Dipartimento di Informatica, Bioingegneria,
Robotica ed Ingegneria dei Sistemi

Computer Vision methods as an aid to visually impaired users

by

Giovanni Fusco

Theses Series **DIBRIS-TH-2013-03**

DIBRIS, Università di Genova
Via Opera Pia, 13 16145 Genova, Italy http://www.dibris.unige.it/

Università degli Studi di Genova

Dipartimento di Informatica, Bioingegneria,
Robotica ed Ingegneria dei Sistemi

Dottorato di Ricerca in Informatica

Ph.D. Thesis in Computer Science

Computer Vision methods as an aid to visually impaired users

by

Giovanni Fusco

February, 2013

Dottorato di Ricerca in Informatica
Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi
Università degli Studi di Genova

DIBRIS, Univ. di Genova
Via Opera Pia, 13
I-16145 Genova, Italy
http://www.dibris.unige.it/

Ph.D. Thesis in Computer Science (S.S.D. INF/01)

Submitted by Giovanni Fusco
DIBRIS, Univ. di Genova
giovanni.fusco@unige.it

Date of submission: February 2013

Title: Computer Vision methods as an aid to visually impaired users

Advisor: Francesca Odone
Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi
Università di Genova
odone@disi.unige.it

Ext. Reviewers:
Massimo Ferri
Dipartimento di Matematica
Università di Bologna
massimo.ferri@unibo.it

Roberto Manduchi
Department of Computer Engineering
University of California Santa Cruz
manduchi@soe.ucsc.edu

Abstract

Visually impaired people experience difficulties in some of the activities of their everyday life, especially in unfamiliar settings due to the impossibility to visually comprehend the surrounding environment. Some of these limits are insurmountable and often people modify their life to adapt to this particular condition, limiting their actions to prevent failures and frustrations. Even in cases in which visually impaired people become self-sufficient in moving and orienting, there remains the difficulty in perceiving the presence of known people in crowded or quiet environments and detecting and interpreting road signs, posters and other spatial references. It follows that the availability of tools capable of performing scene analysis could represent a great improvement in the users' quality of life. In this thesis we focus on the development of applications to support visually impaired people to overcome two of what we believe are the most important issues for their safety and social life: the recognition of known people and the localization of text in the surrounding environment.

Detecting the presence of people and recognizing them or verifying a declared identity would improve both safety and social interactions, while being able to detect text in images makes possible the extraction of a useful source of information improving the understanding of the environment surrounding the visually impaired user.

Despite the vast scientific literature on the topics, both face recognition and text detection are considered challenging problems, especially in a context in which there are no constraints on the data. A variety of factors cause these tasks to remain problematic. Concerning face recognition, a general purpose application has to deal with the motion of the subjects, changes in the lighting conditions and the low quality of the images typically adopted in this context. Regarding text detection, the automatic extraction of text is an extremely challenging problem given the variations of text due to different style, orientation, alignment and scale together with low contrast images and complex textured backgrounds.

Although both problems deal with the classification of objects characterized by dis-

tinctive features, the large variations that affect them make pattern recognition a difficult task to perform on these noisy data.

In this thesis we cope with the problem of recognizing faces in unconstrained settings, proposing a real-time video-to-still face recognition framework that combines elements of the face retrieval paradigm with an architecture derived from an identity verification context, while exploiting the availability of the temporal component in videos. Moreover we propose a novel multi-class feature selection method for the detection of the smallest subset of the most discriminative features unique for all known identities, with the effect of reducing the amount of noise in the data while enhancing the recognition rate.

Concerning the problem of the text detection, we apply inference on the trees of the extremal regions in order to enhance the localization performance, exploiting the hierarchical relationships we define between the connected components in the image.

Table of Contents

Chapter 1 Introduction **6**

Chapter 2 Image-based Assistive Technologies **9**

2.1 Introduction 9

2.2 Technologies for Object Recognition 10

2.3 Technologies for Face Recognition 13

2.4 Technologies for Optical Character Recognition 14

2.5 Technologies for Independent Mobility and Wayfinding 16

2.5.1 Pedestrian Guidance 16

2.5.2 Independent Mobility 17

2.5.3 Vision-Based Electronic Travel Aids 18

2.6 Limits of Assistive Technologies 19

Chapter 3 Methods for Face Recognition **21**

3.1 Introduction 21

3.2 A Face Recognition Pipeline 22

3.3 Face Detection 24

3.4 Face Recognition 24

3.4.1 Holistic Approaches 25

3.4.2 Component-based Approaches 28

3.4.3 Features-based methods 31

3.4.4 Face Recognition from Video 35

Chapter 4 Face Recognition Combining Retrieval and Classification 38

4.1 The Proposed System 38

4.2 Face Detection 40

4.3 Individual Models 41

4.3.1 Choosing the Gallery 41

4.3.2 Models for Face Classification 43

4.4 Module for Face Retrieval 44

4.5 Identity Validation 46

4.6 Experimental assessment 46

4.6.1 MOBO dataset 47

4.6.2 R309 dataset 48

4.6.3 Experimental analysis 49

4.7 Discussion 52

Chapter 5 Simultaneous feature selection in face recognition 55

5.1 Introduction 55

5.2 State of the art 56

5.3 Regularized multi-class feature selection 57

5.3.1 LASSO 58

5.3.2 Group LASSO 59

5.3.3 Multi-class Group LASSO 59

5.4 The proposed pipeline 60

5.5 Experimental analysis 62

5.6 Discussion 67

Chapter 6 Text Spotting Via Inference on the Extremal Region Tree 68

6.1 Introduction 68

6.2 State of the Art 70
6.2.1 Region-based methods 70
6.2.2 Sliding Window-Based Methods 73
6.3 The Proposed Method 74
6.4 ER Classification on Trees 75
6.4.1 Notation 76
6.4.2 Algorithm 1: Inference on ERT 77
6.4.3 Algorithm 2: Winner-Take-All (WTA) 80
6.4.4 Algorithm 3: Neumann-Matas (NM) 80
6.4.5 Features 80
6.5 Experiments and Results 82
6.5.1 Computing the Likelihood Terms 82
6.5.2 Experiments on the ICDAR dataset 82
6.6 Discussion 87

Chapter 7 Conclusions **88**

Bibliography **90**

Chapter 1

Introduction

In unfamiliar settings, visually impaired people experience difficulties accessing the visual information which comprises the vast majority of information in public environments. The availability of tools capable of performing scene analysis would hence represent a great improvement in the users' safety and quality of life. This thesis, started in collaboration with Istituto Chiossone [1] and advanced through discussions with social workers and a users' group of the institute, addresses two computer vision problems of great importance for this community of users face recognition, which would allow users to identify known people in an unfamiliar environment, and text detection in unconstrained environments, which would give users access to useful information, such as road signs, posters and other spatial references, thereby improving users' safety. Besides their application in this domain, these two problems are also challenging open issues for the computer vision community because of the high variability in the appearance of the data that characterize both contexts. Moreover, image-based assistive technologies for visually impaired users based on mobile devices increases the difficulty of these problems. The use of mobile cameras for the acquisition of images, in fact, makes the development of robust image-based applications even more challenging, because of the high variability of the data due to different lighting conditions, unconstrained points of view, perspective distortions, presence of clutter, low quality images, etc. Despite these difficulties, many commercial and prototypical assistive technologies for the detection of objects of interest, obstacles, text and for face recognition are available.

Although the scientific literature on the topics is vast, both face recognition and text detection in unconstrained setting are still open problems. Besides the natural variability of the appearance of the subjects, due for instance to facial expressions and accessories, face recognition on mobile devices becomes more challenging because of the motion of the subjects, changes in the lighting conditions and the low quality of the images. The problem of text detection is extremely challenging because of the variations of text due to different style, orientation, alignment and scale together with low contrast images and complex textured backgrounds.

[1] `www.chiossone.it`

The methods we describe in this thesis have been developed keeping in mind prospective real-time applications to be used on mobile devices by visually impaired users. For this reason, the proposed methods are designed to obtain good performances using a minimum amount of computational resources.

On the problem of face recognition, the contribution of this work is twofold. First, we introduce an efficient real-time video-to-still face recognition system that exploits the advantages of face retrieval and face classification by combining them in a feedback loop.
The second contribution to face recognition is the introduction of a novel method for simultaneous selection of structured features. Identifying the most meaningful subset of features common to all the identities present in the system, we obtain an improvement in recognition performance and a reduction of the space complexity exploiting a sparse representation of the data. Unlike the vast majority of the available approaches for feature selection in face recognition ([ZHL+05, LCA+06, HLW05, SBBW05]), we show that using our method it is possible to perform the selection directly in the multiclass domain without resorting to the intraclass-extraclass binarization procedure ([MJP00]) and achieving better recognition results.

Concerning the text detection, we propose a blob-based text spotting algorithm. In our work, the text candidate blobs are represented by extremal regions [MCUP04], overcoming the difficult task of selecting a threshold for the image binarization. Exploiting the inclusion relationship existing between extremal regions, we propose an efficient and effective Bayesian framework for filtering out "non text" blobs. Unlike [NM12], our method, characterized by a solid theoretical justification, performs the classification of the blobs by maximizing the likelihood of the labeling of the tree, improving detection results.

