $\{A\left(\gamma\right)\}$ occurs in the image at the location represented by the set $L = \bigcup\limits_{\gamma\in\Gamma}\left[I \otimes \left(A\left(\gamma\right), W - A\left(\gamma\right)\right)\right]$. The robust recognition depends on the definition of the pattern. To reduce the computational cost, Zhao and Daut in [149] show that if shape A, $A \subset W$, in image I is separated from other shapes by W and either $\partial A \nsubseteq B$ or $\partial A^c \nsubseteq B^c$, where ∂A is the boundary of A and B is any other shape in the image, then the shape A occurs in the image at, and only at, the location represented by the set $L = I \otimes \left(\partial A, \partial\left(W - A\right)\right)$. The natural corollary is that pattern $\{A\left(\gamma\right)\}$ occurs in image I at locations $L = \bigcup\limits_{\gamma\in\Gamma}\left[I \otimes \left(\partial A\left(\gamma\right), \partial\left(W - A\left(\gamma\right)\right)\right)\right]$. It is also shown in [149] that if upper and lower bounds $\overline{A} \supset A \supset \underline{A}$ on the variation of the shape are known, then it is possible to locate a deformation $\widetilde{A}$ of A inside the set of points $L = I \otimes \left(\partial\underline{A}, \partial\overline{A}^c\right)$. This work is extended to non-deterministic deformations in [34], where probabilistic criteria to select sparse structuring elements are introduced. However, the construction of the structuring elements needs to be done using heuristics, with very great computational cost.

Another solution to the problem of pattern localization allowing for inexact matching is the Blur HMT presented by Bloomberg and Vincent in [11]:

$$X \otimes \left(B^1, B^2; \beta^1, \beta^2\right) = \left(\left(X \oplus \beta^1\right) \ominus B^1\right) \cap \left(\left(X^c \oplus \beta^2\right) \ominus B^2\right) \tag{3.8}$$

where β^1, β^2 are the blur structural elements that introduce the flexibility in the matching process. The radius of β^1, β^2 induce a metric that measures the similitude of the patterns.

These works set the stage of our goals, although they are restricted to binary images. We look forward to obtain a kind of HMT-based operator to localize face instances in gray scale images.

3.3 Gray scale Hit-or-Miss Transform

The binary morphological operations of dilation and erosion are naturally extended to gray scale images by the use of *max*, *min* operators. In [55] and [50] the binary operators are extended to gray scale images via the concept of top surface and umbra of a set. Gray scale dilation and erosion are defined in terms of the dilation and erosion of the umbras.

Suppose A is a subset in the Euclidean N-space. The convention is that $(N-1)$ coordinates are the spatial domain of A, the N-th coordinate is for the surface. $N=3$ for gray scale images. Let $A \subseteq E^N$, $F = \left\{x \in E^{N-1} \left| \exists\, y \in E, (x,y) \in A \right.\right\}$, the *top-surface* of A, denoted by $T[A] : F \to E$ is defined by

$$T[A](x) = \max\left\{y \left| (x,y) \in A \right.\right\}. \tag{3.9}$$

Conversely, a set $A \subseteq E^{N-1} \times E$ is an *umbra* if, and only if, $(x,y) \in A$ implies that $(x,z) \in A$,$\forall\, z \leq y$. For any function $f : F \to E$ defined on $F \subseteq E^{N-1}$, the umbra of f is a set consisting of the surface f and everything below the surface: the umbra of f denoted by $U[f]$, $U[f] \subseteq F \times E$, is defined by

$$U[f] = \left\{(x,y) \in F \times E \left| y \leq f(x) \right.\right\}. \tag{3.10}$$

The gray scale dilation of two functions is defined by the surface of the dilation of their umbras. In the same way, the gray scale erosion is defined as the erosion of the umbras. In the following, let $F, G \subseteq E^{N-1}$ and let $f : F \to E$ and $g : G \to E$.

Definition 4 *The dilation of f by g is denoted by $f \oplus g : F \oplus G \rightarrow E$, and is defined by*

$$f \oplus g = T\left[U\left[f\right] \oplus U\left[g\right]\right]. \tag{3.11}$$

This definition is equivalent to:

$$\left(f \oplus g\right)(x) = \max\left\{f\left(x-z\right) + g\left(z\right) | z \in G, x-z \in F\right\}. \tag{3.12}$$

Definition 5 *The erosion of f by g is denoted by $f \ominus g : F \ominus G \rightarrow E$ and is defined by:*

$$f \ominus g = T\left[U\left[f\right] \ominus U\left[g\right]\right], \tag{3.13}$$

which is equivalent to the definition

$$\left(f \ominus g\right)(x) = \min_{z \in G}\left\{f\left(x+z\right) - g\left(z\right)\right\} \tag{3.14}$$

The basic relationship between surface and umbra is that the surface is the left inverse over the umbra: $T\left[U\left[f\right]\right] = f$. The umbra operation, however, is not an inverse of the surface operation, except when the set is an umbra, then the umbra of the surface of the set is the set itself. The umbra homomorphism theorem states that the operation of taking the umbra is an homomorphism from gray scale morphology to the binary morphology:

$$U\left[f \oplus g\right] = U\left[f\right] \oplus U\left[g\right] \tag{3.15}$$

This theorem allows to prove easily the properties on gray scale morphology using the corresponding results in binary morphology.

The definition of the gray scale HMT does not fit well in this picture, because HMT is not umbra preserving, i.e.

$$U\left[f\right] \otimes \left(U\left[g^1\right], U\left[g^2\right]\right) \tag{3.16}$$

is not an umbra in the general case.

There have been scarce attempts to generalize the definition of the HMT to gray scale images. Among them, Kosravi and Shafer in [69] propose the following definition[2]:

Definition 6 *The gray scale Hit-or-Miss Transform is defined as the sum of two gray scale erosions:*

$$(f \otimes g)(x) = [(f \ominus g)(x)] + [((-f) \ominus (-g))(x)]. \tag{3.17}$$

We will call this transform gray scale HMT (GHMT), following the convention of the authors. It can be seen the result of this operation is an umbra, but it cannot be shown to be umbra preserving: $U[f \otimes g] \neq U[f] \otimes U[g]$. The first erosion tests that the template matches the image from *above*, and the second one, from *below*. The *perfect matching* occurs when both matches happen simultaneously. Using the definition of gray scale erosion and dilation, the above relation can be rewritten into the form:

$$(f \otimes g)(x) = \min_{z \in G} \{f(x+z) - g(z)\} - \max_{z \in G} \{f(x+z) - g(z)\}. \tag{3.18}$$

The last relation says that the result of the GHMT is always negative or at most equal to 0. It can be shown that $f \otimes g$ takes value 0, if and only if $f(x+z) = g(z) + k$ for every $z \in G$, where k is a constant. Matching by the GHMT is invariant to constant shifts in image intensity. However, small perturbations of the pattern, such as those produced by impulsive noise, may lead to very big values of the GHMT. The effective matching distance is given by:

$$d_{GHMT}(f, g)(x) = -(f \otimes g)(x).$$

The authors in [69] report an experimental study that compares the GHMT against normalized correlation for searching templates with small windows of varying length in 1D signals corrupted with impulsive and Gaussian noise. To improve the robust recognition in the presence of noise [69] propose a generalization of the GHMT called Order Statistic Difference (OSD), however the results did not improve much over the normalized correlation. We have tested it as face locator given a set of patterns.

[2] Let us point out an inconsistency in this definition. The Binary HMT is defined as an intersection of erosions. To be consistent with this definition, the GHMT definition would need to be a product of gray scale erosions.

3.4 Gray scale HMT based on Level Sets

We can interpret the image as a topographic map where local elevation corresponds to the gray value in the image (i.e. [54]). The image is formed by piling up the Level Sets, with lowest level at the bottom and the highest on the top. Level sets are called level slices in [35], where they are the basis for the definition of stack filters. Let us consider a gray-level image as a function of its coordinates versus intensity, i.e. $f : F \subset E^N \to \{0, 1, ..., n-1\}$ where n is the highest gray value of the image. The image f can be decomposed into its gray scale sets as given by the formula:

$$S_k(f) = \{x \in F \,|\, f(x) \geq k\}. \tag{3.19}$$

Therefore, Level Sets correspond to binarizations by threshold k. Given the Level Sets, we can reconstruct the original image through the supremum operation:

$$f(x) = \sup_{k=0..n-1} \{x \in S_k(f)\}. \tag{3.20}$$

The gray scale dilation and erosion may be defined in terms of the binary dilation and erosion applied to the Level Sets. That is:

$$(f \oplus g)(x) = \sup_{k=0..n-1} \{x \in [S_k(f) \oplus S_k(g)]\}, \tag{3.21}$$

$$(f \ominus g)(x) = \sup_{k=0..n-1} \{x \in [S_k(f) \ominus S_k(g)]\}. \tag{3.22}$$

The definition that we propose for the gray scale HMT starts with the decomposition of the image f and the structural function g into the binary images defined by their Level Sets. Afterwards we apply the binary HMT at each level set as follows: the k-th level set of the structural function g is used as the structural element of the HMT applied to the k -th level set of the image f. The gray scale Hit-or-Miss Transform can then be expressed as the reconstruction of an image whose level sets are given by these binary HMT. Therefore, the result of the gray scale HMT can be obtained by reconstruction :

Definition 7 *The Level Set gray scale Hit-or-Miss Transform is defined as*

$$\left(f\bar{\otimes}g\right)(x)=\sup_{k=0..n-1}\left(x\in\left[S_k\left(f\right)\otimes S_k\left(g\right)\right]\right), \tag{3.23}$$

where $S_k\left(g\right)$ is the two element partition formed by the foreground (1 pixels) and the background (0 pixels).

We call this operator Level Set Hit-or-Miss Transform (LSHMT), and we denote it $\bar{\otimes}$, to differentiate it from the GHMT described previously. When the searched pattern appears reproduced in the signal, the response is the pattern value at the origin.

$$f\left(y-x\right)=g\left(x\right),\ \ x\in G\Longrightarrow\left(f\bar{\otimes}g\right)\left(y\right)=g\left(0\right)=f\left(y\right). \tag{3.24}$$

It is interesting to note that although the conventional definitions of function erosion and dilation hold, the duality of erosion and dilation in the binary level sets induces an expression of the duality of function erosion and dilation of the form:

$$f\oplus g=n-\left(\left(n-f\right)\ominus\breve{g}\right) \tag{3.25}$$

which is equivalent to the conventional expression of the duality of function erosion and dilation:

$$f\oplus g=-\left(\left(-f\right)\ominus\breve{g}\right) \tag{3.26}$$

where $\breve{g}$ is the functional symmetric of g, $\breve{g}\left(x\right)=g\left(-x\right)$. An undesirable property of the LSMHMT is that constant shifts of the patterns are not detected:

$$f\left(x\right)=g\left(x\right)+k\Longrightarrow\left(f\bar{\otimes}g\right)\left(x\right)=0 \tag{3.27}$$

contrary to the behavior of the GHMT. This handicap could be alleviated taking into account that constant shifts produce at the bottom level sets binary all 1 images that can be discarded, but we will not pursue this here. The results in the experiments induce us to believe that this property is not so important for robust recognition.

It is clear that $\left(f\bar{\otimes}g\right)\left(x\right)\leq\min\left\{f\left(x\right),g\left(x\right)\right\}$. Therefore, the effective matching distance is given by the difference between the image and the results of the LSHMT:

$$d_{LSHMT}\left(f,g\right)\left(x\right)=f\left(x\right)-\left(f\bar{\otimes}g\right)\left(x\right) \tag{3.28}$$

This is the distance that will be used in the face localization experiments to decide is a face is found at location x.

It is to see that the localization of a level set of the structural function in the image does not imply that all preceding level sets have been localized as well, formally:,

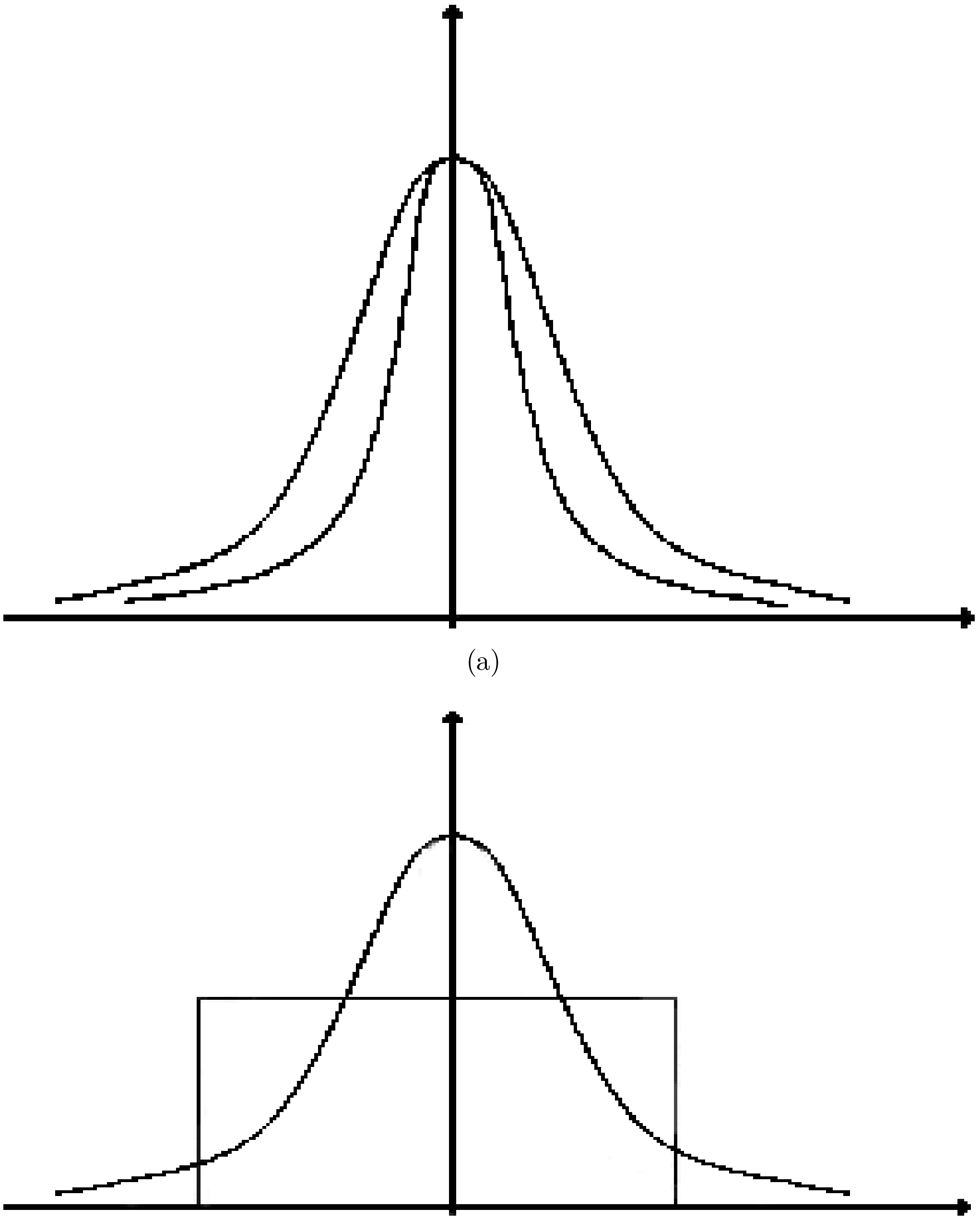
$$\left(S_k\left(f\right)\otimes S_k\left(g\right)\right)\left(x\right)=1\nRightarrow\left(S_{k-1}\left(f\right)\otimes S_{k-1}\left(g\right)\right)\left(x\right)=1 \tag{3.29}$$

Therefore, the LSHMT does not produce a proper structure of level sets. However, our definition does not depend on the continuity of the response over the level sets and has the following interpretation: LSHMT searches for the highest level at which a matching occurs between the two functions. In figure 3-1 we show two cases of the matching of two functions. In the top image, all the level sets of one of the gaussian functions are included in the level sets of the other. The only level set that gives a non-null Hit-or-Miss localization set is the their common peak. Therefore, from the point of view of LSHMT, the matching is perfect. The bottom image shows the match between a Gaussian function and a square function. In this case, the only level set that give non-null Hit-or-Miss localization is the lower intersection between the two curves.

3.5 Experimental Results

We have performed the experimental comparison of GHMT and LSHMT over a set of 19 images. The images contain 2 or 3 persons situated in front of a quite homogenous background, showing large parts of the body and the background. The images are of size 640x480 pixels and represent frontal shots with slight rotations of the head. Persons showing additional facial features (such as eye-glasses or beards) and different facial expressions were also allowed to be part of the experiments. No restrictions about the clothes or the illumination conditions were imposed.

The average face is used as the template for GHMT and LSHMT. The images used for



(a)

(b)

Figure 3-1: Two instances of 1D signals matched by the LSHMT: (a) detection is the top point and (b) detection is the lowest level at which the curves intersect

the face extraction are considered the training set, and the remaining as the test set. The ground truth is given by a set of hand defined rectangles that include most of the faces. These rectangles do not coincide with the face patterns, so that finding the original patterns does not imply 100% recognition of face pixels.

Let $g : G \rightarrow \{0, 1, ..., n-1\}$ be the face pattern (the average face) and $f : F \rightarrow \{0, 1, ..., n-1\}$ the image, the face detection process consists of the computation of the matching distance $d_i (f, g) (x)$, $i = GHMT, LSHMT$ as defined above and the application of a decision threshold to the computed distance. That is, the set of locations of the faces in the image is:

$$L_i (\theta) = \{x \, | d_i (f, g) (x) < \theta\}, i = GHMT, LSHMT. \tag{3.30}$$

Once detected the face locations, the set of detected face pixels in the image is constructed by placing a square of the same size as the face patterns centered at each of the positions in $L_i (\theta)$. The positive detections are the pixels classified as face pixels, negative detections are the pixels not classified as face pixels. *True positive* are the positive detections that correspond to face pixels in the given ground truth. *False positive* are the positive detections that do not correspond to face pixels in the ground truth. *True* and *false negative* are the negative detections that correspond, respectively to no face and face pixels in the ground truth.

We computed the normalized ratios:

$$TP (\theta) = \frac{\# (true\, positive)}{\# (\text{positive } ground\, truth)} \tag{3.31}$$

and

$$FP (\theta) = \frac{\# (false\, positive)}{\# (negative\, ground\, truth)} \tag{3.32}$$

where $\# (A)$ denotes the cardinality of the set A. Both ratios $TP (\theta)$ and $FP (\theta)$ are monotonically increasing functions of θ. The $TP (\theta)$ ratio approaches 1 when all the faces in the image are detected, and the $FP (\theta)$ ratio approaches 1 when all the detected faces correspond to false positives. The plot of $TP (\theta)$ versus $FP (\theta)$ is the so-called Receiving Operator Characteristics (ROC) (the ratio of the true positive pixels to false positive pixels). In figure 3-2 we show the average ROC over the sets of training and test experimental images. This ROC can be used to define the decision threshold and also to compare the behavior of the GHMT and LSHMT operators as face locators. A measure of the optimality of a given detection algorithm is the distance from the point $(1, 0)$ in the (FP, TP) space to its nearest point in the ROC curve. This point indicated the optimal tradeof between false alarms and positive detections that can be achieved by the algorithm.

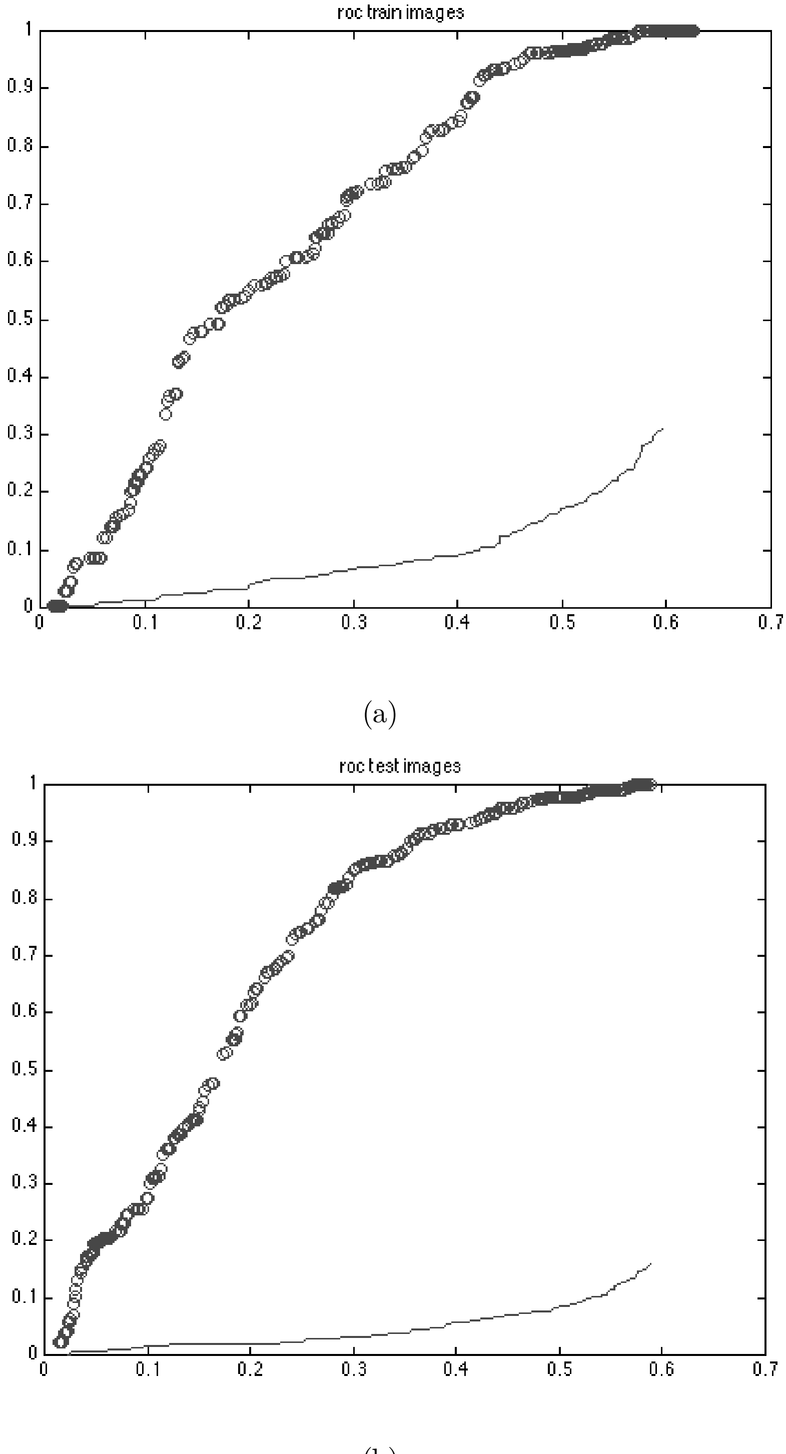
It can be appreciated, that the GHMT performs much worse than the LSHMT. This result is quite surprising, and the magnitude of the difference is quite remarkable. The average ROC of GHMT is very low due to its unability to locate the faces in some of the images. In the figures 3-3 and 3-4 we reproduced some of the face localization instances for the training and test images, respectively. To produce these figures, the decision threshold was set in order to obtain a TP ratio of 0.9.

This result shows the poor ability of the GHMT to cope with small distortions and perturbations of the pattern searched. In preliminary experiments we have verified the perfect matching ability of GHMT. It has a good ability to find exactly the same pattern in the source image, and poor discrimination for close patterns. Our proposition LSHMT has the same localization ability for exact patterns, but it has an enhanced robustness to small distortions and lightning conditions. The case at hand shows this robustness. The average face is a small distortion of each of the face patterns to be found in the images.

3.6 Conclusions

In this chapter we have presented an approach to face localization based on a primitive morphological operator: the Hit-or-Miss Transform. We have reviewed its application as a shape recognition tool in binary images and the proposition by Kanerva and Shaffer of GHMT. We have proposed LSHMT a gray scale HMT defined by the reconstruction of the image after performing the binary HMT with the level sets of the image and the structuring function. Although the experiments show the power of the definition of LSHMT as an object localization tool, we think that there is still much potential for improvement.

The application of BHMT to the level sets combined with the notions of morphological scale-spaces that we will present in the next chapter is an open research track. A related work



(a)

(b)

Figure 3-2: ROC of the ratio of true positive pixels versus the false positive pixels, using the average face pattern: GHMT ('.') and LSHMT ('o') for (a) train images and (b) test images, respectively

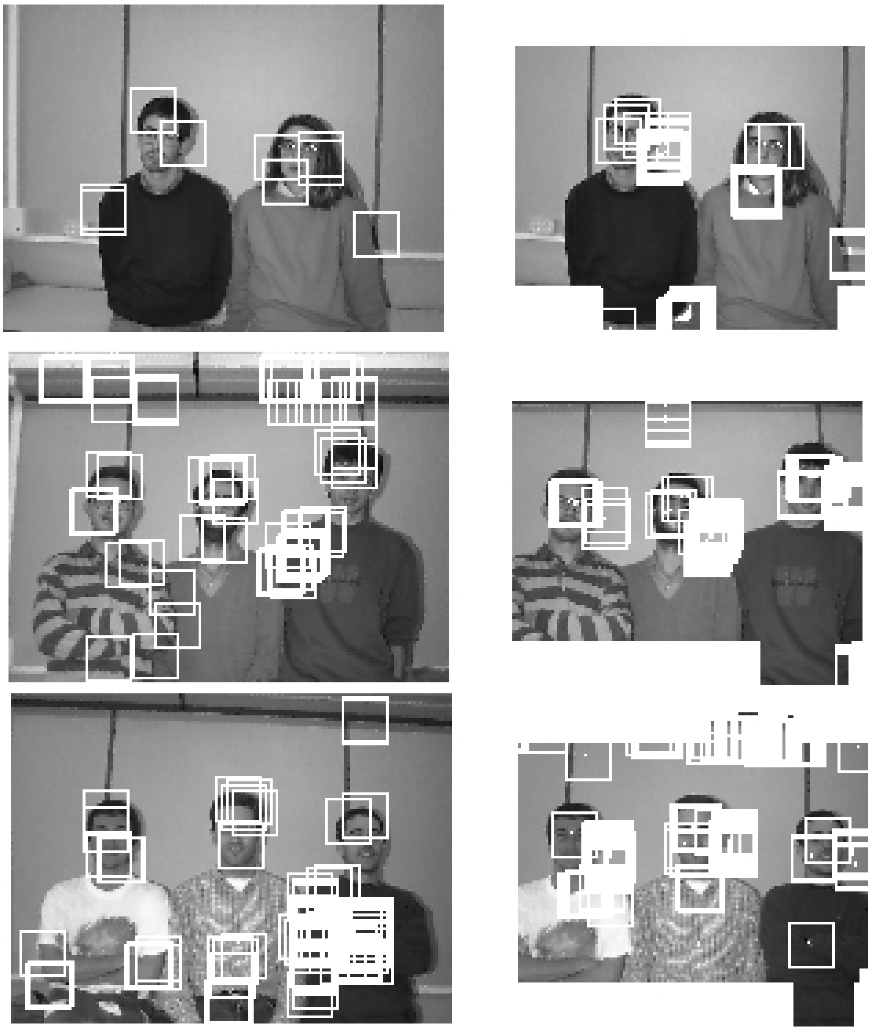


Figure 3-3: Face localization on some training images with the constraint of 90% localization of face pixels: LSMHT (left) and GHMT (right). Training images are the ones that contrbutre face patterns to compute the average face.

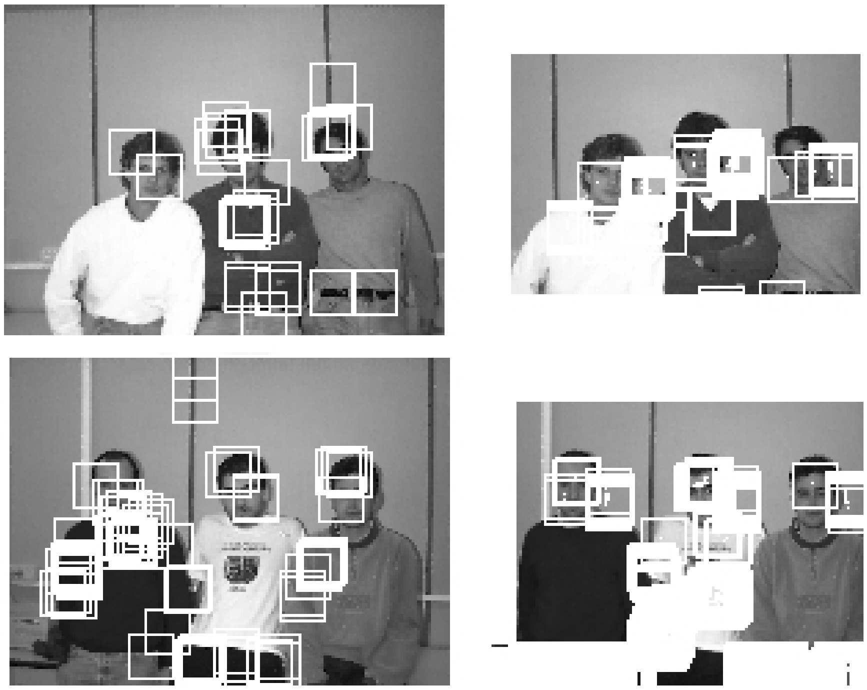


Figure 3-4: Face localization on some test images with the constraint of 90% localization of face pixels: LBHMT (left) and GHMT (right)

is [119]. Patterns are vectors with component values in the interval $[0,1]$. The fuzzy ART neural architecture is interpreted as realizing a kind of HMT whose adaptation comes from the learning of the ART architecture. If possible it would be of interest also to learn the structural functions for the gray scale HMT.

Chapter 4

Face Localization based on Morphological Multiscale Fingerprints

This chapter presents the attempt to apply multiscale morphological features, so-called fingerprints, to characterize faces and then localize them in gray scale images. Foundations for this effort are two sided. On one hand, the Scale-Space approach to image segmentation and its morphological version that define the Multiscale Morphological Fingerprints (MMF) as the image features. On the other hand, the works on graph matching provide the framework for the recognition task based on the features extracted from the image. After the revision of those foundations, we report the results of experiments on face localization.

4.1 An Overview of Scale-Spaces

Nowadays, the scale-space theory is widely accepted as a formalism that plays a major role in computer vision, because of the necessity to specify explicitly the scale visual associated with the objects under observation, and, thus, the appropriate scales to perform the observations and obtain measures. Scale-space theory deals with the formal definition of the concept *scale* σ in terms of signals/images, i.e. how we represent the data at a given scale and how we relate image features from one scale to another. In such a theory, a family of images is built up by blurring the original image depending on a scale parameter σ.

Weickart [137] has rediscovered the first references to linear Gaussian scale. He found that the concept dates back to the 60's and was first invented by Iijima. For some reason, at that time, the idea didn't capture the interest of the scientific community. The notion was popularized later on by Witkin [138]. In his paper, Witkin states that the Gaussian convolution is the unique operator that satisfies general principles of spatial symmetry and scale invariance. Koenderink [71] was first to show that these symmetry and invariance principles are compatible with a causality principle requiring that new details cannot be formed when moving from finer to coarser levels, i.e. every feature in coarse scale (large σ) has to have a cause in fine scale (small σ). He also showed that Gaussian filtering is the Green's function [37] of the differential equation known as the "diffusion equation". Formally: Let us consider a function $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$, and a Gaussian function $g(x,\sigma) : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}$ as the smoothing kernel. Then, the signal smoothed at scale σ, $F : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}$,

$$F(x,\sigma) = \int f(\xi)\, g(x-\xi,\sigma)\, d\xi, \tag{4.1}$$

is a solution of the diffusion equation:

$$\bigtriangledown^2 F = -k\frac{\partial F}{\partial \sigma}, \tag{4.2}$$

taking the image $f(x)$ as the initial condition, i.e. $F(x,0) = f(x)$.

A multiscale analysis generates, from a single initial picture $f(x)$, a sequence of simplified pictures $F(x,\sigma)$, where $F(x,\sigma)$ appears to be a rougher version of $f(x)$ as σ increases. In $F(x,\sigma)$ the details and features like edges and extreme points are kept if their scale exceeds σ. If we denote $S_\sigma(f)$ the set of features (edges, extrema, etc) at scale σ, extracted from $F(x,\sigma)$, the basic properties which any multiscale analysis must satisfy are:

1. Fidelity: $S_\sigma(f) \rightarrow S_0(f)$ as $\sigma \rightarrow 0$.

2. Causality: $S_\sigma(f)$ only depends of (is included in) $S_{\sigma'}(f)$ if $\sigma > \sigma'$.

3. Euclidean invariance: If A is an isometric transformation, then $S_\sigma(f \circ A) = S_\sigma(f) \circ A$.

4. Strong causality: $S_\sigma(f) \subset S_{\sigma'}(f)$ if $\sigma > \sigma'$.

Lindeberg [86] was the first to consider the discrete equivalent of the Gaussian linear scale-space. Instead of specifying a scale-space operator in the continuous domain and then discretizing the continuous operator, Lindeberg *discretized* the scale-space requirements. Fortunately, only for very small scales the two approaches differ significantly.

In mathematical morphology the notion of *scale dependent observations* was introduced for the first time by Matheron [92] in his study on granulometries. The aim of his study was to capture the size distributions of spatial observations. The non-linear scale-dependent operators, dilation and erosion, were later on known as *morphological scale-space operators*. Morphological operators can remove structure from a signal, and therefore they are suitable as scale-space smoothers. An early approach to the definition of morphological scale spaces is given by Chen and Yan [25] have used a scaled disk for the morphological opening of objects in binary images to create a scale-space theorem for zero-crossings of object boundary curvature. Afterwards, these results were extended to general compact and convex structuring elements [64].

On the other hand, Brockett and Maragos [16] were the first to show that *flat* morphological operators can be described in terms of (nonlinear) PDE's. Jackway [62] and van den Boomgaard [13] independently showed that a morphological analogue of the Gaussian linear scale-space does exist: the parabolic erosions and dilations. The same set of basic principles is shown to lead to both the linear Gaussian scale-space and the morphological parabolic scale-space. Van den Boomgaard and Smeulders [14] also showed that the morphological parabolic scale-space can be viewed upon as the solution of a (nonlinear) PDE, just like the linear scale-space is the diffusion equation, described in equation 4.2.

Alvarez and Morel [2] and later on Heijmans and Boomgaard [58] presented an excellent theoretical unification and axiomatization of many multiscale image analysis theories including most of those mentioned above. A more detailed, formal definition of the scale-space together with two particularizations can be found in appendix A.

A redefinition of the basic principles of scale-spaces in a way that is more appropriate for morphological scale spaces, is presented by Boomgaard [15] Let us denote the image by f and the scale-space operator by Ψ^{σ}. Then, F denotes the scale-space, i.e. $F(x, \sigma) = (\Psi^{\sigma} f)(x)$

- **Superposition principle**: Given two images f and g, the linear superposition principle states that: $\Psi^\sigma(\alpha f + \beta g) = \alpha\Psi^\sigma f + \beta\Psi^\sigma g$. In the morphological case, the superposition principle applies to the MAX operator, i.e. $\Psi^\sigma(MAX((\alpha + f), (\beta + g))) = MAX((\alpha + \Psi^\sigma f), (\beta + \Psi^\sigma g))$. A similar relation is valid also for the MIN operator.

- **Translation invariance**: $\Psi^\sigma f_t = (\Psi^\sigma f)_t$, where by f_t we denote the translation of f by t.

- **Rotational invariance**: is obtained with rotational symmetric convolution kernels and rotational symmetric structuring functions.

- **Differential scale** states that $F(x,\ \sigma + d\sigma)$ can be computed from the image scale σ, using only the values taken in a infinitesimal neighborhood of x.

- **Scale invariance** states that any linear scaling of the original image f (both spatially and in intensity), should leave the scale-space unchanged (up to an arbitrary reparametrization of the scale-parameter).

The scale-space causality allows a hierarchical approach to the image analysis. Image features are explored at the coarsest level and the analysis is progressively refined to the finest detail. This means that information can be processed in an order related to its physical relevance. We can start the computational processes making abstraction of low scale information. Low scale information is used afterwards to refine the results. The scale-space operator invariance means that the same results would be obtained regardless of transformations that do not introduce structural deformations..

4.2 Morphological Scale-Space

We follow in this section the work of JAckway [63]. A morphological scale-space can be defined in terms of multiscale erosions and dilations. Let us start recalling the definition of gray scale dilation and erosion.

Definition 8 *Gray scale Dilation: Let us assume that f is the original gray scale image and g is the structuring function, namely $f : D \subset \mathbb{R}^n \to \mathbb{R}$ and $g : E \subset \mathbb{R}^n \to \mathbb{R}$. Then the gray scale* *dilation of f by g is given by:*

$$(f \oplus g)(x) = \sup_{t \in E} \{f(x-t) + g(t)\} \tag{4.3}$$

Definition 9 *Gray scale Erosion: Let us assume that f is the original gray scale image and g is the structuring function, namely $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ and $g : E \subset \mathbb{R}^n \rightarrow \mathbb{R}$. Then the gray scale erosion of f by g is given by:*

$$(f \ominus g)(x) = \inf_{t \in E} \{f(x+t) - g(t)\} \tag{4.4}$$

The result of dilation or erosion depends on the position of the *origin* of the structuring function. In order to avoid level-shifting effects the following restrictions on the structuring function must be imposed [63]: the structuring function must be non-positive and its value at origin is zero

$$\begin{aligned} \sup_{t \in E} \{g(t)\} &= 0 \\ g(0) &= 0. \end{aligned} \tag{4.5}$$

In the construction of a morphological scale-space the notion of scale lies in definition of the structuring function. That means that g is scale-dependent, i.e $g_\sigma : E_\sigma \subset \mathbb{R}^n \rightarrow \mathbb{R}$ is of the form:

$$g_\sigma(x) = |\sigma| \, g\left(|\sigma|^{-1} x\right), \qquad x \in E_\sigma, \ \textit{and} \ \sigma \neq 0 \tag{4.6}$$

A suitable scale-space structuring function is the *sphere* function defined by the following equation:

$$g_\sigma(x) = |\sigma| \, ((1 - \|x/\sigma\|^2)^{1/2} - 1), \quad \|x\| \leqslant \sigma \tag{4.7}$$

To ensure that the erosion and dilation operators induce a scale space that possess the basic properties discussed in the preceding section, further restrictions must be imposed on the structuring function. Let us consider the level sets (threshold sets) of the structuring function

$$\xi_\sigma(t) = \{x \,|\, g_\sigma(x) \geq t\} \tag{4.8}$$

for any $t < 0$. To make use of the scaling notion, we have to ensure teh following conditions for all $t < 0$:

$$|\sigma| \to 0 \Longrightarrow \xi_\sigma(t) \to 0 \tag{4.9}$$

$$|\sigma_1| < |\sigma_2| \Longrightarrow \xi_{\sigma_1}(t) \subset \xi_{\sigma_2}(t) \tag{4.10}$$

$$|\sigma| \to \infty \Longrightarrow \xi_\sigma(t) \supset \{x \mid \ \|x\| \leq R\}, \forall R > 0 \tag{4.11}$$

Condition (4.9) enforces the fidelity property of mulstscale analysis. The structuring object becomes null as the scale goes to zero. Conditions (4.10) and (4.11) together with the increasing property of dilation and erosion operators enforce the causality property of of mulstscale analysis. The result of filtering at higher scales is included in the result of filtering at lower scales. Reconstruction from the level sets induces the equivalent restrictions in the structural functions:

$$|\sigma| \to 0 \Longrightarrow g_\sigma(x) = \begin{cases} 0, & if\ x = 0 \\ -\infty, & if\ x \neq 0 \end{cases} \tag{4.12}$$

$$|\sigma_1| < |\sigma_2| \Longrightarrow g_{\sigma_1}(x) \leq g_{\sigma_2}(x), \ x \in E_{\sigma_1} \tag{4.13}$$

$$|\sigma| \to \infty \Longrightarrow g_\sigma(x) \to 0, \forall x \tag{4.14}$$

Based on structuring functions taht comply with this restrictions, [63] introduces the *Morphological Scale-Space* induced by multiscale dilation-erosion. The scale is considered as a continuum real value. Positive values are interpreted as dilations, negative values are interpreted as erosions.