The core of the thesis contains two parts; the first is devoted to the state of the art and our contributions to face recognition, the second introduces the problem of text detection and and describes our contributions to text detection.

The thesis is structured as follows:

- **Chapter 2**: Gives an overview of the available image-based applications on mobile devices for visually impaired users. We present an overview of the state of the art of the assistive technologies for object, face and character recognition and for independent mobility and wayfinding. We also focus on the challenges that such technologies face and their limitations.

- **Chapter 3**: In this chapter we introduce the problem of face recognition and the challenges characteristic of the domain; we also present an overview of the state of the art.

- **Chapter 4**: Describes the real-time face recognition system we propose. We describe the modules that compose the system and show the results obtained on two datasets.

- **Chapter 5**: Describes our novel method for simultaneous feature selection in multiclass

problems with structured features. We show the validity of the proposed method in the specific case of face recognition, comparing the obtained results with other state of the art feature selection methods on three datasets.

- **Chapter 6**: This chapter introduces the problem of text detection and the challenges a general purpose system has to face; we also describe the proposed method for text detection, comparing our Bayesian approach with different algorithm for the selection of text candidate from the tree of the extremal region. We present the results obtained on the ICDAR 2011 dataset.

Chapter 2

Image-based Assistive Technologies

In this chapter we will give an overview of Computer Vision-based assistive technologies for the visually impaired. We will describe the different contexts in which such technology represents an improvement in the quality of the life for the blind users and which are the limits and the problem to be faced by these assistive technologies. In Section 2.1 we will introduce the problem of Computer Vision and its open challenges, while in Sections 2.2,2.4 and 2.5 we will discuss methods related to the problems of detection of objects of interest, optical character recognition and the detection and recognition of non-text information respectively. In each section we will describe some prototypical and commercial applications. In Section 2.6 we will discuss the limits and the open challenges for vision-based assistive technologies.

2.1 Introduction

Computer Vision (CV) can be defined as "a form of artificial intelligence that strives to make computers able to *see* like normally sighted persons" [MKB10]. Dealing with problems ranging from the image acquisition to the image understanding passing through image processing and analysis, the final objective is to extract interesting visual information from a scene (either a static image or a video); which information is considered interesting is defined by the particular task the user wants to perform. Computer Vision achieves satisfying results when the problems to solve are well constrained, that is when the data are subject to limited variations and the quality of the data is acceptable. When dealing with more complex problems, i.e. when the problems are not characterized by a set of simplifying constrains, CV has to face enormous challenges showing that CV itself is not an entirely solved problem. Despite those limitations, there are tasks such as face detection and recognition, optical character recognition, 3D reconstruction and event recognition that have reached satisfactory solutions used in a wide range of practical applications and commercial products. In the last years, the increasing availability of compact

devices equipped with cameras made possible the spread of computer vision applications on mobile devices, establishing a growing interest in assistive technologies for visually impaired users.

Given the different conditions (different lighting conditions, points of view, perspective distortions, etc.) in which the data can be acquired using a mobile camera, CV has to cope with a set of loosely constrained and hence extremely challenging problems.

In the remainder of this chapter we will give an overview of the mobile assistive technologies focusing on specific tasks and we will discuss the current limits of the Computer Vision in this specific context.

2.2 Technologies for Object Recognition

The recognition of objects from images, that is the problem of finding and recognizing one or several objects of interest, represents one of the biggest challenges in CV. Beside the hard task of engineering features that uniquely characterize each class of objects, there still remain another difficulty: there usually is a high degree of variability between objects of the same class, while objects belonging to different classes may differ by subtle details. Objects of the same kind, in fact, can appear different according to their pose and/or lighting conditions, while on the contrary, these variations can make two objects of different kind look similar. 2.1 A general

Figure 2.1: Example of variation of the appearance of the same object under different point of views.

purpose object detector would then have to model objects without over-fitting on the available training data, while being at the same time capable of discerning different classes of objects. This goal is challenging and hard to achieve and usually the available commercial applications cope with the problem in simplified settings, i.e. specifying a subset of specific objects of interest.

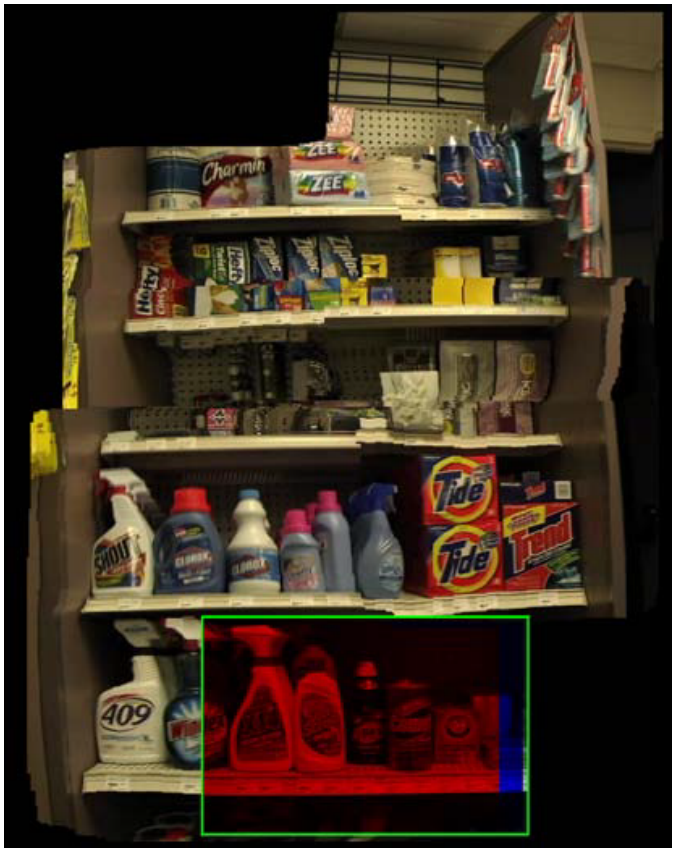


Figure 2.2: Example of detected grocery by ShelfScanner [WCB10]

An interesting method has been proposed in [WCB10], in which the authors introduce *ShelfScanner*, a prototypical object detection system that aims at allowing visually impaired users to find the needed groceries in a store in an autonomous way. ShelfScanner is capable of performing real-time detection of items on a shopping list, exploiting a mosaicing technique that combines different images acquired in a video sequence to create a single coherent image of the scene. The system works well under the assumption that the user sweeps the camera's field of view across grocery shelves.

Another domain specific object recognition system for impaired users is the recognition of currency. The work in [Liu08] describes a prototype camera-phone based system to recognize cur-

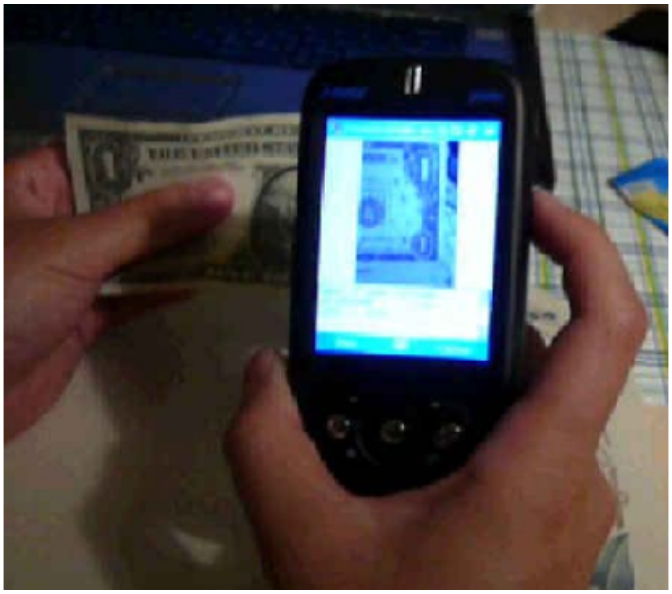


Figure 2.3: Example of currency recognition [Liu08]

rency in real time from video. Taking into account the difficulty for an impaired individual to capture high quality images, the system exploits the video stream processing each captured frame until a good view of the object is available. The same principle is exploited by LookTel Money Reader [1], a commercial iPhone application that recognizes currency in real time.

More generic commercial visual search engines, mainly tailored for normally sighted users, are oMoby [2] , A9's SnapTell [3] and Microsoft's Bing Mobile [4].
Currently the best examples of visual search engine is represented by Google's Goggles [5], capable of recognizing text, landmarks, book covers, artworks and logos. All the application listed before would represent interesting tools to give access to information otherwise inaccessible to blind individuals, but unfortunately their human machine interface limits the use to normally sighted individuals.

[1] www.looktel.com
[2] https://www.iqengines.com/omoby/
[3] http://www.a9.com/whatwedo/visual-search/
[4] http://onmobile.msn.com/en/Products/MobileWeb/BingMobileWeb
[5] http://www.google.com/mobile/goggles/

2.3 Technologies for Face Recognition

One of the hardest challenges for a blind person is the interaction with people in unfamiliar environments. It is, in fact, impossible for a visually impaired person to have clues about the presence of a known person in a room or knowing who are the people in the surrounding environment unless each person give a vocal feedback to the purpose of being identified.

A valuable support for blind users would then be a wearable device capable of locating and

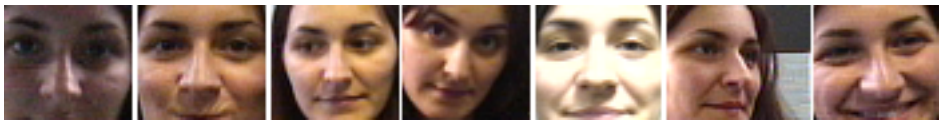


Figure 2.4: Example of face images belonging to the same person under different lighting conditions, poses and facial expressions.

recognizing people in the scene.

Face recognition is an active research field that, similarly to any other general purpose object recognition system, faces many challenges because of the enormous variations in the data [NM98](see Fig. 2.4). In fact, faces are subject to dramatic changes in the aspect due to variation of the pose, expressions, placement of the hair, glasses, lighting conditions, angles of view, partial occlusions, etc.

Despite the hardness of the task, the literature is vast and a variety of methods are available that achieve good performance when operating in constrained conditions, that is with a controlled degree of variability in the data. The most successful field of application for the face recognition is represented by surveillance and in the identity verification context.

Although there is a wide variety of methods working reliably in controlled settings, the reliability of such methods is not sufficient for recognizing people in real-world contexts. Indeed, given the hardness of the task, there are only few works focusing on the development of face recognition tools for visually impaired users.

In [KHR10] the authors present a prototype smart-phone face recognition application to be used in office settings. The proposed system delegates the execution of the algorithmic part for recognition to a server, relying on the availability of a photo database of all employees or students associated with a building. The system takes advantage of the knowledge of the setting by reducing the set of known identities among which recognize the detected faces, announcing the names of the identified people using a text-to-speech synthesizer.

The work of [KLBP05] describes the iCare Interaction Assistant, consisting of a wearable device for assisting visually impaired individuals during social interactions. The proposed prototype system is based on an analog CCD camera mounted on special glasses, using a laptop to execute

Figure 2.5: Camera glasses of the iCare Interaction Assistant [KLBP05]

the face recognition algorithm.
Also related to the problem of face recognition is the work presented in [GKP09]. Here the authors deal with the problem of person localization introducing the *Social Interaction Assistant*, a prototype wearable device that focuses on the localization of a person approaching the users while facing them.

2.4 Technologies for Optical Character Recognition

An Optical Character Recognition (OCR) system translates the content of a scanned document into letters and words, successively available to be stored in files or to be read aloud by text-to-speech softwares. The combination of OCR system with vocal synthesis clearly has the potential for being a great assistive technology by which visually impaired people can access printed information, that is giving the users the possibility to access one of the most important form of communication for nowadays.

After its first electronic prototypical version realized in 1946, many other prototypes have been developed but none of them found a real application in the following 30 years, mainly because of their dimensions and costs. The first commercial product available to the visually impaired community was the Kurzweil Reading Machine in 1979. The system was able to recognize multiple fonts and accessed the text by means of a flat-bed scanner. The produced output was then read aloud using a text-to-speech synthesizer. To ensure an adequate usability of the system by blind users, its testing phase was supported by the National Federation of the Blind (NBF). Ten years later, a non-profit organization called Arkenstone developed OpenBook [6], a reading tools for people with disabilities that was delivered to over 35,000 users.

The increasing availability of cheap mobile devices, equipped with cameras and good computing power, opened the road to the development of portable OCR systems, extending the use of the OCR technology to the recognition of text on commercial signs, street signs and in general all the form of text information, besides from documents. The majority of the available OCR application

[6] `http://www.synapseadaptive.com/arkenstone/Open_Book_Ruby.htm`

Figure 2.6: Result of text recognition [WBB11]

for smart-phones are thought to be used by normally-sighted users. Remarkable applications are ABBYY TextGrabber + Translator [7] and Word Lens [8], which combines OCR and augmented reality technologies to translate and replace with a translated version the text present in a scene.

Unlike normally-sighted oriented applications and desktop OCR, in which the software assumes the cooperation of the users and/or some constraints on the appearance of the text, mobile OCR applications for visually impaired face many difficult challenges. It is straightforward, in fact, to imagine that many of the assumptions made for the flatbed not necessarily hold while using a hand-held camera to access the text. A whole class of challenges is related to the position of the camera with respect to the text. For a visually impaired user, in fact, is impossible to know where to point the camera and if the text is properly framed. Moreover, even when the camera is pointed towards the text, the user could be standing too close – partial cropping and focusing problems – or too far from it – text too small and confusing clutter. Another class of problems is due to the nature of the acquisition system itself. In fact, the low quality of the images acquired by mobile devices due to motion blur or low light conditions dramatically affect the OCR performance.

Many attempts of solving these challenges characterize the recent ongoing computer vision research, under the topic of text detection or text localization in natural images, aiming at finding text in images by means of algorithms robust to clutter ([CCC+11], [WBB11]) and invariant to scale and resolution [SSC11], in order to provide an optimal and less noisy input to OCR systems.

[7] http://www.abbyy.com/textgrabber/
[8] http://questvisual.com/

A commercial mobile OCR systems designed for visually impaired users is the knfb Reader Mobile [9], especially designed for NOKIA cell phones and characterized by ease of use. Another visually impaired users oriented tool is Intel Reader [10], instead, a commercial product consisting in a dedicated portable tablet device.

2.5 Technologies for Independent Mobility and Wayfinding

Our world is populated by many form of information that are inaccessible to visually impaired users, limiting their mobility in unknown and potentially dangerous environments.
Large part of the information is in fact represented by signs, such as traffic lights, crossing lines, government and commercial signs. Being able to access these elements, together with the detection of points of interest and obstacles, would improve the personal safety and the ability to interact with the outdoor environments of visually impaired people.

Signs reading is a challenging problem to solve because of the extreme variability of their appearance and the variety of imaging conditions, and has received little attention in literature. It is worth citing [SWH+05], which introduced *VIDI* (Visual Integration and Dissemination of Information), a prototype system for detecting and recognizing signs in the environment and voice synthesize their contents, and [MHLM05], in which the authors describe a set of algorithms for sign detection and recognition for a wearable system to be used by the blind, capable of recognizing a broad variety of signs.

2.5.1 Pedestrian Guidance

More important for the safety of the blind individuals is the recognition of signs and signals at traffic intersections, such as walk lights and painted crosswalks, to help them to get aligned to the crosswalk and keep the alignment while crossing. In this context, CV can be used to estimate the layout of the intersection and give feedback to the users about the traffic light signals. Several prototypes of this kind of software have been developed on smart-phones, among which *Crosswatch* [ICS09]. Crosswatch is a mobile phone-based system to help visually impaired pedestrians avoiding entering the crosswalk in the wrong direction and straying outside of it by means of audio guidance, in order to correct both translation errors and direction errors. The system has been tested by blind users demonstrating the feasibility of the system. In [ICS10] the authors deal with the problem of designing a computer vision algorithm to rapidly and reliably detect the walk light signal at a traffic light whenever it is present, in order to assist the users before starting entering a crosswalk.

[9] http://www.knfbreader.com/products-mobile.php
[10] http://www.careinnovations.com/products/intel-reader-text-to-speech-technology

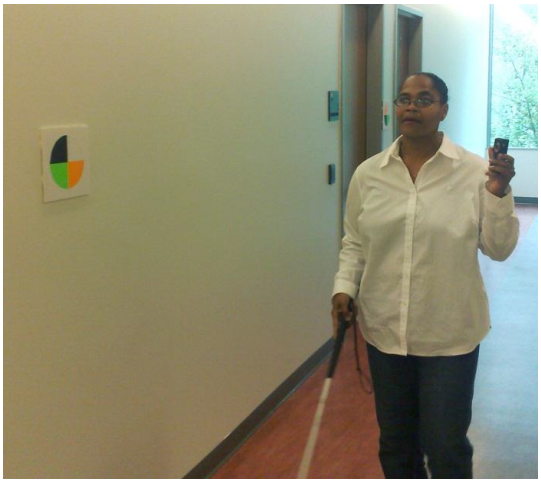


Figure 2.7: An example of multi-colored marker used in wayfinding applications [MKB10]

As for the signs recognition and the mobile OCR, the available softwares for the problem of pedestrian guidance cannot be widely used because of the wide range of operating conditions and the variety of the appearance of signs and traffic signals.

2.5.2 Independent Mobility

Independent mobility is among the hardest challenges faced by blind people. In fact, being able to move independently in unfamiliar or difficult environments requires orientating and wayfinding skills. For instance, in an airport visually impaired individuals can easily orient themselves using braille maps, but they need to physically reach them, hence locate them. An interesting application to support blind people in unfamiliar settings is then to provide a tool capable of detecting special signs that could work as landmarks, extracting spatial informations and providing guidance, orientation and location information by means of sounds or vibrations.

An example of prototype non-textual marker was developed by [CM07]. This special marker, a multi-colored pie-shaped marker (see Fig. 2.7), used for the purpose of guided mobility [MKB10], carries information through the embedded colors. In [BM11] the authors introduce a robust and fast detection algorithm for the multicolored-marker, being able of successfully detecting such markers in cluttered environments.