Definition 10 *Multiscale Dilation-Erosion Scale-Space: With f and g defined as above, the multiscale dilation-erosion scale-space is given by:*

$$f \circledast g_\sigma = \begin{cases} (f \oplus g_\sigma)(x), & if\ \sigma > 0 \\ f(x), & if\ \sigma = 0 \\ (f \ominus g_\sigma)(x), & if\ \sigma < 0 \end{cases} \tag{4.15}$$

For positive scales ($\sigma > 0$), multiscale dilation-erosion corresponds to a gray scale dilation, and for negative scales ($\sigma < 0$), the operation corresponds to a gray scale erosion. As $|\sigma|$

increases, the operator tends to filter out more features producing an increasingly coarse version of the image. The fidelity property of the scale space follows inmediately from the definition. When $|\sigma| \to 0$, the operator converges to the identity operator.

Since the structuring function is non-negative at the origin (we imposed before the condition $g(0) = 0$), the dilation is extensive and the erosion is anti-extensive, i.e. $(f \oplus g) \geq f$ and $(f \ominus g) \leq f$. Thus, we have the result:

$$(f \circledast g_\sigma)_{\sigma<0} \leq f \leq (f \circledast g_\sigma)_{\sigma>0}\,. \tag{4.16}$$

For the remainder of this chapter we will refer to the morphological scale-space as $F : D \subset \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$, where F is given by:

$$F(x,\sigma) = (f \circledast g_\sigma)(x) \tag{4.17}$$

Image features of interest in the Morphological Erosion-Dilation Scale-Space are the extreme points, [63] specializes the formulation of the scale-space properties in terms of these discrete features. First introduces the *continuity property.*

Definition 11 *Scale-space continuity property: An image (signal) feature once presented at a certain scale, it will appear downward, all the way to zero-scale (the original image).*

Scale-space continuity implies that no spurious features appear because of the filtering operator that induces the scale-space. It is a stronger condition than the fidelity property.

Definition 12 *Scale-space monotonic property: The number of features decreases as scale increases:*

$$\#[F(x,\sigma_1)] \leq \#[F(x,\sigma_2)], \qquad \textit{with } \sigma_1 > \sigma_2 > 0$$

where by $\#[F(x,\sigma)]$ we denote the number of features at a given scale.

The monotonic property is concerned with the amount of information that we extract at each scale. It states that this amount decreases monotonically as the scale increases.

Definition 13 *Scale-space causality property: No new signal features can appear at a non-zero scale. If we consider the set of the features in the range $[\sigma_1, \sigma_2]$ as $C = \{(x,\sigma) : x \in \mathbb{R}^n$ and $\sigma_1 \leq \sigma \leq \sigma_2$*

with $0 \leq \sigma_1 \leq \sigma_2$ *then* $C \cap \{(x, \sigma) : \sigma = \sigma_1\} \neq \emptyset$

In other words, the causality property says that a feature at a non-zero scale should have the *cause* at a lower scale, i.e. it *must* appear at all scales downward to zero-scale.

From the pattern recognition point of view, as opposed to the filtering point of view, the scale-space continuity and causality properties, imply that the filtering process that generates the scale-space does not introduce spurious features. The monotonic property implies that the scale allows us to select the amount of information that we will use for classification. Together, the three properties imply that the scale parameter allows to control the quantity and quality of the information used for classification and recognition.

4.3 Morphological Multiscale Fingerprints

The features of interest in the scale-space generated by the multiscale morphological operators are the local extrema of the image intensity. The following proposition relates the position and amplitude of a local maximum (or minimum) in the filtered signal to that in the original one, they are particularizations of the continuity and causality properties of general scale-spaces when the features of interest are the local extrema. We state them without proof..All the proofs of these propositions and theorem can be found in [63].

Proposition 14 *(Continuity) Let us consider that the structuring function have a single maximum at the origin, i.e.* $g(x)$ *is a local maximum implies* $x = 0$ *then:*

1. *If* $\sigma > 0$ *and* $F(x_{\max}, \sigma)$ *is a local maximum then* $f(x_{\max})$ *is a local maximum of* $f(x)$ *and* $F(x_{\max}, \sigma) = f(x_{\max})$;
2. *If* $\sigma < 0$ *and* $F(x_{\min}, \sigma)$ *is a local minimum then* $f(x_{\min})$ *is a local minimum of* $f(x)$ *and* $F(x_{\min}, \sigma) = f(x_{\min})$.

Proposition 15 *(Causality) If we assume again that the structuring function has a single local maximum at the origin, as in the previous proposition, then:*

1. *If* $\sigma_2 > \sigma_1 > 0$ *and* $F(x_{\max}, \sigma_2)$ *is a local maximum then* $F(x_{\max}, \sigma_1)$ *is a local maximum and* $F(x_{\max}, \sigma_1) = F(x_{\max}, \sigma_2)$*;*

2. *If* $\sigma_2 < \sigma_1 < 0$ *and* $F\left(x_{\min}, \sigma_2\right)$ *is a local minimum then* $F\left(x_{\min}, \sigma_1\right)$ *is a local minimum and* $F\left(x_{\min}, \sigma_1\right) = F\left(x_{\min}, \sigma_2\right)$

Let us define the following point sets: $E_{\max} = \{x : f(x)$ is a local maximum$\}$ and $E_{\min} = \{x : f(x)$ is a local minimum$\}$. The fidelity and causality properties of scale spaces are realized in terms of these point sets as it is formalized in the following theorem:

Theorem 16 *Scale-Space Monotonic Property: For any scale sequence* $\sigma_1 < \sigma_2 < 0 < \sigma_3 < \sigma_4$, *the following relations hold:*

$$E_{\min}(f \circledast g_{\sigma_1}) \subseteq E_{\min}(f \circledast g_{\sigma_2}) \subseteq E_{\min}(f) \tag{4.18}$$

and

$$E_{\max}(f \circledast g_{\sigma_4}) \subseteq E_{\max}(f \circledast g_{\sigma_3}) \subseteq E_{\max}(f) \tag{4.19}$$

The *fingerprints* of a scale-space are plots, over scale, of the point sets of signal features:

Definition 17 *Morphological Multiscale Fingerprints:* $E^*(\sigma) = E_{\max}(f \circledast g_\sigma) \cup E_{\min}.(f \circledast g_\sigma)$.

For computational reasons, in practice the *reduced fingerprints* are used instead of the complete fingerprints. The reduced fingerprints correspond to the local maxima of the dilated images in the space scale (positive scales) and the local minima of the eroded images in the scale space (negative scales). The fingerprints of the original image are all the local extrema.

Definition 18 *Reduced Fingerprints:*

$$E_r^*(\sigma) = \begin{cases} E_{\max}(f \oplus g), & if\ \sigma > 0 \\ E_{\max}(f) \cup E_{\min}(f), & if\ \sigma = 0 \\ E_{\min}(f \ominus g), & if\ \sigma < 0 \end{cases} \tag{4.20}$$

We will finish this review about morphological scale-space, just mentioning that in a similar way another morphological scale-space can be built using the operations of closing/opening, as follows:

$$(f \odot g_\sigma)(x) = \begin{cases} (f \bullet g_\sigma)(x), & if\ \sigma > 0 \\ f(x), & if\ \sigma = 0 \\ (f \circ g_\sigma)(x), & if\ \sigma < 0 \end{cases} \tag{4.21}$$

where by $\bullet$ is the *closing* operator and $\circ$ the *opening* operator. We will not insist on this scale-space, as in our experiments we used the morphological scale-space induced by dilation/erosion.

4.4 Morphological Fingerprints and Graph Matching

The attributed relational graphs (ARG) have been used to characterize objects in computer vision applications. The problem of object identification or recognition based on its graph representation becomes the problem of computing the distance between two graphs. A related approach is that of elastic networks [76]. A classic work is the distance defined in [42]. There, the distance is computed as the cost of the transformations suffered by two graphs until they become identical. To compute it, they build-up a lattice of transformations. From empty graphs, up to the best approximation. Distance is computed as the optimal path over this lattice (dynamic programming). Other recent works are [66], [81], [51], [103] and [140]. In our case, recognizing objects based on morphological fingerprints, graphs are defined by fingerprints and their spatial relations. Therefore, object recognition on the basis of morphological fingerprints, is an instance of object recognition based on ARG.

Following the formalization in in [42] that an attributed relational graph is defined in the following way:

Definition 19 *An ARG is a tuple of the form:*

$$G = (N, B, A, E, G_N, G_B) \tag{4.22}$$

where

- N $\left(N = \left\{n_1, .., n_{|N|}\right\}\right)$ *is the set of nodes,*
- B $\left(B = \left\{b_1, .., b_{|B|}\right\}\right)$ *is a set of ordered pairs that represents the arcs in the graph (oriented)* $b = (n_i, n_j)$.

- *A is the alphabet of nodes' attributes.*
- *E is an alphabet of edges' attributes.*
- *G_N is a function or set of functions that generates the nodes' attributes.*
- *G_B is a function or set of functions that generates edges' attributes.*

The description of any object based on the morphological fingerprints of the morphological scale-space analysis of its representative images can be put in the framework of ARG. The set of nodes N is the set of local extrema points in the morphological fingerprint. The alphabet of node attributes are the image position and scale of the local extrema.Therefore:

$$\begin{aligned} A &= \mathbb{Z}^2 \times \mathbb{Z}, \\ a &= \left(\left(x, y\right), \sigma\right) \end{aligned} \tag{4.23}$$

The morphological scale-space fingerprints do not give any direct interpretation of the graph edges B, besides the topological relations of the nodes in the image domain. These relations are implicit in the label of the nodes. The graph could be considered as a complete graph with edges being labelled by the vector between nodes.

$$\begin{aligned} B &= \mathbb{R} \times \left[0, 2\pi\right], \\ b_{ij} &= \left(d_{ij}, \theta_{ij}\right) \end{aligned} \tag{4.24}$$

Also, another approach could be the representation of the topological relation between nodes using a smaller, less than complete, set of edges which may be given, for instance, by the Delaunay triangulation. We have tested this latter approach and found that this structure is rather unstable and suffers great changes with small pose changes of the face. So we do not recommend it. In the experiments described below, we did not use any edge description.

4.5 Graph Matching

Given two graphs of fingerprints, obtained from two different images, G^M and G^T, that represent, respectively, the model and the test, and given a correspondence between the two sets of nodes:

$$\begin{aligned} C_N &\subseteq N^M \times N^T, \\ c_N(k) &= l \Rightarrow \left(n_k^M, n_l^T\right) \in C_N, \end{aligned} \tag{4.25}$$

The correspondence between nodes induces a correspondence between graph edges:.

$$\begin{aligned} C_B &\subseteq B^M \times B^T, \\ c_B(ij) &= kl \Leftrightarrow b_{ij}^M \in B^M \& c_N(i) = k \& c_N(j) = l \& b_{kl}^T \in B^T. \end{aligned} \tag{4.26}$$

We can define the distance between graphs as the measure of the matching degree obtained.. If the graphs are isomorphic and the correspondence between nodes produce an isomorphism, the distance computed should be 0. The graphs are isomorphic when the correspondence between nodes and edges are biunivoque and the diferences between labels are 0. Obviously, this situation is an ideal one and only we can get approximative isomorphisms between subgraphs. The distance between graphs measures the goodness of the approximation to the ideal isomorphism. It can be decomposed into two components:

The first component of the distance between graphs of fingerprints, given a correspondence between graph nodes, is proportional to the degree of matching between nodes and arcs.

$$D_1\left(G^M, G^T\right) = \gamma_1 |S_1| + \gamma_2 |S_2| + \gamma_3 |S_3| + \gamma_4 |S_4| + \gamma_5 |S_5| + \gamma_6 |S_6| \tag{4.27}$$

where the γ_i are weights and S_i are sets that contain (1) the model nodes with no correspondence matching in the test. (2) idem in the test, (3) the model nodes with more than one correspondence in the test, (4) idem in the test, (5) the arcs in the model without correspondence and (6) idem in the test. Formally,

$$S_1 = \left\{ n^M \left| \nexists \left(n^M, n^T\right) \in C_N \right. \right\}, \tag{4.28}$$

$$\begin{aligned}
S_2 &= \left\{n^T \left| \nexists \left(n^M, n^T\right) \in C_N \right.\right\}, \\
S_3 &= \left\{n^M \left| \left|\left\{\left(n^M, n^T\right) \in C_N\right\}\right| > 1 \right.\right\}, \\
S_4 &= \left\{n^T \left| \left|\left\{\left(n^M, n^T\right) \in C_N\right\}\right| > 1 \right.\right\}, \\
S_5 &= \left\{b^M \left| \nexists \left(b^M, b^T\right) \in C_B \right.\right\}, \\
S_6 &= \left\{b^T \left| \nexists \left(b^M, b^T\right) \in C_B \right.\right\}.
\end{aligned}$$

The weights associated with these sets are defined heuristically.

The second component measures the goodness of matching on the basis of node attributes, node positions, nodes scales and arcs labels.

$$\begin{aligned}
D_2\left(G^M, G^T\right) = \; & \gamma_7 \sum_{(k,l)\in C_N} \left\|(x,y)_k^M - (x,y)_l^T\right\| + \\
& \gamma_8 \sum_{(k,l)\in C_N} \left|\sigma_k^M - \sigma_l^T\right| + \\
& \gamma_9 \sum_{((i,j),(k,l))\in C_B} \left(\left|d_{ij}^M - d_{kl}^T\right| + \left|\theta_{ij}^M - \theta_{kl}^T\right|\right)
\end{aligned} \tag{4.29}$$

The distance between the two graphs will be a combination of these distances:

$$D\left(G^M, G^T\right) = D_1\left(G^M, G^T\right) + D_2.\left(G^M, G^T\right) \tag{4.30}$$

Given the definition of the distance between graphs, matching two graphs G^M and G^T can be defined as the optimal isomorphism in the sense of minimizing distance $D\left(G^M, G^T\right)$. That is, to determine the distance between the graphs it is necessary to compute the distance for all the potential correspondences between node sets. The minimum corresponds to the best partial isomorphisms and the distance between the graphs. This is a combinatorial problem that can be solved via exhaustive search, stochastic search (i.e.: Simulated Annealing, Genetic Algorithms) or assuming a deterministic suboptimal but fast approximation. The last approach is the one taken in our work.

4.6 Experimental Results

As an initial evaluation experiment, we have tested the approach face localization based on the Principal Component Analysis (PCA) and the distance to feature space, as described in section 2.2, and the approach based on graph matching of the Morphological Multiscale Fingerprints over a small custom set of 19 images (see appendix B). We have extracted 40 sample faces, and exactly the same faces have been used to compute both the eigenfaces of the PCA transformation and the MMF prototypes. Figure 4-1 shows our custom face database and figure 4-2 shows its reduced fingerprints. The figure 4-3 shows the eigenfaces computed from the face patterns, with its intensity scaled for visualization. The images used for the face extraction are considered the training set, and the remaining as the test set. The ground truth is given by a set of hand-defined rectangles that include most of the faces. The face ground truth and the face patterns do not coincide, as can be verified comparing the faces labeled in figure 4-6 (first column) and the face patterns in figure 4-1. That means that the perfect localization of the face patterns does not imply 100% localization of target faces, even in the so-called training images. This accounts for some of the results that show a high number of false positives even for training images.

Let $\{g_k : G \rightarrow \{0,1,...,n-1\}\}$ be a set of face patterns and $f : F \rightarrow \{0,1,...,n-1\}$ the test image. The face detection process consists of the computation of the matching distance $d_i\left(f,\{g_k\}\right)(x)$, $i = PCA, MMF$ and the application of a decision threshold to this distance. That is, the set of location of the face is $L_i(\theta) = \{x \,|d_i\left(f,\{g_k\}\right)(x) < \theta\}$, $i = PCA, MMF.$

The distance $d_{PCA}\left(f,\{g_k\}\right)(x)$ is the distance to feature space discussed in section 2.2. It is the error between the actual image and the one recovered after the projection into the subspace spawned by the eigenfaces computed from the set of patterns $\{g_k\}$. The distance $d_{MMF}\left(f,\{g_k\}\right)(x)$ is a simplified version of the distance between graphs (4.30). The set of fingerprints at scales $.\left\{\pm\sigma = 2^j\,,\ j = 1..6\right\}$ has been computed. These fingerprints are nodes of the graph G^T and G^M for the image and each of the face patterns, respectively. Neither arcs nor spatial position of the fingerprints are considered. Therefore, $D\left(G^M, G^T\right)$ is reduced, for the experiments reported here, to computing the difference between the sets of fingerprints at each scale of the face pattern and the image. When there are several test patterns, the distance

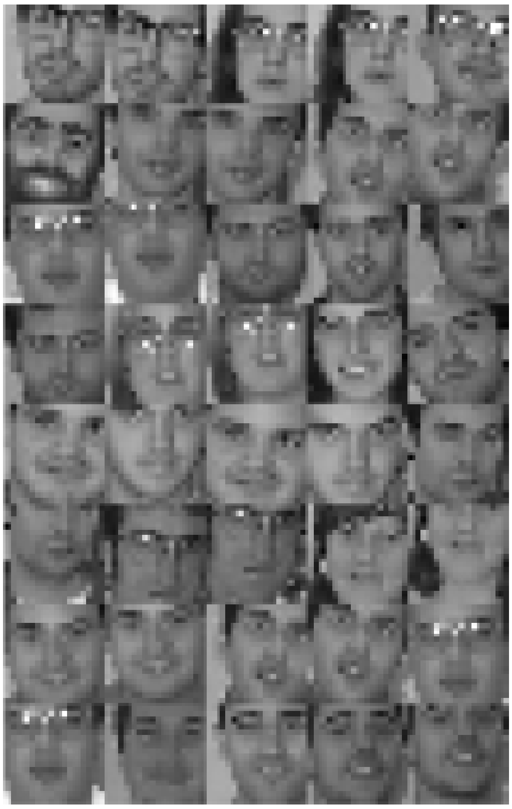


Figure 4-1: Face patterns extracted from our database

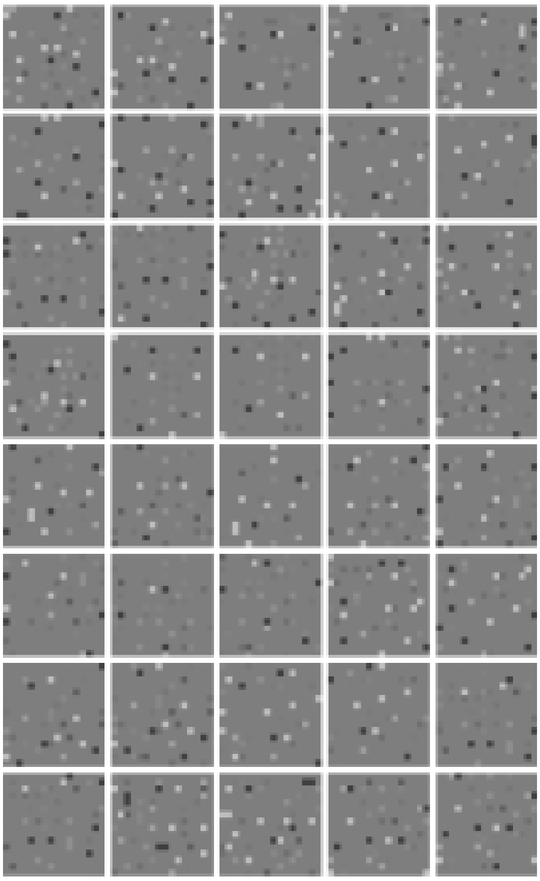


Figure 4-2: The fingerprints corresponding to our face patterns extracted from our image database . Darker points correspond to local minima, lighter points to local maxima. The intermediate graylevel corresponds to no extremal pixels. The intensity is proportional to the scale.