In [TBR+05], the authors propose a prototype location determination system based on two components. The first, the Digital Sign System (DSS), is responsible of the detection (within a range of 3 meters) of special designed tags (see Fig. 2.8), each containing a numeric code that specifies a location data in a spatial database. The second component, the Indoor Guidance System

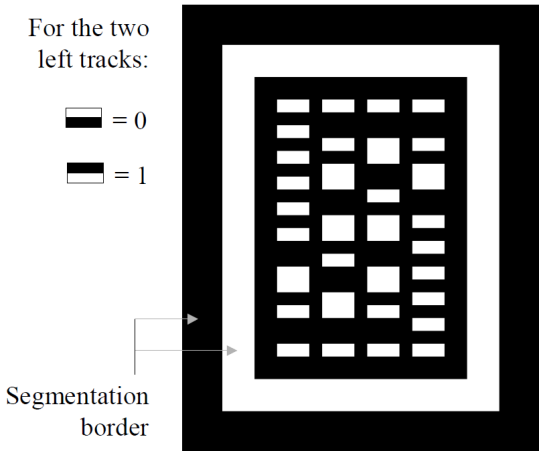


Figure 2.8: An example of DSS tag coding for 1234 [TBR+05]

(IGS), using the DSS codes provides information about the layout of the building and deliver speech-based guidance instructions to the user for wayfinding.

2.5.3 Vision-Based Electronic Travel Aids

Sensory systems for mobility support are technologies aiming at detecting any feature in the environment that could be critical for a safe ambulation, such as obstacles, steps and points of access. The idea behind these applications is to replace or enhance the long cane, the widely used tool that ensure safe ambulation and that represents the baseline against which any Electronic Travel Aid (ETA) system should compete with. In general, these applications are based on depth perception, that can be achieved by means of stereo vision or by active triangulation. The two systems produce the same results, but they have different drawbacks. For instance, stereo vision requires that the surfaces reflect light – hence it cannot operate in dark environments – while active triangulation hardly works under full sunlight; they both fail in case of transparent surfaces.

An example of an ETA system based on active triangulation is the *Teletact* [HFLJ06]. It consists on a laser pointer clipped on a long cane (see Fig. 2.9) that measures the distance to a surface and gives feedback to the user using tactile or acoustic interface.
Another example of a prototype active triangulation system is [YM04], which not only measures the distance, but takes into account the time profile of the distance as well, in order to detect depth discontinuities, usually occurring in presence of steps.

A stereo-vision based device is described in [PMW10]. The prototype system consists of a head-mounted stereo-camera that enables the user to scan the scene. Exploiting the dense 3D data

Figure 2.9: Teletact [HFLJ06]

acquired, using the ”simultaneous localization and mapping” (SLAM) technique, the system builds a vicinity map of the surrounding environment and, based on this map, delivers feedback to the user using a tactile interface.

2.6 Limits of Assistive Technologies

Throughout the chapter it has been stated that image-based assistive technologies, except for some specific constrained applications, are not ready yet to be used extensively and reliably by blind individuals because. The challenges in assistive technologies for blind users are, in fact, many. An acceptable solution should be able to deal with the high variability of the environments and of the objects of interest, should cope with the limited control the users have on the quality of the acquired images while giving feedbacks in a reasonable amount of time, possibly in real-time.
Conciliating all these problems at the same time is usually not feasible. For instance, having a high resolution image would make a sign detection system less sensitive to the distance from the sign or to the size of the sign itself, hence increasing the detection rate, but at the same time it would make the system less performing because of the longer processing time that would reduce the frame rate.
If we still consider the case of a tool for sign detection, the quality of this system may usually be measured in terms of detection rate and false positive rate. Ideally, it would be desirable to have a high detection rate while achieving the lowest false positive rate possible; unluckily those two parameters go in opposite directions, thus usually a compromise between performances and reliability is necessary. It hence follows that achieving a reasonable level of *quality* in image-based assistive technologies is not trivial and even in cases in which Computer Vision achieved satisfying results, there still remains the problem of a missing interface that could help the visually

impaired user to provide to the application an adequate input.

From the point of view of the user, two questions are still unanswered. The first question is how the system should look like, that is how the camera should be carried by the users; the second question is which user interface is the best and which kind of feedbacks the system should give to the users. Concerning the first question, besides cosmetic considerations, whether portable or wearable, an assistive system should be small enough to be carried around easily. For this reason, in the last decades smart phones have become the most common target platform for assistive technologies. Concerning the usability of the systems, it is important that the users could interact easily and efficiently to the applications; this means that no visual feedback should be necessary. The availability of voice-over tools comes in help allowing non-visual interactions with smart phones interfaces, including touch-screen devices. Image-based assistive technologies represent a sort of sensory substitution, in which data coming from the scene are processed and the output is returned to the user using one or more of his or her senses. A poorly designed interface would make an assistive system useless. Common feedbacks are represented by speech, sound, and tactile signal or a combination of them. Unfortunately a general rule about the best combination of feedbacks does not exist, and little research has been conducted on the topic ([MLK+06, RB00, WL+05]).

Chapter 3

Methods for Face Recognition

In this chapter we introduce the problem of face recognition, the challenges it faces and the available solutions. In Sec. 3.1 we give an overview of the difficulties a general purpose face recognition system has to cope with. In Sec. 3.2 we describe the key element of a prototypical face recognition pipeline. Further details about about face detection approaches are given in Sec. 3.3, while in Sec. 3.4 we give an overview of the state of the art concerning the face recognition.

3.1 Introduction

The problem of the face recognition can be expressed in the following form: given an image, the goal is to identify each face present in the scene.

Face recognition has received significant attention in the last decades, mainly for its wide applicability both for security and commercial purposes[1] [Kim05, Moo04, LWF05]. The recognition of faces is a challenging task because of the high variability in the data due to intrinsic and extrinsic causes [GMP00] (see 3.1). Factors affecting intrinsic variation include:

- Facial Expressions;
- Appearance (glasses, hair placement, makeup, beard, etc);
- Aging.

Factors affecting extrinsic variation are:

[1] An updated list of face recognition systems vendors can be found at `http://face-rec.org/vendors/`.

- Pose of the subject, angle of view;
- Lighting;
- Image quality (focus, blurring, noise);

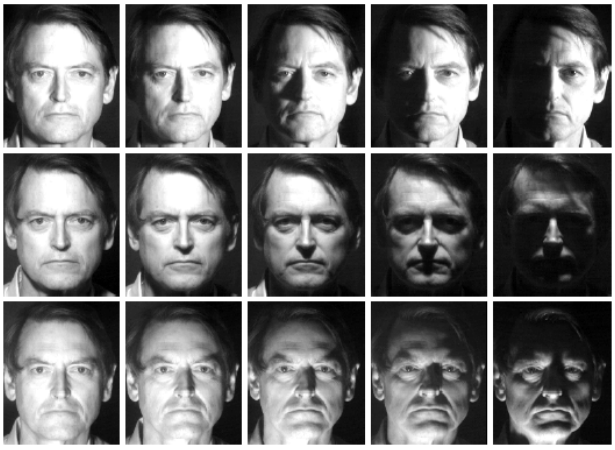


Figure 3.1: Example of variations of a face appearance under different lighting conditions [BHK97].

As a consequence of such variability, the development of a general purpose face recognition system with satisfying reliability represents a big challenge. Additionally, since in many applications the real-time constraint is necessary, the challenge becomes harder because of the reduced computational time requirement, a limitation that is especially true on mobile platforms and that imposes a compromise between recognition rate and system performances.

3.2 A Face Recognition Pipeline

An exemplifying face recognition pipeline (see Fig. 3.2) can be represented by the following steps:

1. Image/Video Acquisition;
2. Face Detection;
3. Face Representation;
4. Recognition.

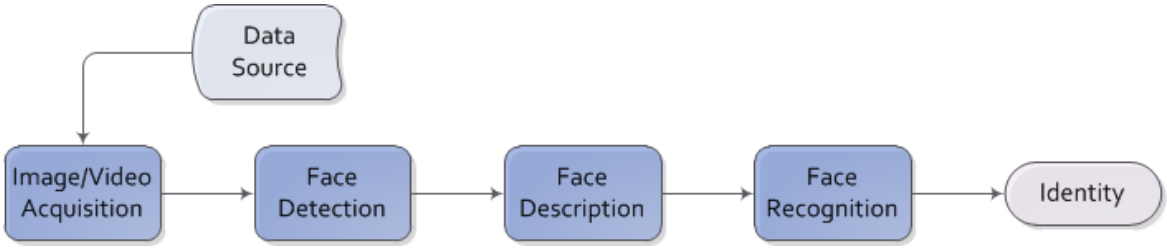


Figure 3.2: A generic face recognition framework.

The acquisition step deals with the problem of grabbing data from a source (video stream). The face detection, instead, is the task of detecting and possibly tracking the position of a face in a scene.

The extraction of patterns for face representation plays an important role in the face recognition pipeline; the ability of the extracted features of separating the different identities and its resistance to variations are crucial to achieve satisfying recognition rates.
The last step consists of the recognition process, usually related to the adopted description of the face patterns, and can be divided in two main approaches: face retrieval and face classification.
In the face retrieval paradigm, given a set of pairs image-identity stored in the system (usually this set is referred to as "gallery"), the recognition process consists in finding the first neighbour (or in less strict contexts the first k-neighbours) of a query image w.r.t. the similarity (conversely the distance) between the extracted features.
The face classification approach, instead, relies on the use of machine learning methods to associate an identity to a query image. The problem of face recognition is a multiclass classification problem in which each known identity represents a class. The literature concerning the solution of multi-class classification problems is vast; among the available algorithms it is worth citing the All-vs-All algorithm, in which a binary classifier for each pair of training labels is built, and the One-vs-All method, in which for each identity a binary classifier is trained using a set of positive examples against a set of negative examples, sampled from all the remaining identities in the system. Besides the difference of techniques the two paradigms employ, there is a difference in the results obtained with the two methods. In the face retrieval paradigm, the output is composed by returning a list of candidate identities, ordered by their similarity to the query image. On the other hand, classification-based methods usually aim at returning a unique answer, that is the identity associated to the classifier whose confidence value is higher then all the others.

A different problem that falls into the family of face recognition methods is identity verification. In identity verification problems the input to the system is an image of a face and a declared identity of the person. The output is the acceptance or the rejection of the identity declared.

3.3 Face Detection

Detecting faces in an image is an easy visual task for humans, but from the Computer Vision standpoint it represents a challenging task. The solution to the problem involves segmentation of face regions or extraction of facial features from an uncontrolled background, regardless of illumination, orientation and camera distance.

Face detection techniques can be divided into two main categories, that differ in the way the a priori face knowledge is used.

The first category consists of techniques that make explicit use of face knowledge and follow the detection methodology in which low level features are derived according to knowledge-based analysis, such as skin color, face geometry and derived measurements.
The techniques that fall in the second group address face detection as a general recognition problem, exploiting pattern recognition methods. Image-based representations of faces [DCC96], for instance intensity images, are directly classified into a face group using training algorithms without feature derivation and analysis. Unlike the feature-based approach, these techniques implicitly incorporate face knowledge into the system through mapping and training schemes.
It has been shown in [HL01] that feature-based methods are applicable for real-time systems where color and motion are available, while image-based approaches are more suitable for processing gray-scale static images. An exhaustive survey on face detection method can be found in [ZZ10a]. [VJ04] presented what can be considered a reference method in the face detection field, because of its effectiveness and its low computational cost. The method consists of a cascade of simple classifiers is built using AdaBoost [FHT01], trained on rectangle features extracted from face and non faces images. A detailed overview of the state of the art about face detection can be found in [ZZ10b].

3.4 Face Recognition

In this section we will introduce the principal approaches found in the literature that cope with the problem of the face recognition. We will describe different ways of extracting facial patterns and different approaches of associating identities to query images, discussing advantages and drawbacks of each method.
An extensive analysis of the face recognition literature can be found in [ZCPR03, JA09, ZG09]. A specific survey about video based recognition is [MD09].

3.4.1 Holistic Approaches

In holistic template-matching systems each template is a prototype face, a gray-scale image, or an abstract reduced-dimensional feature vector that has been obtained through processing the face image as a whole.

Holistic approaches attempt to identify faces using global representations, i.e. descriptions based on the entire image of the face . In the simplest version, the image is represented as a 2D array of intensity values and the recognition is performed by directly comparing the input face against all the other faces in the database. This approach works well under limited circumstances (i.e. fixed illumination, scale, pose, etc.) and due to the correlation-based nature it is sensitive to face orientation, size, variable lighting conditions, background clutter, and noise. The main drawback of this kind of methods is the high dimensionality of the data they treat. To counter this curse of dimensionality, several other schemes have been proposed that employ statistical dimensionality reduction methods to obtain and retain the most meaningful feature before performing recognition. Low-dimensional representations are highly desirable for large databases, fast adaptation, and good generalization.

One of the most popular holistic approaches is the Eigenfaces [TP91], used both for face detection and recognition purposes. The idea behind eigenfaces is to extract the relevant information in a face image, not necessarily related to the notion of face features (such as eyes, nose, lips, etc.), encode it in an efficient way and compare an encoded face with a database of similarly encoded faces. More precisely, the method aims to find the principal components of the distribution of faces (eigenfaces), treating an image as a point in a very high dimensional space, and projecting the faces in the compact subspace spanned by the eigenfaces.
During the training phase, a small set of data is used to estimate the eigenface space with Principal Component Analysis (PCA) [Jol05]. The eigenface space is the best set of projection directions that maximize the total scatter across all the images, obtained maximizing the following functional:

$$J_{PCA}(W_{opt}) = \arg\max_{W} |W^T S_T W| \tag{3.1}$$

where S_T is the total scatter matrix of the training set samples, and W is the matrix whose columns are the orthonormal projection vectors. After the subspace is found all images of the dataset are represented as weight vectors, obtained by projecting the image into the subspace. In the test phase, a new (projected) image is associated to the identity of the image whose projection is closer to it.

Although the good performances achieved by the eigenfaces method in controlled settings, it suffers significantly from variations in the input images. [MP97] overcomes the limits of the eigenfaces, extending the method to facial features (eigenfeatures). The approach can be seen as an enrichment of the eigenfaces, low resolution description of the whole head, by adding high resolution details related to salient face features. The recognition is then performed exploiting a probabilistic similarity to model the intra-personal variations (variations within the same indi-

(a)

(b)

Figure 3.3: (a) An example of Eigenfaces shows the tendency of the principal components to capture lighting direction variations; (b) Example of Fisherfaces on the same training set shows the ability of FLD to discard variations not related to the recognition.

vidual, due to illumination, view-point, expression changes) against the extra-personal variations (difference between an individual and another).
While PCA-based methods detect the features that capture the main directions along which face images differ the most, they don't reduce the intra-class scatter of the faces. This is due to the fact that these methods are unsupervised, so they don't exploit any class membership information. In fact, choosing a criterion that maximizes the total scatter, PCA-based approaches tend to capture undesired intra-class variations, such as those caused by differences in the pose, lighting, facial expressions and other factors [BHK97], with the result that the reprojected points belonging to different classes may overlap to some extent.

In [EC97] the authors use the Linear Discriminant Analysis (LDA) applied to the spatial and the wavelet domain to evaluate the significance of different features of the face to the purpose of the face recognition. The most significant features are detected by eigenvector analysis of the scatter matrices maximizing between-class variations and minimizing within-class variations applying the Fisher's Linear Discriminant criterion (FLD):

$$J_{FLD}(W_{opt}) = \arg\max_{W} |\frac{W^T S_B W}{W^T S_W W}| \tag{3.2}$$

where S_B is the between-class scatter matrix and S_W is the within-class scatter matrix. The use of the class membership information hence allows to find the eigenfeatures that emphasize the variations among faces belonging to different classes with the result of having the points belonging to a class projected far from all the points belonging to all the other classes, while in

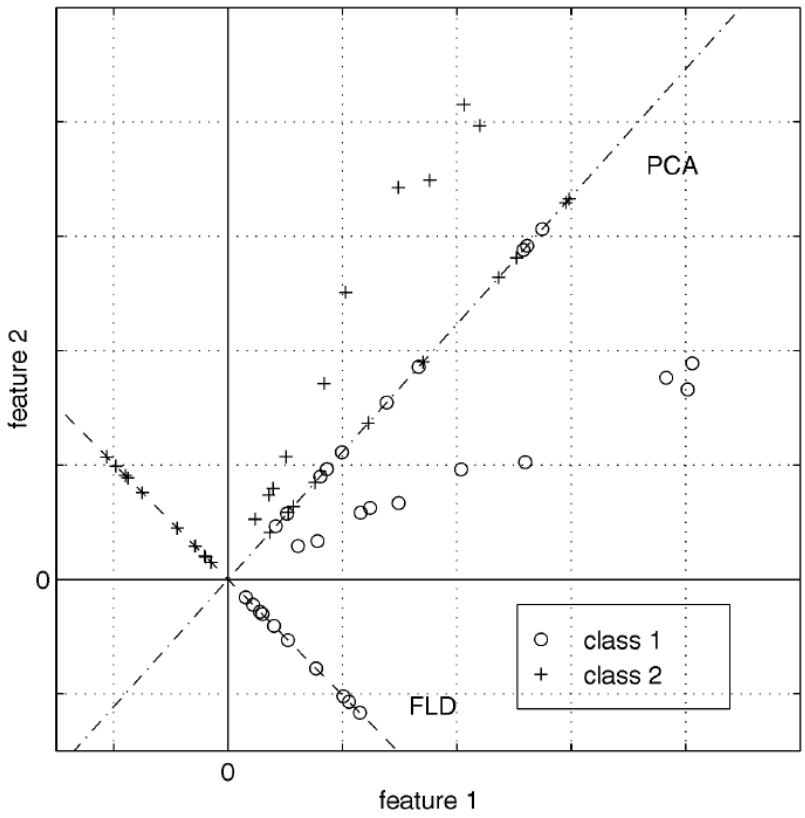


Figure 3.4: A comparison of the discriminative power of PCA and FLD methods [BHK97]. Notice the overlap of the two classes caused by PCA projection.

the same time minimizing the scatter between points belonging to the same class. FLD-based methods suffer from the " small sample size problem" [Fuk90]. The ratio in 3.2 is maximized when the projection vectors of the projection matrix W are the eigenvectors of $S_W^{-1}S_B$, hence when the size of the sample space is larger then the number of sample (a typical condition in face recognition problems) S_W is singular. Many methods have been proposed to solve this problem [TBGL86, ZCK98, HY91, CZY92], but are typically computationally expensive because of the size of the scatter matrix. [SW96] proposed the Fisherface method, a two stage PCA+LDA method, in which PCA is first used to reduce the dimensionality of the scatter matrix so as to make S_W nonsingular in order to successively apply LDA. The optimal projection vector matrix is then $W_{opt} = W_{PCA}W_{FLD}$, where W_{PCA} is obtained from Equation 3.1 and

$$W_{FLD} = \arg\max_W |\frac{|W^T W_{PCA}^T S_B W_{PCA} W|}{|W^T W_{PCA}^T S_W W_{PCA} W|}.$$

However the use of PCA to reduce the dimensionality of S_W has the drawback of potentially remove dimensions that could cast discriminative informations.