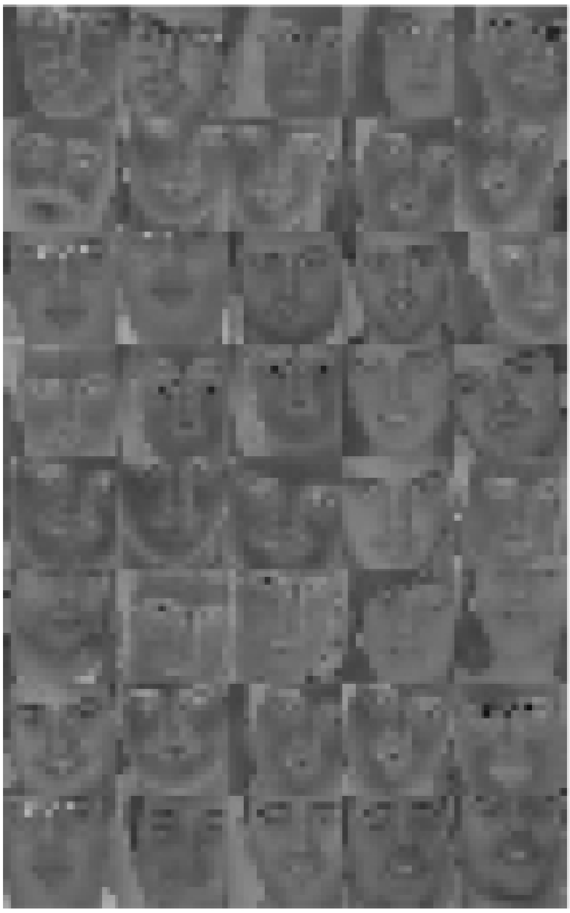


Figure 4-3: Eigenfaces obtained from the patterns in figure 1. The intensity has been shifted and scaled for visualization

to face space is the minimum of the distances to each pattern:

$$d_{MMF}\left(f,\{g_k\}\right)(x)=\min_k D\left(f,g_k\right)(x) \tag{4.31}$$

Once detected the face locations, let be L_i the set of detected face pixels in the image is constructed by placing a square of the same size as the face patterns centered at each of the positions in L_i. The positive detections are the pixels classified as face pixels, negative detections are the pixels not classified as face pixels. *True positive* are the positive detections that correspond to face pixels in the given ground truth. *False positive* are the positive detections that do not correspond to face pixels in the ground truth. *True* and *false negative* are the negative detections that correspond, respectively to no face and face pixels in the ground truth.

We computed the normalized ratios:

$$TP\left(\theta\right)=\frac{\#\left(true\,positive\right)}{\#\left(\text{positive }ground\,truth\right)} \tag{4.32}$$

and

$$FP\left(\theta\right)=\frac{\#\left(false\,positive\right)}{\#\left(negative\,ground\,truth\right)} \tag{4.33}$$

where $\#\left(A\right)$ denotes the cardinality of the set A. Both ratios $TP\left(\theta\right)$ and $FP\left(\theta\right)$ are monotonically increasing functions of θ. The $TP\left(\theta\right)$ ratio approaches 1 when all the faces in the image are detected, and the $FP\left(\theta\right)$ ratio approaches 1 when all the detected faces correspond to false positives. The plot of $TP\left(\theta\right)$ versus $FP\left(\theta\right)$ gives the Receiving Operator Characteristics (ROC) (the ratio of the true positive pixels to false positive pixels). This ROC can be used to define the decision threshold and also to compare the behavior of the PCA and MMF approaches to face localization. A measure of the optimality of a given detection algorithm is the distance from the point $(1,0)$ in the (FP,TP) space to its nearest point in the ROC curve. This point indicated the optimal trade-off between false alarms and positive detections that can be achieved by the algorithm.

In figures 4-4 and 4-5 we plot the ROC that correspond to the face detection with PCA, the MMF based on the nearest-neighbor classification over the whole set of face patterns (MMF-NN), and the MMF based on the mean face pattern (MMF-Mean). Figure 4-4 shows the

ROCs for the training images and figure 4-5 for the test images. Finally, images 4-6 and 4-7 show some examples of face detection instances with the constraint of detecting the 90% of the ground truth face pixels. The figures 4-6 (first column) and 4-7 (first column) present the ground-truth face pixels hand-labeled independently from the face patterns shown in figure 4-1. The results in figures 4-4 and 4-5 may be misleading because $FP(\theta)$ is normalized against the figure background, so that a small variation of the ratio $FP(\theta)$ corresponds to a big increase in the absolute number of false positives. In figures 4-6 (second column) and 4-7 (second column) we show the positive face localization performed using the MMF of the set of face patterns shown in figure 4-2. In figures 4-6 (third column) and 4-7 (third column) the same using PCA. In figures 4-6 (last column) and 4-7 (last column) the localization are performed using the MMF of the mean face pattern.

It can be appreciated in figure 4-4 and 4-5 that the face localization using the MMF consistently performs better than PCA. Both MMF-NN and MMF-Mean cases give higher positive detection rates for the same false positive ratio than PCA. Both approaches eventually detect all the labeled faces in the images, but MMF does with less false positives it than PCA. There are some differences in the ROCs of the MMF-NN and MMF-Mean. In the training images, the MMF-NN shows a sharp rise in recognition up to a level, which corresponds to the rate of true positive detection when the face patterns are 100% detected in the image. Afterwards, MMF-NN performs similar to MMF-Mean. The MMF of the mean face pattern has the same generalization power than the whole set of MMF of face patterns. This conclusion is reinforced by the results over the test images, where MMF-Mean improves over MMF-NN in some regions of the ROC. This is an important computational result because MMF-Mean implied 40 times less parameters than MMF-NN and PCA in this example experiments, and it is 40 times faster.

We have also tested our approach on the CMU image database (see appendix B), which shows a great variety of face poses and illumination conditions. We have followed a similar procedure as before: extracting face patterns and computing the distance in the space of multiscale morphological fingerprints. This time we have not compared with PCA, instead we used the published results (as they are found in the web page) of the CMU face detector as references to compare with our results. Figure 4-8 shows the set of face patterns extracted from twenty images of the CMU database. We have computed the ROC of our approach over the

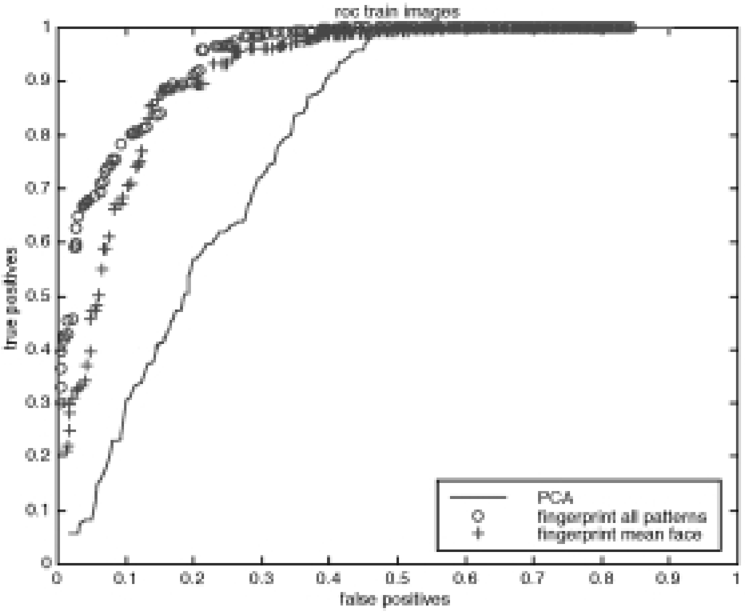


Figure 4-4: ROC for the train images from our database using the fingerprints of the mean face (+), PCA (.) and the combination of all patterns

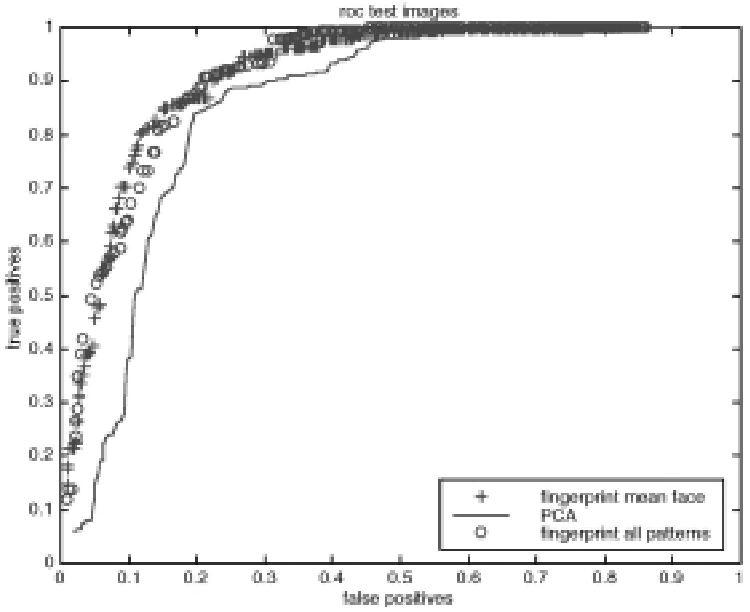


Figure 4-5: ROC for the test images from our database using the fingerprints of the mean face (+), PCA (.) and the combination of all pattern fingerprints (o)

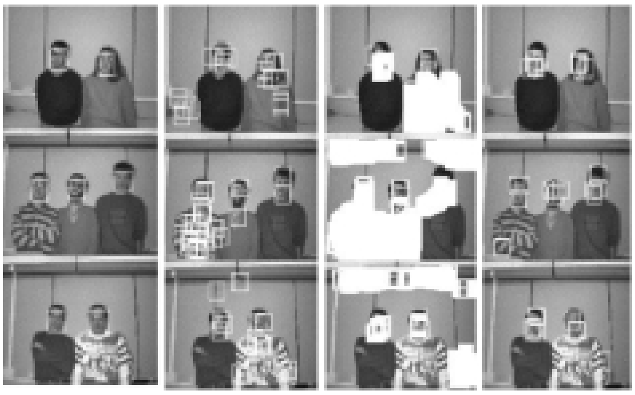


Figure 4-6: Face localization in some of the training images from our database, from which the face patterns have been extracted (from left to right): ground truth face pixels, using the fingerprints of the mean face pattern, using PCA and using a combination of all pattern fingerprints, respectively.

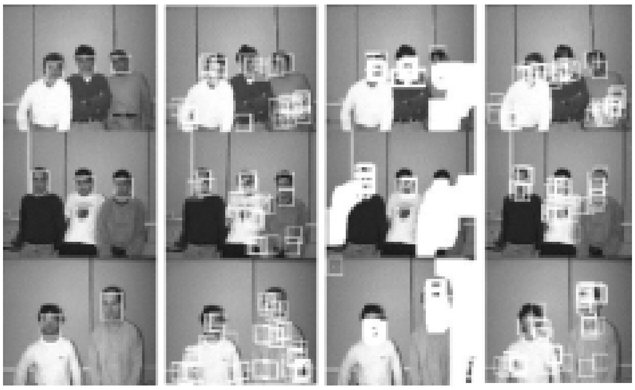


Figure 4-7: Face localization in some of the test images from our database (from left to right): ground truth face pixels, using the fingerprints of the mean face pattern, using PCA and using the combination of all pattern fingerprints, respectively

CMU images. For the computation of this ROC we have employed one hundred images, so the test set contains eighty images. Given that there is no face ground truth defined in the CMU database we have performed our own manual face labeling. The lighter areas in the images in figure 4-10 correspond to the hand-labeled face pixels used to compute the ROC curves. The ROC plotted in figure 4-9 shows similar performances as the one obtained for the small custom database. The original one hundred images selected show a wide variation in scale of the faces. Our technique can cope with small variations in scale. Face detection over wide variations in scale can be done through a multi-resolution approach. To obtain the results shown here, we have applied our technique to the images resized so that the scale factor of the faces in the images, relative to the patterns shown in figure 4-8 , is less than two. On going work deals with the definition of the multi-resolution approach to cope with larger variations in scale.

Finally, in figures 4-10 and 4-11 we give the detection results of our approach and those reported by the CMU system respectively, over a set of images which show a relative variation of pose, illuminations and scale. Note that the images are resized for a proper printing layout. These results show that our approach can be used as an alternative cue in a multicue system for face detection. Our results were obtained setting the decision threshold so as to obtain a 60% of the hand-labeled face pixels detected. Therefore, some of the faces are not detected and, besides, we obtained a small number of false positive faces. Higher detection rates lead also to a higher number of false positive. Most of the faces from figure 4-10 were not included in the face patterns from figure 4-8, so the results can be considered as a generalization capability of our approach.

4.7 Conclusions

We have reviewed a scale-space approach to image analysis based on multiscale erosion-dilation operators. The local extrema (minima and maxima) at each scale are the fingerprints of this scale-space, they constitute the Morphological Multiscale Fingerprints (MMF). A graph matching approach based on these fingerprints have been introduced for pattern recognition. Face localization consists in the graph matching based comparison of fingerprints in the test image and the training patterns. A simplified matching measure that counts the difference between

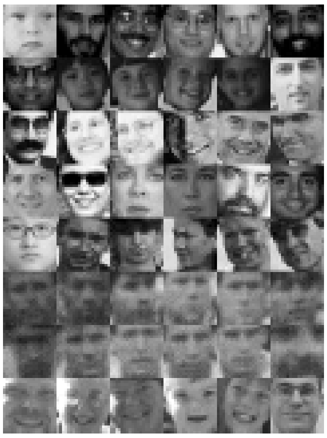


Figure 4-8: Face patterns extracted from the CMU database

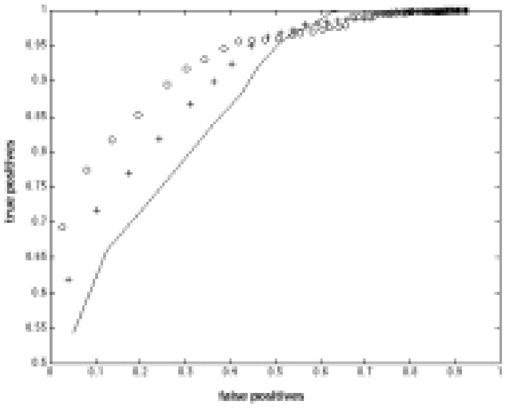


Figure 4-9: ROC computed over the CMU images: (o) images from which the face patterns have been extracted, (.) test images and (+) all the images taken together

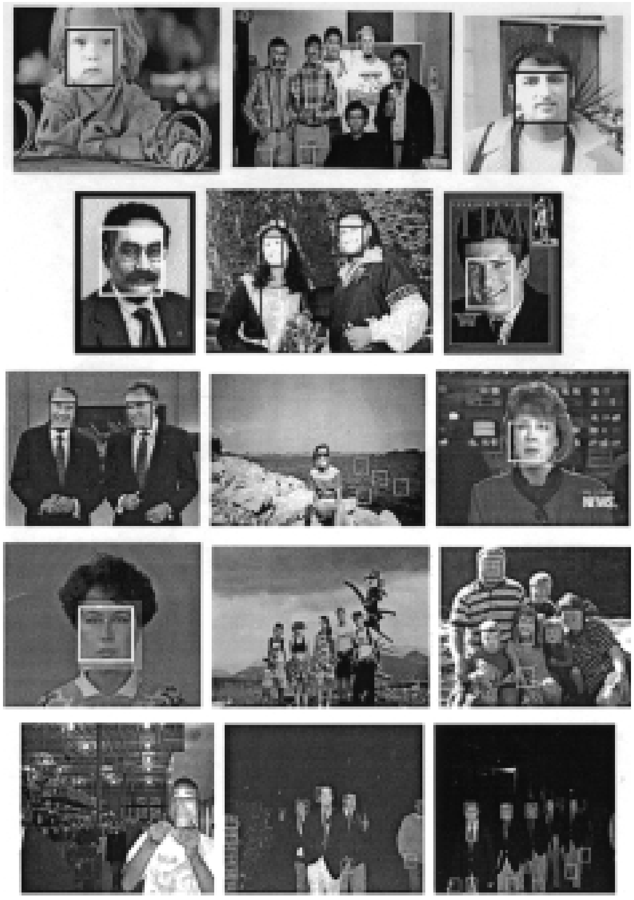


Figure 4-10: Results of our approach over a set of instances from the CMU database. The lighter areas correspond to the face pixels that we have selected to compute the ROC curves. The detection threshold was set so as to ensure a 60% detection of face pixels

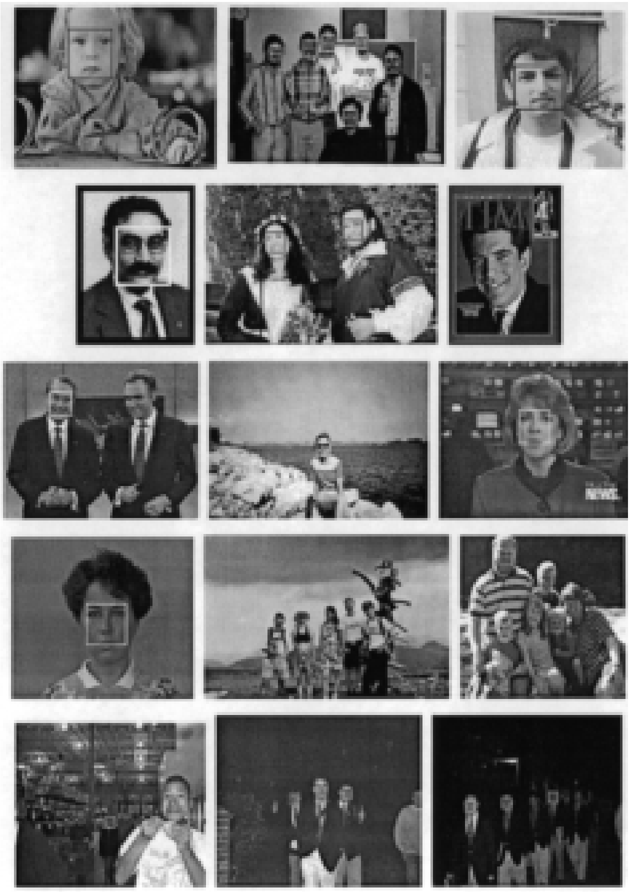


Figure 4-11: Results of the CMU face detection system as published in their website, for instances shown in figure 11

the number of fingerprints at each scale is used in the experiments.

Comparison with the PCA based approach on a small image database is favorable to the MMF approach. Emphasis must be put into the robustness of the approach in the sense that it gives good results despite the uncaring selection of the training patterns.

When the approach is applied to the CMU database, we find results comparable with the carefully trained system of the group of Kanade [114].

Future works must apply more sophisticated graph matching measures and strategies that may result in enhanced robustness and scale and rotation invariant detection.

Chapter 5

Face Localization with Morphological Associative Memories

In this chapter we focus on the potential application of a novel neural network architecture, based on mathematical morphology ideas, to the task of face localization. We give a review of the theory that supports them and then we present our approach based on the morphological heteroassociative memories and morphological scale-space.

5.1 Introduction

We start the presentation of Morphological Associative Memories (introduced in [112] and [113]) making a comparison with the classical ones. The mathematical framework for morphological associative memories is given by the homomorphism between the algebraic lattice defined by the addition and product operations, and the algebraic lattice defined by the supremum (infimum) and addition operators. In the classical definition of neural networks, the output of a neuron is computed as a non-linear filtering of a linear combination of its inputs. The equations below describe the neuron behavior:

$$\tau_i(t+1) = \sum_{j=1}^{n} a_j(t) w_{ij}, \tag{5.1}$$

$$a_i(t+1) = f(\tau_i(t+1) - \theta_i), \tag{5.2}$$

where $\tau_i(t+1)$ denotes the excitation of the neuron i at time $(t+1)$, $a_j(t)$ the output value of the jth neuron at time t, w_{ij} is the weight value of the link between the neurons i and j, n the number of inputs, f the activation function and θ_i the threshold.

To equivalent morphological neural network definitions are obtained by substitution of multiplication by addition, and substitution of the addition by Max or Min operators. The equation describing the morphological neuron state is given by the following equations:

$$\tau_i(t+1) = \max_{j=1..n} \{a_j(t) + w_{ij}\}, \tag{5.3}$$

$$\tau_i(t+1) = \min_{j=1..n} \{a_j(t) + w_{ij}\}. \tag{5.4}$$

Usually, for morphological neural networks, the activation function is the identity function.

From a mathematical point of view, the operations described by the equations (5.3) and (5.4) are based on the algebraic lattice structure, $(\mathbb{R}, \max, \min, +)$. At the same time, the equations (5.3) and (5.4) represent the basic operations of dilation and erosion from mathematical morphology. That's why we call this type of networks, morphological neural networks.

Writing the equations (5.3) and (5.4) in matrix form, we obtain:

$$T = W \bigtriangledown a, \tag{5.5}$$

and respectively,

$$T = M \bigtriangleup a \tag{5.6}$$

where a is the vector of neuron activations, W or M are the weight matrices and T is the output vector.