In [CLK+00] the authors propose the Null Space method, based on a modified FLD criterion. First all the images are projected onto the null space of S_W, in order to obtain a new within-class scatter that is a zero matrix, and then LDA is applied to the projected samples to obtain W_{opt} maximizing

$$J_{MFLD} = \arg\max_W \frac{|W^T S_B W|}{|W^T S_T W|}.$$

Because global features are highly sensitive to appearance and pose variations, holistic approaches are limited to recognize frontal views of faces under the constraints of having the same pose, fixed scale and no occlusions.

In [WYG+09] the recognition is posed as the problem of classifying using multiple linear regression models exploiting sparse signal representation. Under the assumption that samples belonging to a specific object class lie on a linear subspace [BHK97], the authors define models for each identity to be recognized by using the downsampled gallery images.

In [WW10] the intrinsicface model is introduced. The model is realized considering three kind of differences that face images convey: facial difference shared by all the faces, individuality difference and intrapersonal difference. A new dimensionality reduction method is introduced, called Intrinsic Discriminant Analysis (IDA) that tries to classify face images by the maximization of the individuality difference, while minimizing the intrapersonal difference.

3.4.2 Component-based Approaches

The idea behind component-based recognition is to localize key features of the face, such as eyes, mouth and nose, in order to compensate for pose changes and overcome the limitations of global methods. In fact, for small rotations, the changes in these detected features are small compared to the changes that affect the whole face image and the flexible geometrical relation between the components can be learned by means of a flexible face model to be used in the classification stage.

[Kan73] represents one of the first attempts to face recognition by facial features. The method was based on the extraction of 16 facial scale-independent parameters, using Euclidean distance measure for matching. In [BP93], inspired by Kanade's approach, 35 geometric features are computed (see Fig. 3.5). The authors assume that the feature vectors for a single identity are distributed according to a Gaussian distribution. Different identities are characterized by their own average value, while the distribution is common. The shape of the distribution is estimated by using all the examples in the training set:

$$\Sigma = \frac{1}{N} \sum_{i=1}^{N} \Sigma_i$$

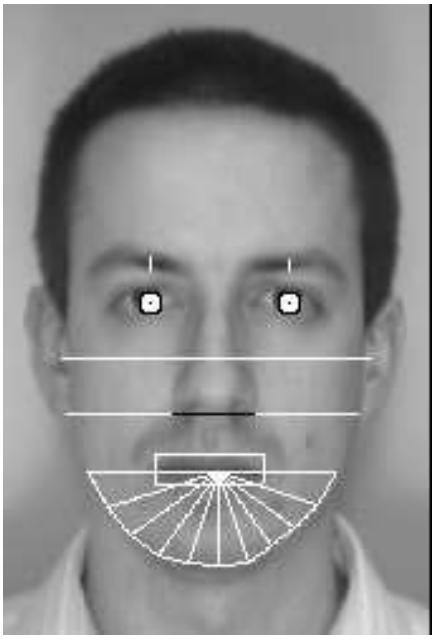


Figure 3.5: Geometrical features used in [BP93]

where Σ_i is the estimate of the covariance matrix computed from the data of the $i-th$ identity. Given a measurement $\mathbf{x}$ estimated from a new face, the probability that $\mathbf{x}$ belongs to the identity j is given by the following distance:

$$\Delta_j(\mathbf{x}) = (\mathbf{x} - \mathbf{m}_j)^T \Sigma^{-1} (\mathbf{x} - \mathbf{m}_j)$$

where $\mathbf{m}_j$ is the average vector representing the $j-th$ person. The new image is then associated to the nearest neighbor identity in the database, i.e. the identity that maximizes the probability of the measurement vector. [YHC92] introduces the use of deformable templates to detect salient face features. Features of interest are described by parametrized templates that interact with the image by adjusting their parameters in order to minimize an energy function. Other feature extraction techniques involve Hough transform methods [Nix85], Reisfeld's symmetry operator [Rei93] and Graf's filtering and morphological operations [GCPC95].
All the above cited techniques heavily rely on heuristics, such as reducing the search subspace by means of geometrical constraints, and the models also need some degree of tolerance since they can never fit the structures in the image perfectly. The choice of the tolerance value is difficult, since large values tend to make the detection of features less precise, negatively affecting the overall recognition process that relies on precise localization of said features. In general, algorithms for automatic extraction of features do not provide a high degree of accuracy and require a considerable amount of computational capacity.
The elastic bunch graph matching method [WFKvdM97] is another well known feature-based approach. A graph for an individual face is generated starting from a set of detected fiducial

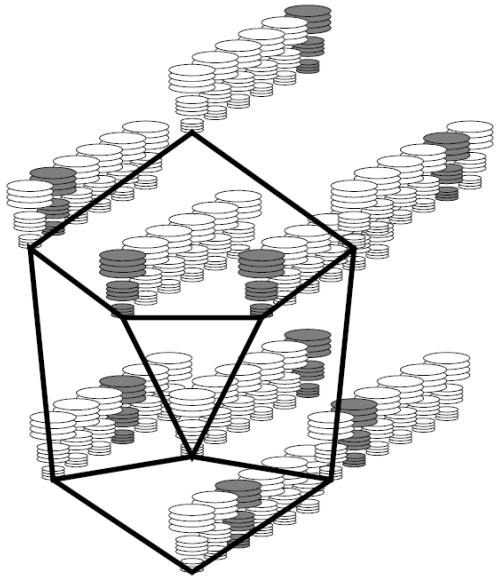


Figure 3.6: A sketch of a Face Bunch Graph (FBG) [WFKvdM97]. The FBG combines together different face graphs of an individual. Its nodes, called bunches, are labeled with sets of jets and its edges are labeled with averages of distance vectors. During comparisons to a face graph, the best fitting jet in each bunch, colored in grey, is selected independently.

points. Each fiducial point is then a node of a full connected graph, labeled with a *jet*, that is a set of Gabor filters' [D$^+$85] responses applied to the neighborhood of a fiducial point. Each node is connected by an arch labeled with the distance between the corresponding fiducial points. During the training phase, a representative set of face graphs needs to be manually generated and then combined into an individual face bunch graph, a particular stack-like structure in which jets corresponding to the same salient feature extracted from different face images of the same person are stacked together. Once the face bunch graph has been generated, the system can automatically generate new graphs from new face images by Elastic Bunch Graph Matching. The recognition of a new face image is performed comparing its newly generated graph to those of all the known identities and picking the closest one.

In general the FBG-based methods achieve good performances and is among the best performing methods in the FERET evaluation [PMRR00]. However the method has the serious drawback of requiring the manual placement of the graph for the first 70 faces before the automatic elastic graph matching becomes sufficiently reliable [Suk00].
In [HLC10] the authors propose a method to recognize expressive faces using only one neutral face image per subject in training, combining PCA classification and optical flow prior information into a probabilistic framework.

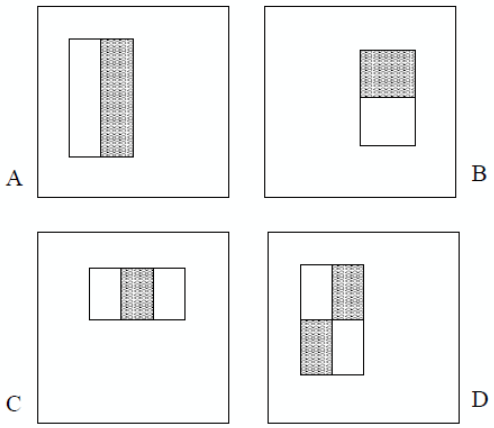
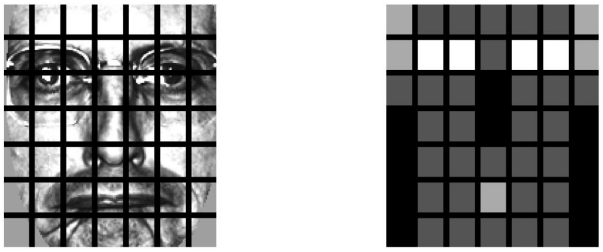


Figure 3.7: Example of four different types of rectangle features [VJ01].