The operators $\bigtriangledown$ and $\bigtriangleup$ are called *matrix Max-product* and *matrix Min-product,* respectively. For two matrices chosen of proper size, A and B, we may define their *Max-product* $(C = A \bigtriangledown B)$ and *Min-product* $(C = A \bigtriangleup B)$ as:

$$C = A \bigtriangledown B = [c_{ij}] \Leftrightarrow c_{ij} = \max_{k=1..n} \{a_{ik} + b_{kj}\} \tag{5.7}$$

$$C = A \bigtriangleup B = [c_{ij}] \Leftrightarrow c_{ij} = \min_{k=1..n} \{a_{ik} + b_{kj}\} \tag{5.8}$$

5.2 Morphological Associative Memories

Morphological associative memories were proposed in [112] and [113] as a special kind of neural network whose goal is to store and recall a set of input-output pattern pairs. Let's consider we have k vector pairs $\left\{(x^1, y^1), (x^2, y^2), ..., (x^k, y^k)\right\}$, where $x^\xi \in \mathbb{R}^n$ and $y^\xi \in \mathbb{R}^m$ for any $\xi = 1..k$. With the pair of matrices (X, Y) we can build two types of associative memories, characterized by their weight matrices $W_{XY} = (w_{ij})$ and $M_{XY} = (m_{ij})$, respectively:

$$w_{ij} = \min_{l=1..k} \left\{y_i^\xi - x_j^\xi\right\}, \quad \xi = 1..k \tag{5.9}$$

$$m_{ij} = \max_{l=1..k} \left\{y_i^\xi - x_j^\xi\right\}, \quad \xi = 1..k \tag{5.10}$$

Matrices W and M are known as *Heteroassociative Morphological Memory* (HMM). From the previous relations, it follows that the matrices W and M are lower and upper bounds of Max/Min products, for $\forall\xi$, i.e. $W_{XY} \leq y^\xi \times \left(-x^\xi\right) \leq M_{XY}$, where by × we denote any of the operators $\bigtriangledown$ or $\bigtriangleup$, respectively. Therefore, the following bounds on the output patterns hold

$$\forall\xi : W_{XY} \bigtriangledown x^\xi \leq y^\xi \leq M_{XY} \bigtriangleup x^\xi \tag{5.11}$$

that can be rewritten in matrix form:

$$W_{XY} \bigtriangledown X \leq Y \leq M_{XY} \bigtriangleup X. \tag{5.12}$$

A particular situation related with these morphological associative memories occurs when $X = Y$. In this case, we deal with *Morphological Autoassociative Memory* (AMM). A very interesting property of AMM, is that they always provide perfect recall for any number of memorized patterns, i.e.

$$W_{XX} \bigtriangledown X = X \text{ and } M_{XX} \bigtriangleup X = X, \tag{5.13}$$

respectively. For this reason, we can say that, the autoassociative memories have unlimited storage capacity. Another advantage that results from here is that the recall process occurs in 1 step and it can be implemented in parallel.

At this point, a question that arises naturally is: what about corrupted versions of input patterns? Between which limits can vary a corrupted pattern such that recall of the stored patterns can still be guaranteed? Let us start defining the kind of perturbations that affect the patterns from a morphological point of view.

Definition 20 *Let be $\widetilde{\mathbf{x}}^{\gamma}$ a noisy version of $\mathbf{x}^{\gamma}$. If $\widetilde{\mathbf{x}}^{\gamma} \leq \mathbf{x}^{\gamma}$ then $\widetilde{\mathbf{x}}^{\gamma}$ is an eroded version of $\mathbf{x}^{\gamma}$, or $\widetilde{\mathbf{x}}^{\gamma}$ is subjected to erosive noise. If $\widetilde{\mathbf{x}}^{\gamma} \geq \mathbf{x}^{\gamma}$ then $\widetilde{\mathbf{x}}^{\gamma}$ is a dilated version of $\mathbf{x}^{\gamma}$, or $\widetilde{\mathbf{x}}^{\gamma}$ is subjected to dilatative noise.*

When the corrupted pattern presents only one of these types of noise, the problem of recalling the correct output is not difficult to solve. It can be shown that W associative memories cope well with erosive noise, while M associative memories cope well with dilative noise. However, real life noise is both erosive and dilatative. A more complex approach should be taken when we deal with patterns that present a mixture of erosive and dilatative noise. In this situation, a straightforward approach can't be applied. More general robustness is achieved through a composition of autoassociative and heteroassociative memories based on the definition of *kernel* patterns [112] and [113].

For instance, let us follow the reasoning and the definition of kernels in the case of the W memories. We can define a memory W such that it can associate randomly corrupted versions of the pattern x^{ξ} with a sub-pattern z^{ξ} specially designed from x^{ξ}. Pattern z^{ξ} represents a special dilated version of x^{ξ}, that fulfills the following requirements: It greater than the original pattern, and no other pattern is strictly lesser than it

$$z^{\xi} \gg x^{\xi} \text{ and } z^{\xi} \vee x^{\gamma} = \mathbf{1}; \xi \neq \gamma$$

($\mathbf{1}$ is the unit vector) The last condition is written assuming the Boolean case. The memory W is defined as an autoassociative memory for the patterns z^{ξ}, $\xi = 1, ..., k$ so that $W \bigtriangledown z^{\xi} = z^{\xi}$. Since W is robust against erosive noise, and most corrupted version $\tilde{x}^{\xi}$of x^{ξ} will be seen like an eroded version of z^{ξ}, we can have most of the times $W \bigtriangledown \tilde{x}^{\xi} = z^{\xi}$. Furthermore, the same relation is also true for x^{ξ} , because x^{ξ} is an eroded version of z^{ξ}, too. In other words, $W \bigtriangledown x^{\xi} = z^{\xi}$. In fact this relation holds as long as $z^{\xi} \gg \widetilde{x}^{\xi}$ is true.

To implement the output mapping another memory M is defined like an heteroassociative memory that associates each input pattern z^{ξ} with the output pattern y^{ξ}. If the patterns z^{ξ} are appropriately chose, we have that the composition of the M and W memories realizes the desired heteroassociative memory:

$$M \bigtriangleup (W \bigtriangledown x^{\xi}) = M \bigtriangleup z^{\xi} = y^{\xi} \tag{5.14}$$

and, also, that this heteroassociative memory is robust against corrupted inputs, it recalls the desired output:

$$M \bigtriangleup (W \bigtriangledown \tilde{x}^{\xi}) = M \bigtriangleup z^{\xi} = y^{\xi} \tag{5.15}$$

for randomly corrupted patterns $\widetilde{x}^{\xi}$ for which $z^{\xi} \gg \widetilde{x}^{\xi}$. This discussion is summarized in the following definition.

Definition 21 *Let $Z = (\ z^1,\ z^2, ...,\ z^k)$ be an $n \times k$ matrix. We say that Z is a kernel for (X, Y) iif the following conditions are satisfied:*

1. $W_{ZZ} \bigtriangledown X = Z$
2. $M_{ZY} \bigtriangleup Z = Y$

It follows that if Z is a kernel for (X, Y) then we have:

$$M_{ZY} \bigtriangleup (W \bigtriangledown X) = M_{ZY} \bigtriangleup Z = Y \tag{5.16}$$

The following theorem gives a clue of finding the kernels Z.

Theorem 22 *If Z is a kernel for (X, Y) then $Z \geq X$.*

If Z is a kernel for (X, Y) then the column vector $z^{\xi}, \xi = 1, ..., k$ are called *kernel vectors* for (X, Y). According to the above theorem, it results that $z^{\xi} \geq x^{\xi}$. Thus, the kernel vectors represent dilated versions of the patterns $\{x^1, x^2, ..., x^k\}$.

The cases in which $Z = X$ (or even $z^\xi = x^\xi$, for some arbitrary ξ) need to be avoided. This conclusion leads us to the notion of *proper kernels.*

Definition 23 *If Z is a kernel for (X, Y) with the property $z^\xi \neq x^\xi \ \forall \xi$, then Z is called a proper kernel for (X, Y).*

The notion of proper kernel does not imply that $z^\xi > x^\xi \ \forall \xi$. It only implies that for each $\xi = 1, 2, ..., k$ there exists an index $i_\xi \in \{1, ..., n\}$ such that $z^\xi_{i_\xi} \neq x^\xi_{i_\xi}$. In other words, each kernel vector z^ξ has at least one coordinate that differs (it is strictly grater than) from the corresponding coordinate of x^ξ.

If $z^\gamma \geq \tilde{x}^\gamma \geq x^\gamma$ then,

$$z^\gamma = W_{ZZ} \bigtriangledown z^\gamma \geq W_{ZZ} \bigtriangledown \tilde{x}^\gamma \geq W_{ZZ} \bigtriangledown x^\gamma = z^\gamma \tag{5.17}$$

and hence, $W_{ZZ} \bigtriangledown \tilde{x}^\gamma = z^\gamma$. Thus, if Z is a kernel for (X, Y) and $z^\gamma \geq \tilde{x}^\gamma \geq x^\gamma$ then we are guaranteed that

$$M_{ZY} \bigtriangleup (W_{ZZ} \bigtriangledown \tilde{x}^\gamma) = y^\gamma \tag{5.18}$$

Therefore, a dilated version $\tilde{x}^\gamma$ of x^γ satisfying the inequality $z^\gamma \geq \tilde{x}^\gamma$ will be correctly associated with the pattern y^γ. The next theorem will establish the conditions that guarantee perfect recall from the corrupted input patterns $\tilde{x}^\gamma$.

Theorem 24 *If Z is a kernel for (X, Y) and $\tilde{x}^\gamma$ is a corrupted version of the pattern x^γ such that $z^\gamma_j \geq \tilde{x}_j \forall j = 1, ..., n$ and $z^\gamma_j = \tilde{x}_j$ whenever $w_{ZZ}(i, j) = z^\gamma_i - z^\gamma_j$, then $W_{ZZ} \bigtriangledown \tilde{x}^\gamma = z^\gamma$.*

The last two theorems give some guides on the nature of kernel vectors and the rough steps to build them. A first step is to dilate the pattern vectors x^ξ, saving only a few of the original pattern values. In the Boolean case, one needs to turn most pattern values of the vector x^ξ to one, saving only a few unique pattern values to build the kernel vector z^ξ such that $z^\xi \vee x^\gamma = \mathbf{1}$ (unit vector) whenever $\gamma \neq \xi$.

An additional benefit of using kernel patterns is the increase in storage capacity for associations (X, Y). This is due to the fact that W_{ZZ} is a perfect recall memory for any finite number of pattern vectors z^{ξ} and because of the sparseness of dilated patterns (i.e., in the Boolean case most pattern values will be one), the storage capacity of M_{ZY} can be greatly increased.

However, we see two main drawbacks in the systematic application of this kernel framework:

- The size of the morphological autoassociative memories that grows quadratically with the size of the input and kernel patterns. That prevents its use in practical applications involving even small size images and
- The construction of the kernels is an artisan work. It does not exist any systematic procedure of construction.
- There is no formal characterization of the amount of noise that these constructions can support.

5.3 Heteroassociative memories and morphological scale-space

Given a set of patterns, the definition of a systematic procedure for the kernel extraction is still an open research track. However, our main problem is the computational complexity. We go back to the heteroassociative memory's theory searching for solutions to real life applications which are computationaly feasible. To add robustness to the heteroassociative morphological memories (HMM), we will follow a construction strategy based on ideas extractedextracted from the multiscale theory. Let us start by recalling some theoretical background on HMM [112], [113].

The weight matrices are lower and upper bounds of the Max/Min products $\forall\xi; W_{XY} \leq y^{\xi} \times \left(-x^{\xi}\right)' \leq M_{XY}$ and therefore the following bounds on the output patterns hold $\forall\xi, W_{XY} \bigtriangledown x^{\xi} \leq y^{\xi} \leq M_{XY} \bigtriangleup x^{\xi}$, that can be rewritten .

$$W_{XY} \bigtriangledown X \leq Y \leq M_{XY} \bigtriangleup X. \tag{5.19}$$

Definition 25 *Given a set of input-output pairs,* (X, Y) *, a matrix* A *is a* $\bigtriangledown$*-perfect (*$\bigtriangleup$*-perfect) memory for* (X, Y) *if* $A \bigtriangledown X = Y$ *(*$A \bigtriangleup X = Y$*).*

It can be proven that if A and B are $\bigtriangledown$ -perfect and $\bigtriangleup$-perfect memories for (X, Y) then

W_{XY} and M_{XY} are also $\bigtriangledown$-perfect and $\bigtriangleup$-perfect, respectively:

$$A \leq W_{XY} \leq M_{XY} \leq B \tag{5.20}$$

$$W_{XY} \bigtriangledown X = Y = M_{XY} \bigtriangleup X \tag{5.21}$$

Put it the other way, the M and W memories are perfect by construction, unless some interfering between the stored patterns occurs Conditions for perfect recall of noise free input-output pairs [112], [113] on the memories are given by the following theorem:

Theorem 26 ((Perfect recall HMM)) *W_{XY} is a $\bigtriangledown$-perfect memory if and only if, for all ξ the matrix $\left[y^{\xi} \times \left(-x^{\xi}\right)'\right] - W_{XY}$ contains a zero at each row. Similarly M_{XY} is a $\bigtriangleup$-perfect if and only if for all ξ the matrix $\left[y^{\xi} \times \left(-x^{\xi}\right)'\right] - M_{XY}$ contains a zero at each row. These conditions are rewritten for W_{XY} and M_{XY}, respectively, as follows:*

$$W_{XY} \bigtriangledown X = Y \Longleftrightarrow \forall\gamma\forall i\exists j; x_j^{\gamma} = \bigvee_{\xi=1}^{k} \left(x_j^{\xi} - y_i^{\xi}\right) + y_i^{\gamma},\ \textit{and} \tag{5.22}$$

$$M_{XY} \bigtriangleup X = Y \Longleftrightarrow \forall\gamma\forall i\exists j; x_j^{\gamma} = \bigwedge_{\xi=1}^{k} \left(x_j^{\xi} - y_i^{\xi}\right) + y_i^{\gamma}. \tag{5.23}$$

Let us translate this for the special case of images, as inputs, and orthogonal binary codes, as outputs, i.e. $y^{\xi} \cdot y^{\gamma} = \delta_{\xi\gamma}$ and $y^{\xi}, y^{\gamma} \in \{0,1\}^{N}$. We will assume a coding of the form

$$\left(y_{\xi}^{\xi} = 1, y_{\gamma}^{\xi} = 0, \gamma \neq \xi\right). \tag{5.24}$$

Then, condition 5.22 for perfect recall of the W_{XY} matrix splits into two conditions. For each input vector γ,

$$\forall i \neq \gamma\ \exists j; x_j^{\gamma} = \bigvee_{\xi=1}^{k} \left(x_j^{\xi} - y_i^{\xi}\right) \geq \bigvee_{\xi=1}^{k} x_j^{\xi} \tag{5.25}$$

for the components or the output vector that must be zero because of the orthogonal coding (5.24),

$$i = \gamma\ \exists j; x_j^{\gamma} = \bigvee_{\xi=1}^{k} \left(x_j^{\xi} - y_i^{\xi}\right) + 1 \geq \left(\bigvee_{\xi=1}^{k} x_j^{\xi}\right) + 1 \tag{5.26}$$

for the single component of the output vector that must be one.

Therefore, in order that image and output code $\left(x^{\xi}, y^{\xi}\right)$ can be properly stored in a W_{XY} memory, and the output y^{ξ} recalled by $W_{XY} \bigtriangledown x^{\xi}$ there must exist some pixel positions whose values are greater in this image than the maximum value of these pixels in the remaining images. In other words, we can not store patterns for which $x^{\xi} < x^{\gamma} - 1$ for any other $\gamma \neq \xi$. The dual assertion is true for M_{XY} memory. In order for the pair $\left(x^{\xi}, y^{\xi}\right)$ to be stored in M_{XY} and the output y^{ξ} recalled by $M_{XY} \bigtriangleup x^{\xi}$, then there must exist some pixels whose values are lower in this image than the minimum value found in the remaining images.

We will call pseudo-kernel the set of indices that ensure the recognition with W_{XY}. $K_M\left(x^{\gamma}\right) = \left\{j \left| x_j^{\gamma} > x_j^{\xi},\ \gamma \neq \xi \right.\right\}$.. This is the set of distinctive pixels that ensure that image x^{γ} and its class code can be stored and retrieved in a noiseless setting. The dual pseudo-kernel $K_M\left(x^{\gamma}\right) = \left\{j \left| x_j^{\gamma} < x_j^{\xi},\ \gamma \neq \xi \right.\right\}$ corresponds to the set of indices that ensure storage and recall with the M heteroassociative memory.

The theorem on the perfect recall of the HMM can be reduced in the special case of orthogonal binary output codes to the following corollary:

Corollary 27 *Given input-output pairs (X, Y) with the output being a set of orthogonal binary vectors, the conditions for perfect recall of the W_{XY} and M_{XY} memories are that for all the input patterns , respectively, $K_M\left(x^{\gamma}\right) \neq \emptyset$ and $K_W\left(x^{\gamma}\right) \neq \emptyset$.*

Coming back to the issue of noisy patterns the next theorem states the conditions for perfect recall of the output..

Theorem 28 (Perfect recall for noisy patterns.) *The conditions for the perfect recall of y^{γ} given a noisy copy $\widetilde{x}^{\gamma}$ of x^{γ} for W_{XY}, that is, the conditions under which $W_{XY} \bigtriangledown \widetilde{x}^{\gamma} = y^{\gamma}$ are as follows:*

$$\begin{aligned} \forall j; \widetilde{x}_j^{\gamma} &\leq x_j^{\gamma} \vee \bigwedge_i \left(\bigvee_{\xi \neq \gamma} \left(x_j^{\xi} - y_i^{\xi} + y_i^{\gamma} \right) \right) \quad \textit{and} \qquad (5.27) \\ \forall i \exists j_i; \widetilde{x}_{j_i}^{\gamma} &= x_{j_i}^{\gamma} \vee \left(\bigvee_{\xi \neq \gamma} \left(x_{j_i}^{\xi} - y_i^{\xi} + y_i^{\gamma} \right) \right). \end{aligned}$$

Similarly for the perfect recall of y^γ given a noisy copy $\widetilde{x}^\gamma$ of x^γ for M_{XY}, that is, the conditions under which $M_{XY} \bigtriangleup \widetilde{x}^\gamma = y^\gamma$ are as follows:

$$\begin{aligned} \forall j; \widetilde{x}_j^\gamma &\geq x_j^\gamma \wedge \bigvee_i \left(\bigwedge_{\xi \neq \gamma} \left(x_j^\xi - y_i^\xi + y_i^\gamma \right) \right) \quad \textit{and} \\ \forall i \exists j_i; \widetilde{x}_{j_i}^\gamma &= x_{j_i}^\gamma \wedge \left(\bigwedge_{\xi \neq \gamma} \left(x_{j_i}^\xi - y_i^\xi + y_i^\gamma \right) \right). \end{aligned} \tag{5.28}$$

These conditions (5.28), (5.27) are the basis for our approach. The condition (5.27) states that the matrix W_{XY} is robust against controlled erosions of the stored input patterns while condition (5.28) states that the matrix M_{XY} is robust against controlled dilations of the input patterns. Therefore if we store in the W matrix a set of dilated patterns, the input could be considered most of the times as an erosion of the stored pattern. Conversely, if we store in the M matrix a set of eroded patterns, the input could be considered most of the times a dilation of the stored pattern.

In the case of orthogonal binary output vectors, the theorem of perfect recall for noisy patterns can be again simplified to the following corollary:

Corollary 29 *Given (X, Y) input-output pairs, with the output a set of binary orthogonal vectors. Let it be $\widetilde{x}^\gamma$ a noisy copy of x^γ. The conditions under which $W_{XY} \bigtriangledown \widetilde{x}^\gamma = y^\gamma$ are that $\widetilde{x}^\gamma \leq x^\gamma$ and $\exists j \in K_W(x^\gamma)$ such that $\widetilde{x}_j^\gamma = x_j^\gamma$. The conditions under which $M_{XY} \bigtriangleup \widetilde{x}^\gamma = y^\gamma$ are that $\widetilde{x}^\gamma \geq x^\gamma$ and $\exists j \in K_M(x^\gamma)$ such that $\widetilde{x}_j^\gamma = x_j^\gamma$.*

We will consider these matrices as approximations to the ideal memory of all the distorted versions of the input data, so that their output is an approximation to the response of this ideal memory.

We apply a scale space approach to systematize the construction of the matrices and to characterize the robustness of the process.