3.4.3 Features-based methods

Local features deal with the problem of finding efficient and discriminative facial appearance descriptors that can be considered reliable even in case of large variations in illumination, pose, facial expression, aging, partial occlusions. In the last years, representations based on pooling of local appearance descriptors have drawn increasing attention because of their capacity in capturing small appearance details while remaining resistant to registration errors owing to local pooling.

These methods are also attractive since it has been shown that human visual perception is well-adapted to extract and pool local structural information (also referred to as micro-patterns) from images.
We have already introduced the use of Gabor jets in [WFKvdM97] for the detection of salient features in face images. In [LGP+05], instead, Gabor wavelets are used to extract desirable facial features characterized by spatial frequency, spatial locality and orientation selectivity to deal with the variations due to illumination and facial expression changes. The extracted high-dimensional Gabor features are then projected onto a lower-dimensionality subspace detected by means of PCA. The reprojected features are then use to train a Support Vector Machine (SVM) [Vap99] to perform the recognition tasks.
In [JV03] rectangle features are used to describe the training data. Rectangle features [VJ01] have the advantage of capturing well many simple patterns in image pixels, such as edge, ridge, symmetries and other characteristics of the intensity signal. A rectangle feature is defined as a filter $\phi_i(I)$ applied to an image I computed by summing the intensities of all pixels in the dark regions and subtracting the sum of the intensities of all pixel in the light regions (see Fig. 3.7). The recognition problem is solved assigning to a new image I_1 the identity associated to a stored



(a) (b)

Figure 3.8: (a) A face image divided into 7×7 regions. (b) The weights corresponding to the regions used for the weighted χ^2 distance measure [AHP06].

image I_2 that maximizes a face similarity function defined as a sum of features f_i:

$$F(I_1, I_2) = \sum_{i=i}^{N} f_i(I_1, I_2).$$

A feature consists of a filter that is applied on both input images:

$$f_i(I_1, I_2) = \begin{cases} \alpha & \text{if } |\phi_i(I_1) - \phi_i(I_2)| > t_i \\ \beta & \text{otherwise} \end{cases}$$

where $t_i \in \mathcal{R}$ is a feature threshold. AdaBoost [FHT01] is used to learn the best set of features f_i and the best values for thresholds t_i and weights α, β maximizing the recognition performances.

[AHP04, AHP06] introduced the use of Local Binary Pattern (LBP) operator [OPH96] for the purpose of face recognition. The LBP operator is a widely used texture descriptor due to its excellent performance and it has been shown to be highly discriminative and, most importantly, invariant to monotonic gray-level changes. It is also characterized by a high computational efficiency that makes it suitable for fast image analysis tasks.

The LBP operator computes a label for each pixel of an image by thresholding their 3×3 neighborhood with the center value and considering the result as a binary number. Formally, the LBP operator can be defined as follows:

$$LBP\,(x_c, y_c) = \sum_{n=0}^{7} 2^n s(i_n - i_c)$$

where n runs over the 8 neighbors of the central pixel c, i_c and i_n are the gray-level values at c and n locations, and $s(u)$ is 1 if $u \geq 0$ and 0 otherwise. An example of LBP encoding process is illustrated in figure 3.9

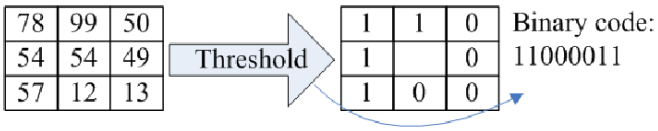


Figure 3.9: Example of basic LBP operator [TT07]

The texture description is then represented by the histogram of the labels.

Two extensions to the original LBP operator have been introduced by [OPM02]. The first is the generalization of the operator to use neighborhoods of different sizes. Said P the number of sampling points on a circle of radius R, we will refer to the generalized LBP operator as $LBP_{P,R}$.
Another extension to the original operator uses so called uniform patterns. An LBP is called uniform if it contains at most two bitwise transitions between 0 and 1 or vice versa when the binary string is considered circular. We refer to the generalized uniform LBP operator as follows: $LBP^{2u}_{P,R}$. Uniform patterns provide a vast majority of the 3×3 texture patterns and the most frequent detected patterns correspond to primitive micro-features, such as edges, corners, and spots. Given a LBP labeled image $f_l(x, y))$, its histogram can be defined as:

$$H_i = \sum_{x,y} \delta\{f_l(x, y) = i\}, i = 1, \cdots, n - 1,$$

where n is the number of LBP labels and

$$\delta A = \begin{cases} 1, \text{A is true} \\ 0, \text{A is false.} \end{cases}$$

The histogram contains information about the distribution of the local micro-patterns but discards the spatial information, that happens to be useful for an efficient face representation. In [AHP06] the image is divided in a set of regular regions $R_0, R_1, \cdots, R_m$, giving raise to the definition of spatially enhanced histogram:

$$H_{i,j} = \sum_{x,y} \delta\{f_l(x, y) = i\}\delta\{(x, y) \in R_j\}, i = 1, \cdots, n - 1, j = 0, \cdots, m - 1.$$

The final face description is then obtained concatenating the various $H_{i,j}$ to form a unique histogram (see Fig. 3.10).
The resulting histogram contains the description of the face on three different levels of locality: the patterns detected at pixel level, the distribution of the patterns extracted from each region and the concatenation of such histograms to build a global description.

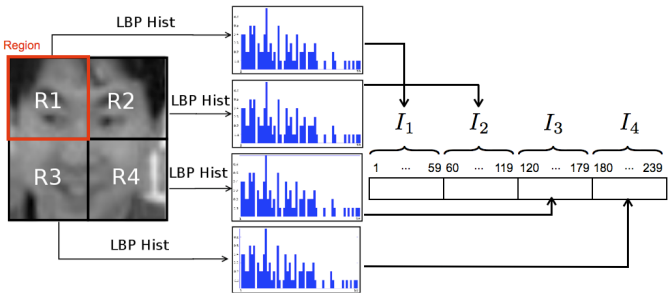


Figure 3.10: An example of spatially enhanced histogram computed from 4 regions of the image

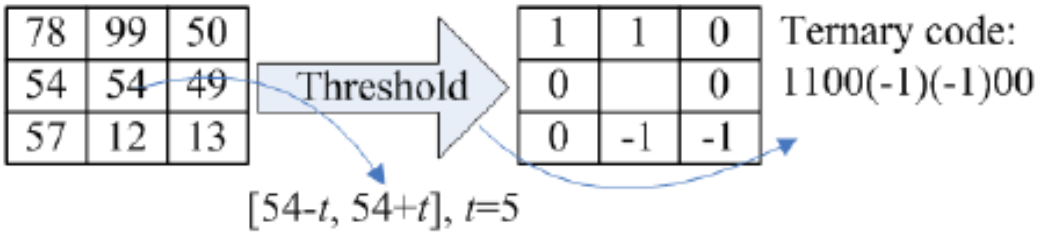


Figure 3.11: Example of LTP operator [TT07].

Under the assumption that certain regions in a face are more relevant for the face recognition in humans, two images are then compared by means of a weighted Chi square distance:

$$\chi^2(\mathbf{x}, \xi) = \sum_{i,j} w_j \frac{(x_{i,j} - \xi_{i,j})^2}{x_{i,j} + \xi_{i,j}}$$

where $\mathbf{x}$ and ξ are the normalized enhanced histograms to be compared, i and j refer to the ith bin in the histogram corresponding to the jth local region. The scalar w_j is the empirically estimated weight for region j.

In [TT07] Local Ternary Patterns (Fig. 3.11), an extension of LBP, are introduced to the purpose of reducing the sensitivity to noise of the original formulation of LBPs. In their work, the authors propose not to use enhanced histograms in order to better take into account the spatial resolution otherwise reduced using the approach of [AHP06]. The similarity between two images X and Y is measured using a similarity metric that for each LTP pixel code in X tests the existence of a similar code at a nearby position in image Y, using a weighting function that decrease with the distance:

$$D(X, Y) = \sum_{pixels(i,j) \text{ of Y}} w(d_X^{k_Y(i,j)}(i, j))$$

where d^k is the distance transform image respect to a LTP code b^k, $k_Y(i, j)$ is the code value of pixel (i, j) in image Y and $w()$ is any user-defined function giving penalty according to the

distance from the nearest matching code in X. In their work both the Gaussian similarity metric and truncated linear distances have been discovered performing well.

[BLGT06] represents one of the first method using SIFT features [Low04], in which the authors investigate the use of the scale invariant features for face recognition. The authors observe that the classification is more accurate when the extracted features are selected and grouped according to the face geometry, driven by the knowledge of the position of few facial landmarks.

To conclude this overview of local features-based approaches, we cite two biologically inspired approaches. [MW08], inspired by the feed-forward model of the primate visual object recognition pathway proposed by [RP$^+$99] (R&P Model), proposes an extension to the R&P model, introducing a new set of facial identification features, called S2 facial features (S2FF), a modification to the R&P C2 features used for object recognition.