If we construct the W memory with patterns dilated with a structural function of scale σ, that comply with the conditions given in chapter 4 for an scaling function, we preserve the recognition of the original patterns if the kernel of the patterns contain local maxima of scale σ or higher. By the continuity proposition of the morphological scale-space, the local extrema neither are displaced nor changed in value. Therefore, if we erode/dilate all the patterns with the same structural function their relative values will be preserved at the points were the local extrema of the images are placed. We can put it as the following proposition:

Proposition 30 *Given a set of pairs* (X, Y) *such that we can construct a* W_{XY} *memory which is*$\bigtriangledown$*-perfect , let us denote the dilated patterns as* $x^{\gamma,\sigma} = x^{\gamma} \oplus g_{\sigma}$; $(X^{\sigma}, Y) = \{(x^{\gamma,\sigma}, y^{\gamma})\}$. *Then,*

$$W_{X^{\sigma},Y} \bigtriangledown x^{\gamma} = y^{\gamma} \text{ if and only if } E\left(x^{\gamma},\sigma\right) \cap K_W\left(x^{\gamma}\right) \neq \emptyset, \tag{5.29}$$

where $E\left(x^{\gamma},\sigma\right)$ *is the set of local maxima at the scale* σ.

The proof of this proposition follows from the fact that patterns are only distinguished by their pseudo-kernels. These pseudo-kernels are not changed by morphological dilation with an structural function that complies with the properties of a scale-space. Therefore, as long as the pseudo-kernels are (at least partially) preserved, the input-output can be still recalled from the memories constructed from patterns dilated with the scaling function. An dual proposition can proven for the M HMM.

This construction gives also a characterization of the kind of noise that can be accepted. Let be $\widetilde{x}^{\gamma}$ such that $\widetilde{x}^{\gamma} \leq x^{\gamma,\sigma}$ and $\widetilde{x}_j^{\gamma} = x_j^{\gamma}$, for some $j \in E\left(x^{\gamma},\sigma\right) \cap K_W\left(x^{\gamma}\right)$. Then, $W_{X^{\sigma},Y} \bigtriangledown \widetilde{x}^{\gamma} = y^{\gamma}$. We have a construction of the W memory whose tolerance to dilatative noise, on top of its intrinsic robustness to the erosive noise, is characterized by the scale σ of the dilation..

Still there is no guarantee that the pseudo-kernel of the images will contain any local extrema points.

5.4 Multiscale HMM

Given a set of input patterns X and a set of output class encoding Y. We build a set of HMM $\{M_{XY}^{\sigma}, W_{XY}^{\sigma}; \sigma = 1, 2, ...\mathrm{s}\}$ where each M_{XY}^{σ} is constructed from output and the input patterns eroded with an spherical structural object of scale σ, and each W_{XY}^{σ} is constructed from the outputs and input patterns dilated with an spherical structural object of scale σ. Given a test input pattern $\mathbf{x}$, we compute the response of the memories at the different scales:

$$\mathbf{y}^M = \bigcup_{\sigma=1}^{\mathbf{s}} \left(M_{XY}^{\sigma} \bigtriangleup \mathbf{x}\right) \tag{5.30}$$

and

$$\mathbf{y}^W = \bigcup_{\sigma=1}^{\mathbf{s}} \left(W_{XY}^{\sigma} \bigtriangledown \mathbf{x}\right) \tag{5.31}$$

The final output is the intersection of these multiscale responses:

$$\mathbf{y} = \mathbf{y}^M \bigcap \mathbf{y}^W. \tag{5.32}$$

In the case of face localization, the output is the classification of the image block as a face, which is given as a block of white pixels whenever the input image block is identified with any of the stored face patterns.

5.5 Experimental Results

A set face patterns is selected as the representatives of the face class. In the experiments reported here the set of face patterns is the one presented in figure 5-1. This small set shows several interesting features: faces are of different sizes, background has been manually removed, there is no precise registration of face features (some of the faces are rotated), and there is no intensity normalization (equalization or any other illumination compensation). Therefore, building this set of patterns corresponds to an almost casual browsing and picking of face images in the database.

As described in the previous section, the M and W HMMs are built up to classify each image block as one of the face patterns. If it fails, the response is arbitrary and we consider the image block as a non-face block. The HMMs output are orthogonal binary vectors encoding the face pattern. Both memories are convolved with the image to search for faces. At each pixel the positive classification with one of the M^{σ} memories produces a square of face pixels of a size that is the half of the image block, and centered at this pixel position. The recognition at the different scales is added into an M-recognition binary image. The same process applies to the W^{σ} memories. The intersection of the face pixels recognized with each HMM is the final result, which is superimposed to the original image.

We have performed initial studies over a small database of 20 images with a varying range of scales. The average ROC curve over all the images relating the true and false positives obtained with scale ranges varying from $s = 1$ up to $s = 13$ is shown in figure 5-2. It can be appreciated that the approach obtains a high recognition rate (over 85%) with very small false recognition rates (less than 5%). As the scale range increases we reach the 100% of face recognition at the pixel level. Face pixels were labelled manually in a process which is independent of the selection of the face patterns. We give in figure 5-3 some images with recognition results at scale 5. We have also applied this approach, with the same set of face patterns, to the CMU database. Results were irregular, but in several cases it was good. We show in figure 5-4 some original images of the CMU and in figure 5-5 the results of the face localization with the face patterns presented in figure 5-1. The results were acceptable if we take into account that no size transformation has been applied for these concrete images. That is, the size of the face patterns and that of the faces in the images are not very similar. Also, results with very different illumination conditions are acceptable.

5.6 Conclusions

In this chapter we have taken as the computational tool, the Morphological Associative Memories proposed by Ritter and Sussner. Their appealing is their potential for real time applications, as long as they only require integer and Max/Min operations. We propose a multiscale construction as a mean to improve the robustness of the HMM's. This approach is based on the continuity property of multiscale erosion-dilation scale-spaces. We present experimental results on our custom image database and the CMU image database.

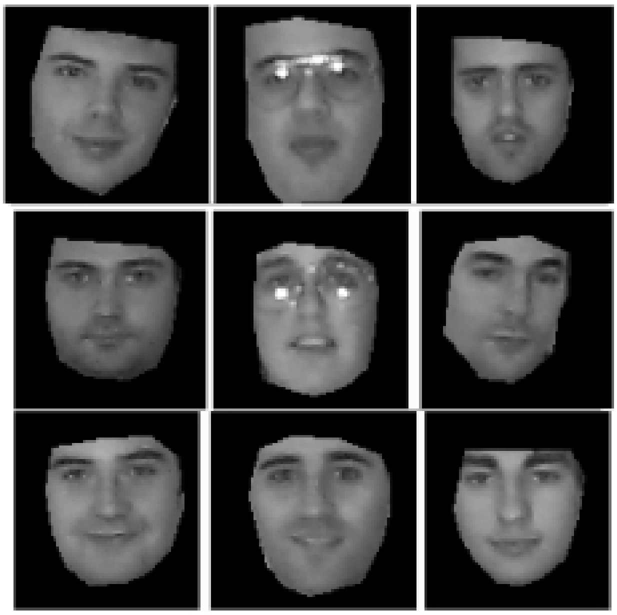


Figure 5-1: Face patterns used in our experiments with morphological heteroassociative memories as face localizers.

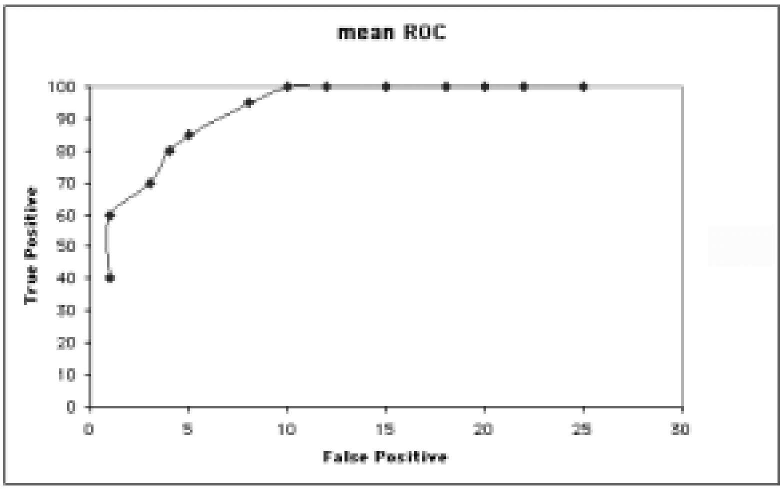


Figure 5-2: Average ROC curve for all images obtained varying the scale parameter.

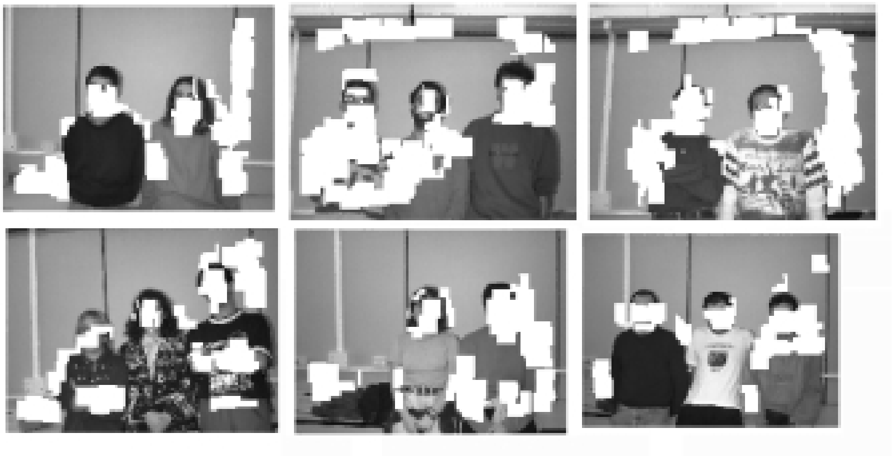


Figure 5-3: Some results of face localization using patterns eroded/dilated to scales up to 5

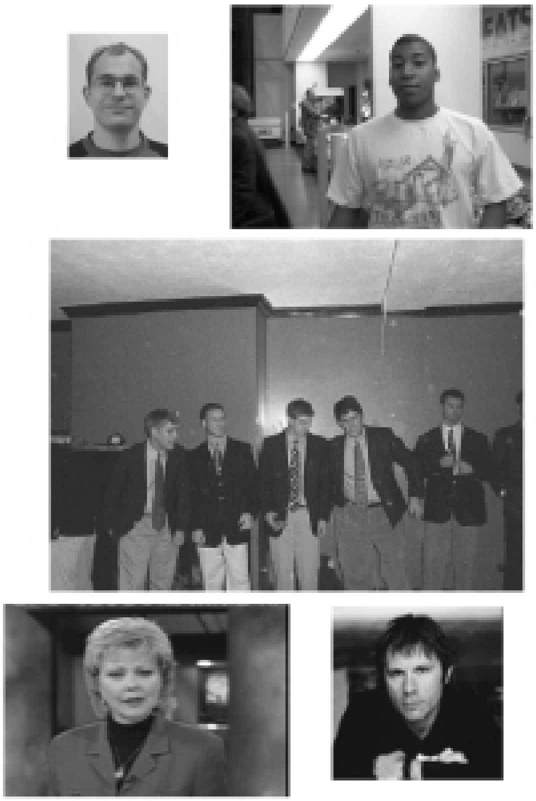


Figure 5-4: Some original images of the CMU used for face localization

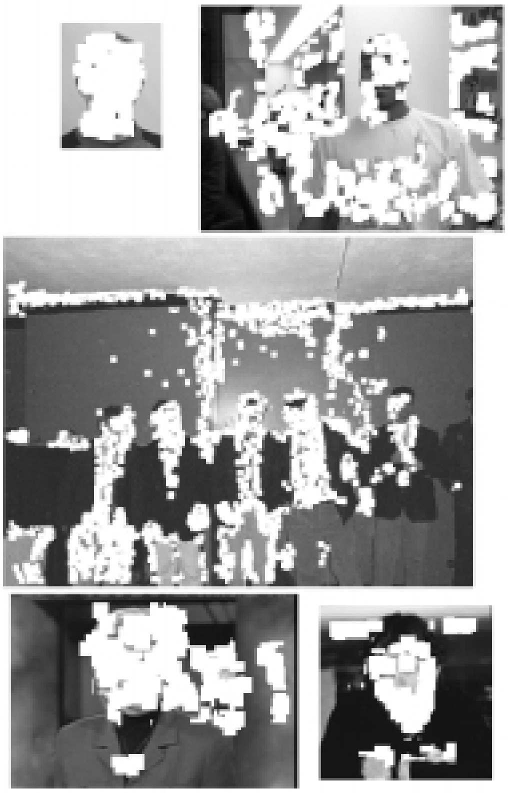


Figure 5-5: Results of the face localization with the HMM approach on the CMU images, using the face patterns displayed in fig. 5.1.

Chapter 6

Conclusions and Future Work

Face localization is a very interesting and difficult instance of object localization in images, under several circumstances. The real time systems use motion and color information to produce the segmentation of the image, reducing the search space, often producing a cropped image of enough quality to be the input to face recognition modules. However, this kind of techniques works well in very restricted conditions. The general face localization systems must deal with illumination variations, pose and scale variations and so on. In this very general setting, face localization is still an open problem. It is our belief that face detection must be based on a collection of clues, and so the study of different alternative detection algorithms that can be combined or fused is of interest.

The present thesis contributes to the solution of this problem exploring a set of approaches based on mathematical morphology. The first approach is based on the gray scale extension of the classical Hit-or-Miss Transform (HMT). The HMT has a radical appeal as the basic localization operator in mathematical morphology applied to image analysis. However in its fundamental formulation it is a very stiff operator, without the smaller tolerance to even the smaller perturbations of the searched patterns. Nevertheless, its appeal remains and several authors have proposed it for object recognition and object localization. Among them, some authors have proposed generalizations of the HMT to gray scale images. We tested this Gray scale HMT (GHMT) on the localization problem. Besides, we ourselves have proposed a generalization of the definition of the HMT to gray scale images. This generalization is based on the decomposition of the image in level sets and the reconstruction of the image from the level sets.

We call this transform Level Set based HMT (LSHMT) and we compare it with the GHMT on the task of face localization. We found our proposed LSHMT to be superior, and we think that it deserves further consideration. Among the potential improvements to this approach are the consideration of morphology in color spaces and the introduction of scale space techniques through the Blur HMT. The Blur HMT is an attempt to produce a measure of the uncertainty of the matching obtained with a HMT whose structural objects suffer controlled dilations. The extension of the ideas of the Blur HMT to the gray scale domain is straighforward in our approach based on the level sets.

The second approach is based on the detection of image features which consist of the Morphological Multiscale Fingerprints. The Scale Space approach to image analysis proposes the study of versions filtered at different scales in order to determine the true scale of each of the objects in the image. In Linear Scale Spaces the features of interest are the zero crossings of the filtered images, which determine the localization of the image edges or discontinuities. This approach has been very successful in several task of computer vision (i.e.: stereo correspondences). There have been propositions of Morphological Scale Spaces. Among them the proposition of Jackway of a Scale Space based on Multiscale Erosion-Dilation operators has drawn our attention. In this Scale Space the features of interest are the local extrema of the image. We propose to use these features for recognition based on a simplified graph matching paradigm. The matching based on the number of fingerprints found in at each scale is enough to give good detection results, although restricted in terms of the robustness to scale and rotation changes. The results of this approach have been published elsewhere, however we consider that still there is plenty of room for improvement in this approach. The consideration of spatial information in the graphs edges, improved matching strategies and computation of the fingerprints in color images, based on color space morphology could give more accurate results and provide simple and elegant solutions to the scale and rotation problems.

The third and last approach tested are the so called Morphological Associative Memories, proposed originally by Ritter, Sussner and Diaz de Leon. The aim of the Heteroassociative Memories is the storage and retrieval of a set of input-output patterns. The approach is very powerful, but degrades greatly in the presence of noise. The authors characterized the sensitivity and robustness of the approach in terms of erosive and dilatative noise. Depending on its construction, the Heteroassociative Memories are robust against a kind of noise and sensitive to the dual noise. To obtain robustness to the general simultaneously erosive and dilatative noise, they propose a construction based on the abstract notion of kernel. We find that it is possible to obtain a degree of robustness by the application of Scale Space ideas to the construction of the Heteroassociative Memories. The morphological scale space theory ensures that the construction of Heteroassociative Morphological Memories with eroded/dilated patterns following a scale space are robust to erosive and dilatative noise bounded by the scale of the erosion/dilation structural element. We are also applying this strategy to the problem of vision based self-localization in mobile robots, an application that falls outside the scope of this thesis. The improvements for the detection results based on Associative Morphological Memories would come from the use of color space morphology, which would provide color and structure matching simultaneously.

We feel at the end of this work that, although much has been achieved, there is still a lot of possibilities for exploration and for improvement. We hope that others will find exciting and follow the opportunities for research opened by this work.

Appendix A

Scale-Space: An Algebraic Framework ([58])

Scale-spaces are a central concept in the mathematical foundations of computer vision. In this appendix we briefly review the algebraic framework proposed in [58] that sets the most abstract and general formulation of scale-spaces.

Let f be the image at scale zero and let $T(s)$ be the operator such that $T(s)f$ is the observation at scale s. The family of operators $\{T(s)\}_{s>0}$ is known as a scale-space, also the collection of images $\{T(s)f\}_{s>0}$ is known as a scale-space.

The incremental construction of a scale-space follows from the so-called *atlas principle*. In words, it says that if we observe an image at scale s and take this observation as the input for an observation at the scale t, an observation at scale $r \geq \max(s,t)$ results. The scale is a positive scalar corresponding to the intuitive notion of size. In [58], the generic scale-space operator has the form $T_\psi(t) = S(t)\psi S(t)^{-1}$ subject to the condition that $T_\psi(t)$ obeys the semigroup property.

Let $\Im$ be the family of images under consideration. In what follows, an image will be defined as a mapping, $f : \mathbb{R}^n \rightarrow \mathbb{R}$, i.e.e $\Im = Fun(\mathbb{R}^n)$.

Definition 31 *A one-parameter family $S = \{S(t)\,|t > 0\}$ of operators on $\Im$ is called a scaling (or multiplication) if:*

(i) $S(1) = id$

(ii) $S(t)\,S(s) = S(ts) \qquad s,t > 0$

This means in particular that S is a commutative group and that the inverse of $S(t)$ is given by

$$S^{-1}(t) = S(1/t)\,, t > 0 \tag{A.1}$$

If S_1, S_2 are two scalings whose members commute mutually, i.e. $S_1(t)\,S_2(s) = S_2(s)\;S_1(t)$ for $s,t > 0$, then the composition $\{S_2(s)\,S_1(t)\,|t > 0\}$ is a scaling, too. In particular, if $S(t)$ defines a scaling, then for $p \in \mathbb{R}$,

$$S^p(t) = S(t^p)\,, t > 0 \tag{A.2}$$

does so as well. We denote this scaling by S^p.

The next proposition provides a method to construct new scalings from existing ones.

Proposition 32 *If S is a scaling on $\Im$ and $\lambda : \Im \to \Im$ is an invertible operator, then S_λ given by $S_\lambda = \lambda^{-1}S(t)\,\lambda$, $t > 0$ is also a scaling.*

The two parameters of scalings $S^{p,q}$ given $p, q \in \mathbb{R}$ is defined as:

$$S^{p,q}(t)\,f = t^q f(\cdot/t^p) \quad \text{for } f \in Fun\left(\mathbb{R}^d\right), t > 0 \tag{A.3}$$

includes many of the scalings of interest as special cases. For instance, $p = 1, q = 0$ gives the spatial scaling $S^{1,0}(t)\,f = f(\cdot/t)$.

The *atlas principle* introduced by Koenderink [71], shows how it is possible to build a scale-space incrementally. In other words, if we consider an image at scale s and consider this the input for an observation at scale t, then an image at a some higher scale r should result:

$$T(t)\,T(s) = T(r) \tag{A.4}$$

From a mathematical point of view, this condition is equivalent with the requirement that the family of scale-space operators $T(t)$ is a semigroup.

Definition 33 *Let us consider $I = (0, \infty)$. We say that a binary operator $\dotplus$ on I defines a commutative semigroup, if the following conditions are satisfied:*

(i) $\dot{+}$ is commutative. For $s, t \in I$ we have $s \dot{+} t = t \dot{+} s$

(ii) $\dot{+}$ is associative. For $r, s, t \in I$ we have $(r \dot{+} s) \dot{+} t = r \dot{+} (s \dot{+} t)$

Examples are the linear $s \dot{+} t = s + t$ and supremal $s \dot{+} t = s \vee t$ cases.

Definition 34 *Let us again consider the set $I = (0, \infty)$ with the operator $\dot{+}$. If $(I, \dot{+})$ defines a semigroup and if, furthermore, the following monotonic condition is satisfied:*

$$s \leqslant t \Rightarrow s \dot{+} r \leqslant t \dot{+} r \qquad \text{with } r, s, t \in I \tag{A.5}$$

then we say that $(I, \dot{+}, \leqslant)$ is a linearly ordered semigroup.

Both $(I, +, \leqslant)$ and $(I, \vee, \leqslant)$ are linearly ordered semigroups.

Definition 35 *Let us consider that $(I, \dot{+}, \leqslant)$is a linearly ordered semigroup and S is a scaling on $\Im$. The family $\{T(t)\}_{t>0}$ of operators on $\Im$ is called an $(S, \dot{+}) -$ scale space if the following two conditions are satisfied:*

$$T(t)\, T(s) = T(t \dot{+} s) \qquad \text{with } s, t > 0 \tag{A.6}$$

$$T(t)\, S(t) = S(t)\, T(1) \qquad \text{with } t > 0 \tag{A.7}$$

If now we substitute $T(1)$ by ψ and $T(t)$ by $T_\psi(t)$ (to emphasize the dependence on ψ) then, the relation A.7 can be rewritten as:

$$T_\psi(t) = S(t)\, \psi S^{-1}(t) \qquad \text{with } t > 0 \tag{A.8}$$

In this case we say the operator ψ *induces* the $(S, \dot{+}) - \textit{scale space}\ \{T_\psi(t)\}_{t>0}$. The relation A.8 tells us that the same operator ψ is applied at different scales, ranging from small scales when t is small to large scale when t is large. Depending on the choice of image operator ψ, several types of scale-spaces can result: linear or morphological.

Linear scale-spaces are generated by linear convolution operators on $L^2(\mathbb{R}^d)$ given by:

$$\psi(f) = K \star f = \int_{\mathbb{R}^d} K(\cdot - y)\, f(y)\, dy \tag{A.9}$$

which are linear, translation invariant operators. $K \in L^2\left(\mathbb{R}^d\right)$ is called the *convolution kernel* or *impulse response*. We assume that K is mass-preserving, that is:

$$\int_{\mathbb{R}^d} K(x)\, dx = 1 \tag{A.10}$$

Considering the scaling $S^{p,q}$, the scale-space operator $T_\psi(t) = S(t)\,\psi S^{-1}(t)$ is a linear transformation invariant operator and thus $T_\psi(t)$ is a convolution as well:

$$T_\psi(t)\, f = K_t \star f \tag{A.11}$$

where the convolution kernel takes the form:

$$K_t(x) = t^{-pd} K\left(x/t^p\right). \tag{A.12}$$

The semigroup condition $T_\psi(t)\, T_\psi(s) = T_\psi(t \dot{+} s)$ amounts to the following condition on the kernels K_t :

$$K_t \star K_s = K_{t\ \ s}\,, \qquad\qquad s, t > 0 \tag{A.13}$$

It follows from the study of this equation that the Fourier Transforms of the convolution kernels are of the form:

$$\widehat{K}(\xi) = (2\pi)^{-\frac{d}{2}} \exp\left(-a\,|\xi|^k\right), \quad a, k > 0. \tag{A.14}$$

The well-known Gaussian scale-space follows from the quadratic scale $S(t)\, f(x) = f\left(x/\sqrt{t}\right)$,

$$\widehat{K}(\xi) = (2\pi)^{-\frac{d}{2}} \exp\left(-a\,|\xi|^2\right). \tag{A.15}$$

The corresponding convolution kernel K is the Gaussian function:

$$\widehat{K}(x) = (4\pi a)^{-\frac{d}{2}} \exp\left(-\frac{|x|^2}{4a}\right). \tag{A.16}$$

When $a = \frac{1}{2}$ then the image operator is given by:

$$\psi(f)(x) = (2\pi)^{-\frac{d}{2}} \int f(x-y) \exp\left(-\frac{|y|^2}{2}\right) dy, \tag{A.17}$$

and the induced scale-space is given by:

$$(T_\psi(t) f)(x) = (2\pi t)^{-\frac{d}{2}} \int f(x-y) \exp\left(-\frac{|y|^2}{2t}\right) dy. \tag{A.18}$$

To obtain morphological scale-spaces based on erosion and dilation, first we consider the morphological operators of erosion and dilation $(\varepsilon_b, \delta_b)$ as a translation invariant adjunction on $Fun\left(\mathbb{R}^d\right)$:

$$\varepsilon_b = \bigwedge_{h \in \mathbb{R}^d} [f(x-h) + b(h)] \tag{A.19}$$

$$\delta_b = \bigvee_{h \in \mathbb{R}^d} [f(x+h) - b(h)], \tag{A.20}$$

where function b is the structuring function. Erosion is equivalent to the so-called *infimal convolution* $\varepsilon_b(f)(x) = (f \boxminus b)(x)$. It is assumed that the structuring function b is lower semi-continuous, convex and satisfies $b(x) > -\infty$, for every x. Given a scaling $S = S^{p,q}$ it can be shown that

$$T_\varepsilon(t) f = f \boxminus S(t) b. \tag{A.21}$$

When b is an indicator function i.e. $b = I_B$ then $T_\varepsilon(t)$ is independent of t and we have the flat erosion:

$$\varepsilon(f) = \bigwedge_{h \in B} f(x-h). \tag{A.22}$$

The semigroup condition $T_\varepsilon(t) T_\varepsilon(s) = T_\varepsilon(t \dotplus s)$ amounts to the following condition on the structuring function b:

$$S(t) b \boxminus S(\varepsilon) b = S(t \dotplus s) b \tag{A.23}$$

which leads to the equation:

$$t^q b^*\left(t^{p-q}\xi\right) + s^q b^*\left(s^{p-q}\xi\right) = (t \dotplus s)^q b^*\left(\left((t \dotplus s)^{p-q}\right)\xi\right), \tag{A.24}$$

where $b^*(\cdot)$ is the conjugate of the function $b(\cdot)$.

When $\dot{+}$ is $+$ and $(p = 1, q = 1)$ the above relation is trivially satisfied and $\varepsilon(f) = f \boxminus b$ induces an $\left(s^{1,1}, +\right) -$ *scale space* for every convex function b.

The study of other scale-spaces generated by other parameter settings of the scaling and other semigroup operators is the subject of current research in the mathematical morphology literature.

Appendix B

Methodological Remarks

We have used two set of images as the training and testing databases. The first is a custom database taken using an early model of a digital photographic camera: the Apple Quicktake. The images include several faces of small aspect ratio, each face covers about 10% of the image surface. The images were taken in front of a smooth surface, with flash at a similar distance from the subjects. The original color images were reduced to gray scale. In the image below, some instances from this database are depicted in the figure B-1.

The second database is the one from Carnegie Mellon University. It is a collection of images from several sources, of several sizes and showing faces at several scales. It was used first by [114] to test their face localization system, and has been used as benchmark by many other researchers (see [26], for instance). The database consists of four datasets, called Test Set A, Test Set B and Test Set C and the Rotated Test Set. Test Set B was provided by Kah-Kay Sung and Tomaso Poggio from the AI Lab at MIT, and Test Sets A and C were collected at CMU. The test sets A, B and C altogether form the Upright Test Set. This test set consists of 130 images, showing in total 511 faces at different scales and with very slight rotations (up to 10^o), wearing beards and glasses. These images were taken from the World Wide Web, shots from TV and scanned photographs. It contains images with single/multiple face/s per image or no faces at all. The Rotated Test Set is built up from 50 images, containing 223 faces, at which 210 are at angles of more than 10^o from upright. We have made no use of Rotated Test Set.

Some instances from the CMU database used in our experiments are shown in the image

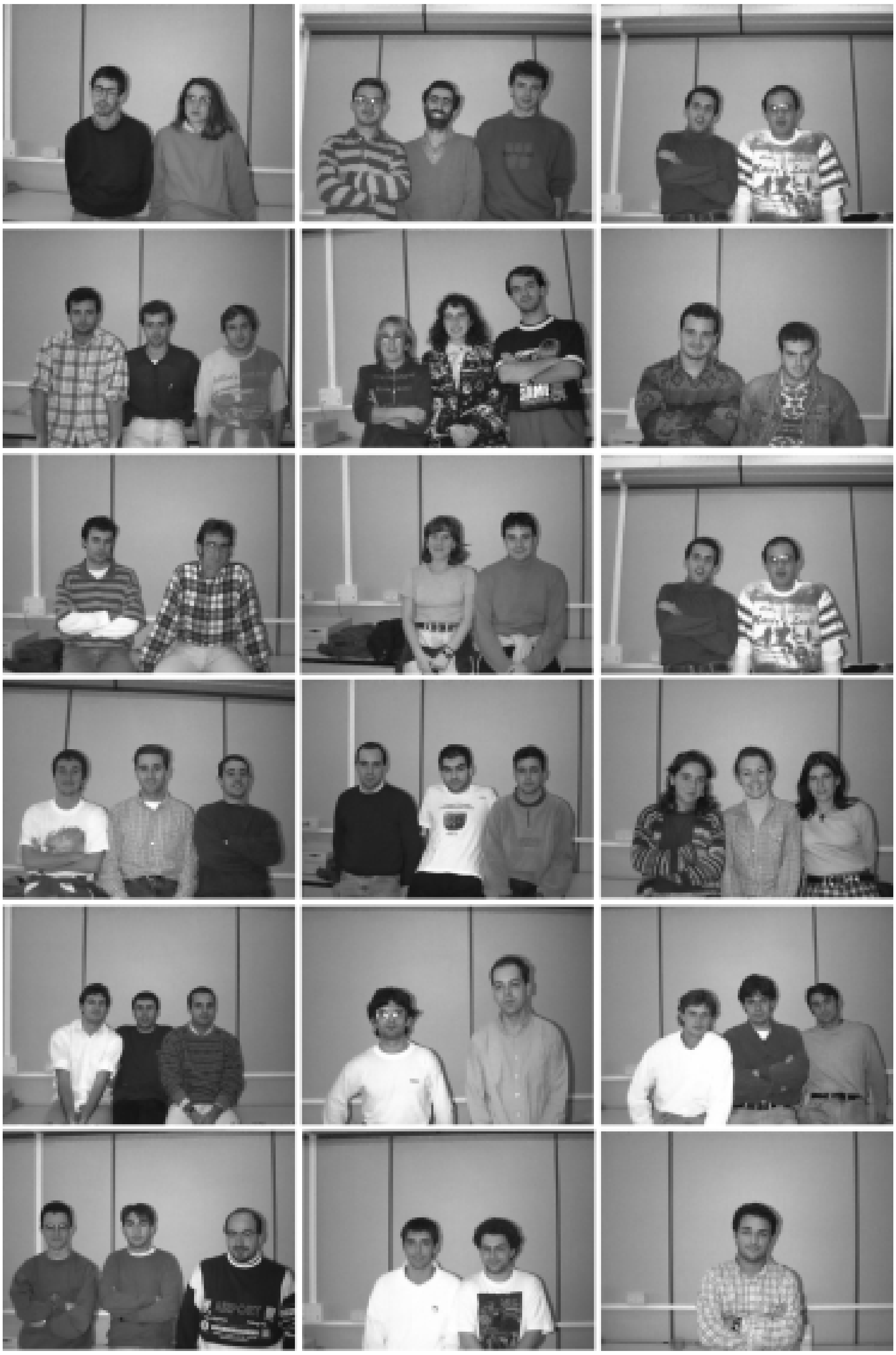


Figure B-1: Some instances from our database

B-2.

Most of the results found in the literature are given on the basis of the number of correct detected faces and false positives.. In our work, the results refer to the classification of the image on a pixel basis. Therefore, visual results show the face pixels highlighted and the numerical results are usually given as Receiving Characteristic Operator (ROC) that relate the false positive and true positive rates depending on the threshold value used for the decision. Most of the times, our experiments refer to a single scale detection. Scale-invariant detection can be achieved through a multi-resolution pyramid approach. The ground truth for the evaluation of the algorithms is given the hand-labeling of face pixels. The labeling is given by rectangular shapes defined over images.

As the emphasis in this thesis is on the naive and straighforward construction of the face detection systems, no sophisticated pattern selection procedures have been applied. Neither preprocessing of images nor boostrapping strategy have been used. We call training set the set of images used to extract the face patterns. The face pattern extraction is simply the selection (with the mouse) of a point near the center of the face. The extracted face is the square of fixed size sorrounding this point. There is no procedural relation between the ground truth determined on the image for evaluation purposes and the set of face patterns. The same face is slightly different when considered as the source of a face pattern and when considered as the ground truth for evaluation. This difference is not intentional, it is only the fruit of performing separate procedures for these tasks.

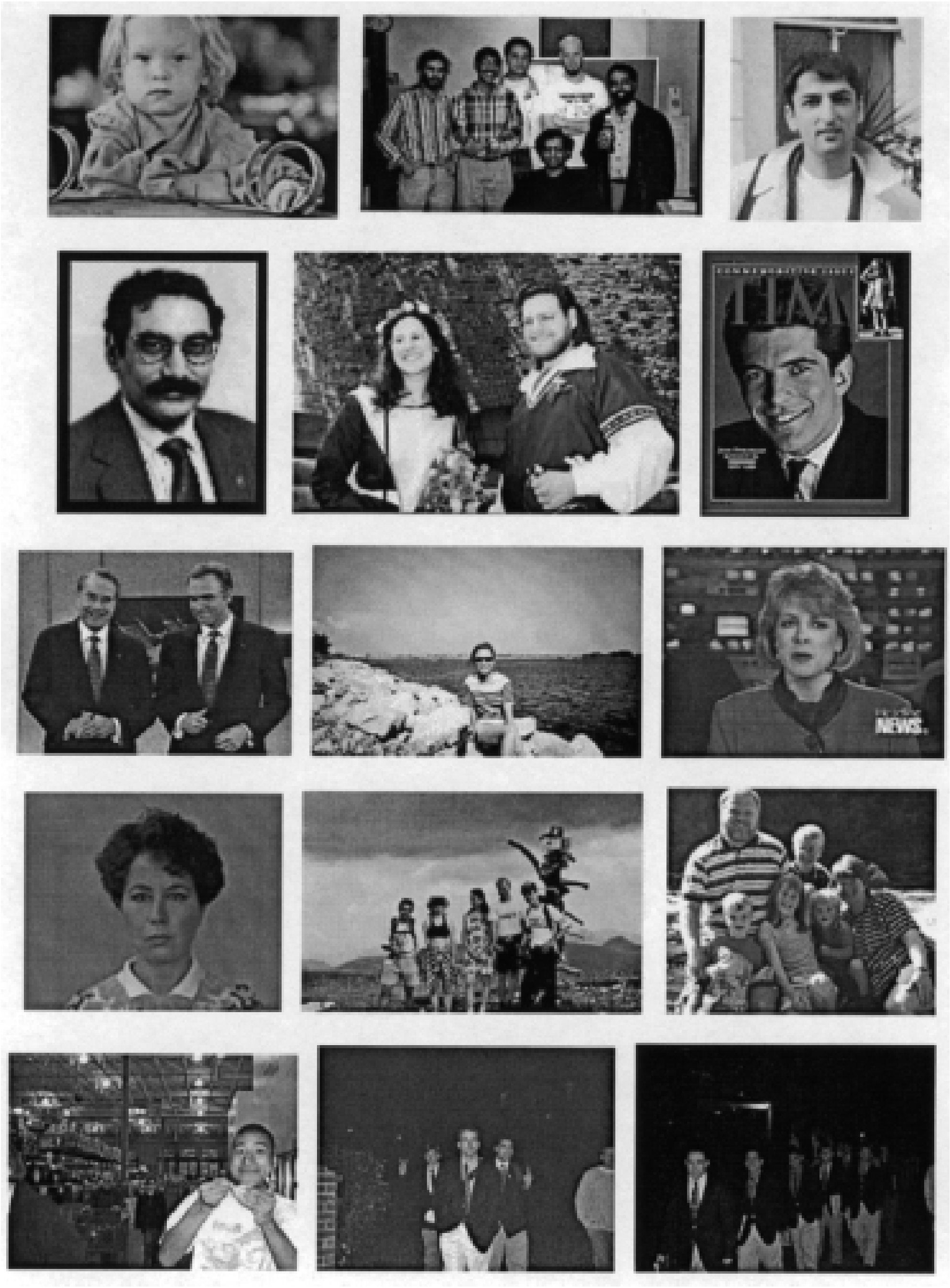


Figure B-2: Some instances from the CMU database

Bibliography

[1] *AIBO homepage.* Available on-line at: http://www.aibo.com

[2] Alvarez, L. and J. M. Morel. Formalization and computational aspects of image analysis. *Acta Numerica*, pp.1-59, 1994

[3] *Ananova Homepage.* Available on-line at: http://www.ananova.com

[4] Arons, B.. A review of the cocktail party effect. *Journal of The American Voice I/O Society*, 12:35-50, 1992

[5] Bakker, P. and Y. Kuniyoshi. Robot see, robot do: an overview of robot imitation. *Proc. of AISB'96 Workshop on Lerning in Robots and Animals*, pp. 3-11, Brighton, United Kingdom, 1996

[6] Bala, L.-P., K. Talmi and J. Liu. Automatic detection and tracking of faces and facial features in video sequences. *1997 Picture Coding Symposium*, pp. *unavailable,* Berlin, Germany, 1997

[7] Ballard, P. and G. C. Stockman. Controlling a Computer via Facial Aspect. *IEEE Trans. on Systems, Man, and Cybernetics*, 25(4):669–677, 1995

[8] Banon, G. J. F. and J. Barrera. Minimal representation for translation invariant set mappings by mathematical morphology. *SIAM J. Appl. Math.*, 51:1782-1798, 1991

[9] Basu, S., N. Oliver and A. Pentland. 3D Modeling and tracking of human lip motions. *Proc. of Int'l Conf. on Computer Vision* , pp. not available, Bombay, India, January 4-7, 1998

[10] Betke, M. and J. Kawai. Gaze detection via selforganizing gray-scale units. *Proc. of the Int'l Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pp. 70-76, Kerkyra, Greece, 1999

[11] Bloomberg, D. S. and L. Vincent. Pattern matching using the blur hit-miss transform. *Journal of Electronic Imaging*, 9(2):140-150, 2000

[12] Bobick, A. F., S. S. Intille, J. W. Davis, F. Baird, C. S. Pinhanez, L. W. Campbell, Y. A. Ivanov, A Schütte and A. Wilson. The Kidsroom. *Communications of the ACM*, 43(3):60-61, 2000

[13] Boomgaard, R. v. d.. *Mathematical morphology: extensions towards computer vision.* Ph D thesis, University of Amsterdam, The Netherlands, 1992

[14] Boomgaard, R. v. d. and A. W. M. Smeulders. The morphological structure of images: the differential equations of morphological scale-space. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16:1101-1113, 1994

[15] Boomgaard, R. v. d., A. W. M. Smeulders and J. G. M. Schavemaker. Image segmentation in morphological scale-space. *Proc. of Int'l Symposium on Mathematical Morphology* , pp. 15-16, 1994

[16] Brockett, R. and P. Maragos. Evolution equations for continous-scale morphological filtering. *IEEE Trans. on Signal Processing*, 42(12):3377-3386, 1994

[17] Bruce, V. and A. Young. Understanding face recognition. *British Journal of Psychology*, 77:305-327, 1986

[18] Brunelli, R. and D. Falavigna. Person identification using multiple cues. *IEEE Transactions on PAMI*, 17(10):955-966, 1995

[19] Brunelli, R., D. Falavigna, T. Poggio and L. Stringa. Automatic person recognition by acoustic and geometric features. *Machine Vision and Applications*, 8:317-325, 1995

[20] Brunelli, R. and O. Mich. SpotIt! An interactive identikit system. *Graphical models and image processing*, 58(5):399-404, 1996

[21] Brunnström, K., J.O. Eklundth and T. Uhlin. Active fixation for scene exploration. *International Journal of Computer Science*, 17:137-162, 1996

[22] Bülthoff, H. H., S. Y. Edelman and M. J. Tarr. How are three-dimensional objects represented in the brain? Available as *AI Memo No. 1479*, MIT, 1994

[23] Burl, M. C., T. K. Leung and P. Perona. Face localization via shape statistics. *Proc. of First Int'l. Workshop on Automatic Face and Gesture Recognition*, pp. *unavailable,* Zurich, Switzerland, 1995

[24] Chelappa, R., C. L. Wilson and A. Sirohey. Human and machine recognition of faces: A survey. *Proc. of IEEE*, 83(5):704-740, 1995

[25] Chen, M.-H. and P.-F. Yan. A multiscaling approach based on morphological filtering. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11:694-700, 1989

[26] Colmenarez, A. J. and T. S. Huang. Face detection with information-based maximum discrimination. *Proc. of Int'l Conf. on Computer Vision and Pattern Recognition*, pp. 782-787, Puerto Rico, 1997

[27] Crimmins, T. R. and W. R. Brown. Image algebra and automatic shape recognition. *IEEE Trans.onAerospaceElectron.SystemsAES-21*, pp. 66-69, 1985.

[28] Darrell, T., I. Essa and A. Pentland. Task-specific gesture analysis in real-time using interpolated views. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(12):1236-1242, 1996

[29] Daugman, J. *How iris recognition works.* Available on-line at: http://www.cl.cam.ac.uk/~jgd1000/irisrecog.pdf

[30] Daugman, J. Wavelet demodulation codes, statistical independence and pattern recognition. *Proc. of 2nd IMA-IP Conf. (Institute of Mathematics and its Applications)*, pp.244-260, London, UK, 1999

[31] Daugman, J. Biometric decision landscapes. Available as: *Technical Report TR482*, Computer Laboratory, University of Cambridge, 1999

[32] Dennett, D. *Brainchildren: Essays on Designing Minds.* Cambridge (Massachusetts), MIT Press, USA, 1998

[33] Dougherty, E. R. Optimal mean-absolute-error filtering of gray-scale signals by the morphological Hit-or-Miss Transform. *Journal of Mathematical Imaging and Vision*, 4:255-271, 1994

[34] Dougherty, E. R. and D. Zhao. Model-based characterization of statistically optimal design for morphological shape recognition algorithms via the Hit-or-Miss Transform. *Journal of Visual Communication and Image Representation*, 3(2):147-160, 1992

[35] Dougherty, E. R. and J. T. Astola (eds.). *Nonlineal Filters for Image Processing*, SPIE Press/IEEE Press, 1999

[36] Dror, I. E., F. L. Florer, D. Rios and M. Zagaeski. Using artificial bat sonar neural networks for complex pattern recognition: recognizing faces and the speed of a moving target. *Biological Cybernetics*, 74:331-338, 1996

[37] DuChateau, P. and D. W. Zachmann. *Theory and Problems of Partial Differential Equations.* Schaum's Outline Series, McGraw Hill, New York, USA, 1986

[38] Duda, R. O. and P. E. Hart. *Pattern Classification and Scene Analysis*,Wiley, New York, 1973

[39] *eTrue.* Homepage at: http://www.etrue.com

[40] Ekman, P. and W. V. Friesen. *Facial Action Coding System.* Consulting Psychologists Press, Palo Alto (CA), USA, 1978

[41] Ekman, P. *Emotion in the Human Face.* Cambridge University Press, Cambridge, United Kingdom, 1982

[42] Eshera, M. A. and K.S. Fu. A graph distance measure for image analysis. *IEEE Trans. on SMC,* 14(3):398-408, 1984

[43] Essa, I. A. Computers seeing people. *AI Magazine*, 20(2):69-82, 1999

[44] Essa, I. and A. Pentland. Facial Expression recognition using a dynamic model and motion energy. *Proc. of Int'l Conf. on Computer Vision*, pp. 360-367, Washington D.C., USA, 1995

[45] Fischler, M. A. and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. on Computers*, vol. COM-22, pp. 67-92, 1973

[46] Fisher, R. B. Is computer vision still AI?. *Lecture Notes in Computer Science*, 1159:239-253, 1996

[47] Fukunaga, K. *Introduction to statistical pattern recognition.* Academic Press, 1990

[48] Gardner, H. *Frames of Mind: The Theory of Multiple Intelligences.* Basic Books, New York, 1993

[49] Gee, A. and R. Cipolla. Non-intrusive gaze tracking for human-computer interaction. *Proc. of Mechatronics and Machine Vision in Practice*, pp.112-117, Toowoomba, Australia, 1994

[50] Giardina, C. R. and E. R. Dougherty. *Morphological Methods in Image and Signal Processing.* Prentice Hall, New Jersey, USA, 1988

[51] Gold, S and A. Ramgarajam. A graduated assignment algorithm for grap matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(4):377-388, 1996

[52] Goleman, D. *Emotional Intelligence.* Bantam, New York, USA, 1995

[53] Gonzalez, R. C. and R. E. Woods. *Digital Image Processing.* Addison-Wesley, New York, 1992

[54] Guichard, F. and J.-M. Morel. Image iterative smoothing and P.D.E.'s. *Notes de Cours Du Centre Émile Borel*, Institut Henri Poincaré, Paris, France, 14 septembre 1998 - 18 decémbre 1998

[55] Haralick, R. M. and L. G. Shapiro. *Computer and Robot Vision.* Addison-Wesley, New York, 1992

[56] Haro, A., M. Flickner and I. Essa. Detecting and Tracking Eyes By Using Their Physiological Properties, Dynamics, and Appearance. Available as: *Technical Report GIT-GVU-TR-99-46*, College of Computing, Georgia Institute of Technology, 1999

[57] Healey, J. and R. Picard. StartleCam: a cybernetic wearable camera. *Proc. of the Int'l Symposium on Wearable Computers*, pp. not available, Pittsburgh, USA, October 1998

[58] Heijmans, H. J. A. M. and R. van den Boomgaard. Algebraic framework for linear and morphological scale-spaces. Available as *Technical Report PNA-R0003*, CWI, Amsterdam, The Netherlands, February 2000

[59] *History of the AI.* (spanish only). Available on-line at http://www-gth.die.upm.es/~ macias/doc/pubs/aircenter99/www.aircenter.net/19601970.html

[60] Hong, L. and A. Jain. Integrating faces and fingerprints for personal identification. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(12):1295-1307

[61] *Iridian Technologies.* Homepage at http://www.iridiantech.com

[62] Jackway, P. T. Morphological scale-space. *Proc. of 11th IAPR Int'l Conf. on Pattern Recognition*, IEEE Computer Society Press, pp. 252-255, 1992

[63] Jackway, P. T. Morphological scale-space with application to three-dimensional object recognition. *Ph. D. Thesis*, Queensland University of Technology, Australia, 1994

[64] Jang, B. K. and R. T. Chin. Shape analysis using morphological scale-space. *Proc. of 25th Ann. Conf. on Information Sciences Systems*, pp.1-4, 1991

[65] Juell, P. and R. Marsh. A hierarchical neural network for human face detection. *Pattern Recognition*, 29(5):781-787, 1996

[66] Kälviäinen H. and E. Oja. Comparisons of attributed graph matching algorithms for computer vision. *Proc. of STEP-90, Finnish Artificial Intelligence Symposium*, pp.354-368, Oulu, Finland, June 1990

[67] Kanade, T. *Computer Recognition of Human Faces.* Birkhuser Verlag, Stuttgart, Germany, 1977

[68] Kelly, K. *Out of Control.* Addison-Wesley, New York, USA, 1994

[69] Khosravi, M. and R. W. Schafer. Template matching based on a gray scale Hit-or-Miss Transform. *IEEE Trans. on Image Processing* , 5(6):1060-1066, 1996

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

[71] Koenderink, J. The structure of images. *Biological Cybernetics*, 50:363-370, 1984

[72] Kondo, T. and H. Yan. Automatic human face detection and recognition under non-uniform illumination. *Pattern Recognition*, 32:1707-1718, 1999

[73] Kotropoulos, C., A. Tefas and I. Pittas. Frontal face authentication using morphological elastic graph matching. *IEEE Trans. on Image Processing*, 9(4):555-560, 2000

[74] Kotropoulos, C., A. Tefas and I. Pitas. Morphological elastic graph matching applied to frontal face authentication under well-controlled and real conditions. *Pattern Recognition*, 33:1935-1947, 2000

[75] Kuniyoshi, Y., M. Inaba and H. Inoue. Learning by watching: extracting reusable task knowledge from visual observation of human performance. *IEEE Trans. on Robotics and Automation*, 10(6):799-822, 1994

[76] Lades, M., J. C. Vorbrüggen, J. Buhmann, J. Lange, C. v. d. Malsburg, R. P. Würtz and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. on Computers* , 42(3):300-311, 1993

[77] Lanitis, A., C. J. Taylor and T. F. Cootes. Automatic face identification system using flexible appearance models. *Image and Vision Computing*, 13(5):393-401, 1995

[78] Lee, C. H., J. S. Kim and K. H. Park. Automatic human face location in a complex backgroundusing motion and color information. *Pattern Recognition*, 29(11):1877-1889, 1996

[79] Leung, T. K., M. C. Burl and P. Perona. Finding faces in cluttered scenes using random labeled graph matching. *Proc. of 5th Int'l. Conf. on Computer Vision*, IEEE Computer Society Press, pp. 637-644, 1995

[80] Lew, M. S. and N. Huijsmans. Information theory and face detection. *Proc. of Int'l Conference on Pattern Recognition*, pp.601-605, Vienna, Austria, 1996

[81] Li, S. Z. Matching: invariant to translations, rotations and scale changes. *Pattern Recognition*, 25(6):583-594, 1992

[82] Lin, C.-H. and J.-L. Wu. Automatic facial feature extraction by genetic algorithms. *IEEE Trans. on Image Processing*, 8(6):834-844, 1999

[83] Lin, S.-H., S.-Y. Kung and L.-J. Lin. Face recognition/detection by probabilistic decision-based neural network. *IEEE Trans. on Neural Networks*, 8(1):114-132, 1997

[84] Littmann, E. and H. Ritter. Adaptive color segmentation - a comparison of neural and statistical methods. *IEEE Trans. on Neural Networks*, 8(1):175-185, 1997

[85] Littmann, E., A. Drees and H. Ritter. Robot guidance by human pointing gestures. *Proc. of NICROSP*, pp. not available, 1996. Available online at: ftp://neuro.informatik.uni-ulm.de/ni/enno/littmann.nicrosp96.ps.gz

[86] Lindeberg, T. *Scale-Space Theory in Computer Vision.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994

[87] Luckman, A. J., N. M. Allison, A. W. Ellis and B. M. Flude. Familiar face recognition: a comparative study of a connectionist model and human performance. *Neurocomputing*, 7:3-27, 1995

[88] Marcone, G., A. Fusi, G. Stoppani and G. Orlandi. Human face recognition: automatic face detection. *Proc. of 4th Bayona (Pre COST #254) Workshop on Intelligent Methods in Signal Processing and Communications*, pp.75-80, Bayona-Vigo, 1996

[89] Marrin, T. and R. W. Picard. Analysis of affective musical expression with the conductor's jacket. *Proc. of the XII Colloquium on Musical Informatics*, pp. not available, Gorizia, Italy, 1998

[90] Mase, K. and A. Pentland. Automatic lipreading by optical-flow analysis. *Systems and Computer in Japan*, 22(6):67–76, 1991

[91] Mase, K. Recognition of facial expression from optical flow. *IEICE Transactions*, *Special issue on Computer Vision and its Applications*, E. 74(10):3474–3483, 1991

[92] Matheron, G. *Random Sets and Integral Geometry.* John Wiley and Sons, New York, USA, 1975

[93] McCarthy, J. What is Artificial Intelligence. *Notes on AI*, Available on-line at: http://www-formal.stanford.edu/jmc/whatisai.html

[94] Mckevitt, P. and J. G. Gammack. The sensitive interface. *Artificial Intelligence Review*, 10:275-298, 1996

[95] Moghaddam, B. and A. Pentland. An automatic system for model-based coding of faces. Available as: *Technical Report TR-317*, Perceptual Computing Group, MIT Media Laboratory, 1995

[96] Moghaddam, B. and A. Pentland. Probabilistic visual learning for object representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):696-710, 1997

[97] Nam, Y. and K. Wohn. Recognition of SpaceTime Hand-Gestures using HMMs. *Proc. of Int'l Workshop on Integration of Gestures in Language and Speech*, pp. 175-184, 1996

[98] Ohnishi, N. and N. Sugie. Visual-auditory interfaces for machines that serve humans. *Robotics and Autonomous Systems*, 18:243-249, 1996

[99] Pankanti, S., R. M. Bolle and A. Jain. Biometrics: the future of identification. *Computer*, 33(2):46-49, 2000

[100] Park S. H., I. D. Yun and S. U. Lee. Color image segmentation based on 3-D clustering: morphological approach. *Pattern Recognition*, 31(8):1061-1076, 1998

[101] Pavlovic, V. I., R. Sharma and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction: a review. *IEEE Trans. on Patterns Analysis and Machine Intelligence*, 19(7):677-695, 1997

[102] Pearson, D. E. Developments in model-based video coding. *Proceedings of the IEEE*, 83(6):892-906, 1995

[103] Pelillo M., K. Siddiqi and S. W. Zucker. Matching hierarchical structures using association graphs. *IEEE Trans. on Pattern Analysis and Machine Intelligence,* 21(11):1105-1119, 1999

[104] Penev, P. S. and J. J. Atick. Local feature analysis: a general statistical theory for object representation. *Network: Computation in Neural Systems*, 7:477-500, 1996

[105] Pentland, A. Looking at people: sensing for ubiquitous and wearable computing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(1):107-119, 2000

[106] Pentland, A., R. Picard and P. Maes. Smart rooms, desks, and clothes: toward seamlessly networked living. *British Telecommunications Engineering*, 15:168-172, 1996

[107] Pfeifer, R. and P. Verschure. The challenge of autonomous agents: pitfalls and how to avoid them. *Proc. on Int'l Workshop on Autonomous Agents*, pp. not available, Corsendonk, Belgium, 1991

[108] Picard, R. W. *Affective Computing.* MIT Press, USA, 1997

[109] Potamianos, G., H.P. Graf, and E. Cosatto. An image transform approach for HMM based automatic lipreading. *Proc. of Int'l. Conf on Image Processing*, pp. not available, Chicago, USA, 1998

[110] Quek, F. K. H. Eyes in the interface. *Image and Vision Computing*, 13(6):511-525, 1995

[111] *Prosopagnosia Homepage.* Available on-line at http://www.anything-balloons.com/glenn

[112] Ritter, G. X. and P. Sussner, J.L. Diaz de Leon. Morphological Associative Memories. *IEEE Transactions on Neural Networks*, 9(2):281-292, 1998

[113] Ritter, G. X. and P. Sussner. An Introduction to Morphological Neural Networks. *Proc. of Int'l Conference on Pattern Recognition*, pp. 709-717, 1996

[114] Rowley, H. A., S. Baluja and T. Kanade. Neural network-based face detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence.* 20(1):23-38, 1998

[115] Rowley, H. A., S. Baluja and T. Kanade. Rotation invariant neural network-based face detection. Available as: *Technical Report CMU-CS-97-201*, School of Computer Science, Carnegie Mellon University, December 1997

[116] Rybak, I. A., V. I. Gusakova, A. V. Golovan, L. N. Podladchikova and N. A. Shevtsova. A model of attention-guided visual perception and recognition. *Vision Research*, 38:2387-2400, 1998

[117] Saber, E. and A. M. Tekalp. Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions. *Pattern Recognition Letters*, 19:669-680, 1998

[118] Serra, J. *Image Analysis and Mathematical Morphology* . Academic Press, London, United Kingdom, 1982

[119] Sharpe, J. P., N. Sungar and K. M. Johnson. Adaptive Resonance Theory and self-organizing morphological kernels. *SPIE Vol. 2300*, pp. 37-45, 1994

[120] Sieberg, D. (CNN Tech Editor). *Iris recognition at airports uses eye-catching technology.* Available on-line at: http://www.cnn.com/2000/TECH/computing/07/24/iris.explainer/index.html

[121] Soille, P. Morphological Image Analysis. *Principles and Applications*, Springer-Verlag, 1999

[122] Somaie, A. A. and S. S. Ipson. A human face profile identification system using 1-D real Fourier descriptors. *Int'l Journal of Infrared and Millimeter Waves*, 16(8):1285-1298, 1995

[123] Srihari, R. K. Use of captions and other collateral text in understanding photographs. *Artificial Intelligence Review*, 8:409-430, 1994-1995

[124] Sung, K.-K. and T. Poggio. Example-based learning for view-based human face detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence.* 20(1):39-51, 1998

[125] Takahashi, K., T. Sakaguchi, T. Minami and O. Nakamura. Description and matching of density variation for personal identification through facial images. *Proc. of SPIE: Visual Communication and Image Processing*, 1360:1694-1704, 1990

[126] Toyama, K. "Look, Ma - No Hands!" - Hands-free cursor control with real-time 3D face tracking. *Proc. of Int'l Workshop on Perceptual User Interfaces*, pp. not available, San Francisco, California, USA, 4-6 Nov. 1998

[127] Troje, N. F. and H. H. Bülthoff. Face recognition under varying poses: the role of texture and shape. *Vision Research*, 36(12):1761-1771, 1996

[128] Turing, A. Computing Machinery and Intelligence. *Mind*, 59(236):433-460, 1950

[129] Turk, M. and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71-86, 1991

[130] Turk. M. and G. Robertson. Perceptual user interfaces. *Communications of the ACM*, 43(3)33-34, 2000

[131] Vilaplana, V., F. Marqués, P. Salembier and L. Garrido. Region-based segmentation and tracking of human faces. *Proc. of 9th European Signal Processing Conf. EUSIPCO'98*, pp. 311-314, Rhodes, Greece, September 1998

[132] *Visionics.* Homepage at: http://www.visionics.com

[133] Waibel, A., M. T. Vo, P. Duchnowski and S. Manke. Multimodal interfaces. *Artificial Intelligence Review*, 10:299-319, 1996

[134] Wang, J.-G. and E. Sung. Frontal-view face detection and facial feature extraction using color and morphological operations. *Pattern Recognition Letters*, 20:1053-1068, 1999

[135] Wang, J. and T. Tan. A new face detection method based on shape information. *Pattern Recognition Letters*, 21:463-471, 2000

[136] Wechler, H.. P. J. Philips, U. Bruce, F. Fogelman Soulié, T.S. Huang (eds) Face rocognition: From theory to applications. Springer Verlag, 1998

[137] Weickart, J., S. Ishikawa and A. Imiya. Linear scale-space has first been proposed in Japan. *Journal of Mathematical Imaging and Vision*, 10(3):237-252, 1997

[138] Witkin, A. P. Scale-space filtering. *Proc. of Int'l Joint Conf. on Artificial Intelligence.* Morgan-Kaufmann (eds.), pp. 1019-1022, Palo Alto, California, USA, 1983

[139] Wong, C., D. Kortenkamp and M. Speich. A mobile robot that recognize people. *Proc. of 7th IEEE Int'l. Conf. on Tools with Artificial Intelligence*, pp. not-available, Washington, USA, 1995

[140] Wong, E. K. Model matching in robot vision by subgraph isomorphism. *Pattern Recognition*, 25(3):287-303, 1992

[141] Wren, C., A. Azarbayejani, T. Darrell and A. Pentland. Pfinder: Real-time tracking of human body. *IEEE Trans. on Pattern Analysis and Machine Intellgence*, 19(7):780-785

[142] Yacoob, Y. and L. Davis. Computing spatio-temporal representations of human faces. *Proc. of Int'l. Conf. on Computer Vision and Pattern Recognition*, pp. 70-75, Washington D.C., USA, 1994

[143] Yang, G. and T. S. Huang. Human face detection in a complex background. *Pattern Recognition*, 27(1):53-63, 1994

[144] Yang, M. H. *Hand Gesture Recognition and Face Detection in Images.* Ph.D. Thesis, Computer Science Department, University of Illinois at Urbana-Champaign, July 2000. Available on line as Technical Report UIUCDCS-R-2000-2151 at ftp://ftp.cs.uiuc.edu/pub/dept/tech_reports/2000/

[145] Yokoo, Y. and M. Hagiwara. Human faces detection method using genetic algorithms. *Proc. of Int'l Conf. on Evolutionary Computation*, pp.113-118, Nagoya, Japan, 1996

[146] Yoo, T.-W. and I.-S. Oh. A fast algorithm for tracking human faces based on chromatic histograms. *Pattern Recognition Letters*, 20:967-978, 1999

[147] Yow, K. C. and R. Cipolla. Finding initial estimates of the human face location. *Proc. of 2nd Asian Conf. on Computer Vision*, vol. 3, pp. 514–518, Singapore, 1995

[148] Yu, K., X. Jiang and H. Bunke. Face recognition by facial profile analysis. *Proc. of First Int'l. Workshop on Automatic Face and Gesture Recognition*, pp. 208 - 213, Zurich, Switzerland, 1995

[149] Zhao, D. and D. G. Daut. Morphological Hit-or-Miss Transformation for shape recognition. *Journal of Visual Communication and Image Representation*, 2(3):230-243, 1991