A more recent biologically inspired method is [DHG$^+$10]. The system they propose emulates the biological strategies of the human visual system, representing the face using dual retina texture and color features. The similarity between faces is then measured using a hierarchical cue-fusion method, outperforming state of-the- art approaches, such as LDA and LBP.

3.4.4 Face Recognition from Video

In the last decades, automatic methods for detecting faces and validating identities from videos have received growing attention, because of interesting applications in the context of the video surveillance, but also because of the large diffusion of mobile devices with embedded cameras. Compared to still images, the size and the quality of the frames are usually inferior, but it is possible to exploit the temporal component to detect coherence in the patterns and obtain a better recognition rate.

In [CKZ02] the authors use a probabilistic framework adopting time series state space model to describe the kinematics contained in the observed videos. During the recognition phase, confidence values are accumulated for each identity in the system, and the recognition decision is made according to the degeneracy of the posterior probability of the identity variable.

[LC03] apply adaptive temporal Hidden Markov Model to perform face recognition from video. Each frame in a video is considered as one observation and is reduced to a low-dimensional feature vector by means of PCA. During the training phase, an HMM learns the statistics and the temporal dynamics of a sequence (see Fig. 3.12). In the recognition phase, the temporal characteristics of the face sequence are analyzed by the HMM corresponding to each subject and the HMM with the highest likelihood score provides the identity in the sequence.

In [Agg04] the Auto-Regressive and Moving Average (ARMA) model was adopted to model a moving face as a linear dynamical system and perform video to video face recognition. The recognition is performed measuring the similarity between ARMA models using subspace angles

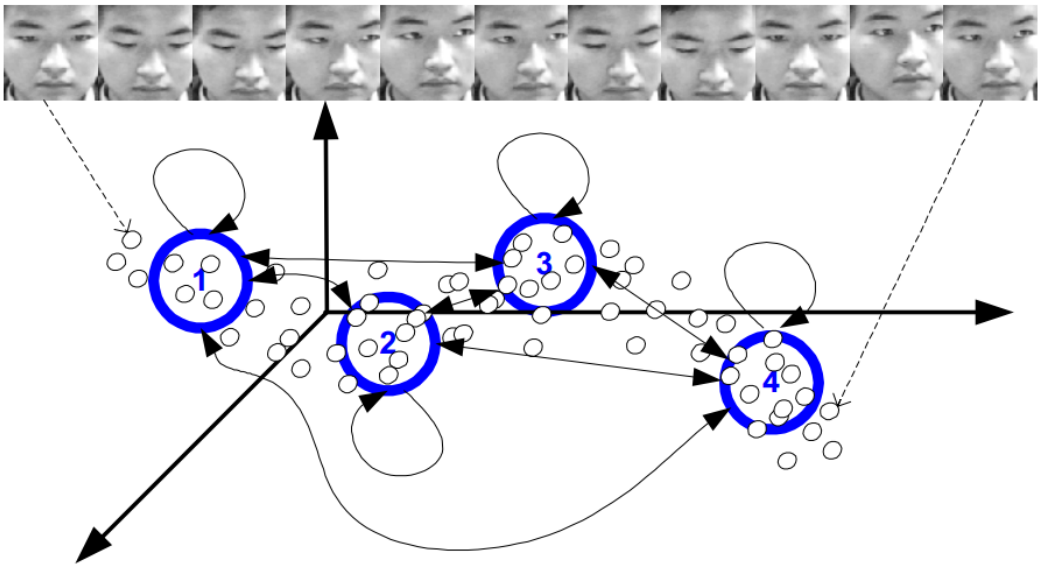


Figure 3.12: Example of temporal HMM for modeling a face sequence [LC03].

based distance metrics .

[CT10] introduces a geometric approach in which each test and training example is a set of images of an individual's face and the dissimilarity between images set is measured by geometric distances between convex models.

Recently in [WCM+11] the authors propose a patch-based face image quality assessment algorithm that measures the similarity of a face image to a probabilistic face model to select the highest quality subset of images from a video. The recognition is then performed using LBP descriptors, discarding the temporal component.

[ZP07] and [HMPL07], instead, treat the problem of face recognition from videos modeling the face as a dynamic texture, introducing two extensions to LBP descriptor. In [ZP07] Volume Local Binary Patterns combines motion and appearance, concatenating the LBP histograms computed from three orthogonal planes (Fig. 3.14). In [HMPL07] the authors propose an approach for face recognition from videos that uses local spatio-temporal features. A sequence is treated as a set of volumes from which to extract local histograms of LBP code occurrences, named Extended Volume LBP (EVLBP) (Fig. 3.14). Using AdaBoost, only the most significant features are selected and used for recognition.

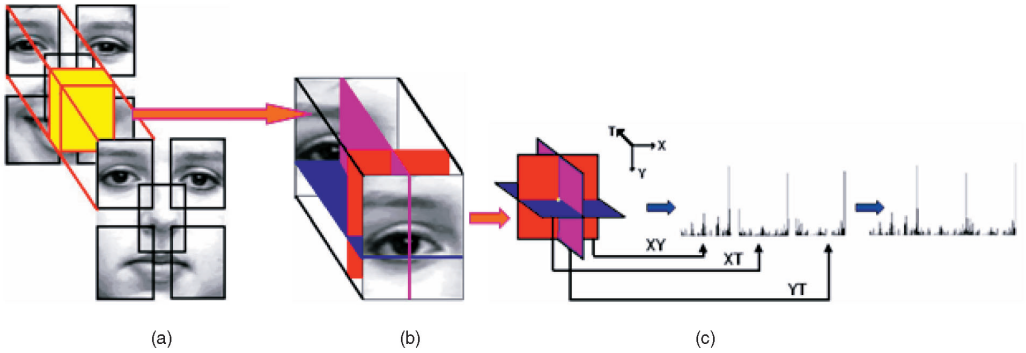


Figure 3.13: Example of VLBP[ZP07].

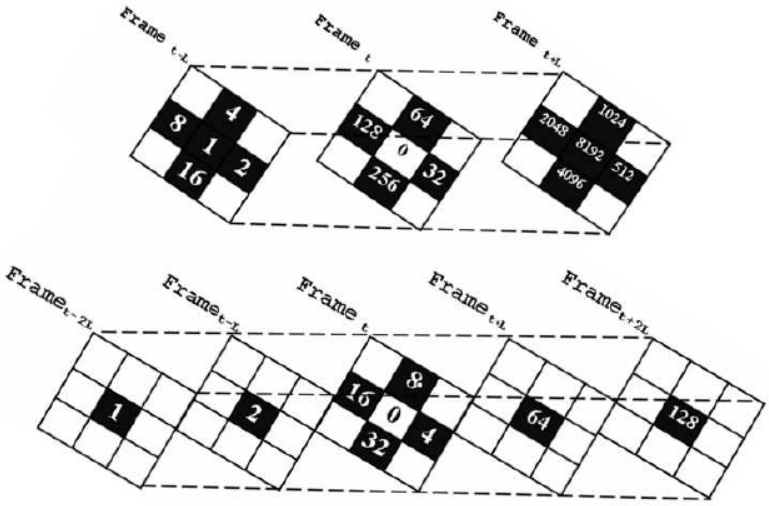


Figure 3.14: Example of EVLBP operator [HMPL07].

Chapter 4

Face Recognition Combining Retrieval and Classification

In this chapter we describe the novel video-to-still real time face recognition system we propose. The method consists in the combination of face matching and identity verification techniques to obtain a more robust recognition system, while also exploiting the temporal component in the video to refine the recognition hypothesis over the time. In Section 4.1 we introduce the proposed system, while in Sections 4.2 we briefly introduce the adopted method for face detection. In Section 4.3 we describe the approaches adopted to extract the face patterns and to model the identities, while in Sections 4.4 and 4.5 we describe the two recognition modules that compose the system. Finally in Section 4.6 we present the experiments we conducted and the obtained results.

4.1 The Proposed System

The problem of the face recognition, as previously stated in Section 3.2, can be faced by face retrieval or face classification approach. In general, the two approaches have different advantages and disadvantages, mainly related to the quality and quantity of the training data and to the way they exploit such data. The face retrieval, being based on the measurement of the similarity between a query face and the faces in a gallery, does not need a proper training phase and can produce reasonable results using few data. On the other hand, the retrieval performances are dependent on how varied the gallery is, that is how heterogeneous the gallery images for a given identity are.
Face classification, exploiting the generalization property of machine learning methods, is less data-dependent than retrieval. However, a large amount of data and a time consuming training phase is necessary to achieve the required generalization capability of the classifiers.

The idea behind the method we introduce in this chapter is to take advantage of the good properties of the face retrieval and classification approaches, while compensating their disadvantages. More specifically, we would like the system to be able to reasonably perform the recognition task in cases in which the data are not varied enough or the quantity of the available data is not sufficient to train a robust classifier for the identities.
Hence, the system we propose combines elements of the face retrieval paradigm with an architecture deriving from the identity verification context, achieving real time face recognition from video. More specifically, the method includes a video-to-still face matching approach that exploits the temporal information to refine the probe query achieving an enhancement of the recognition performance. The retrieval provides (possibly multiple) hypothesis on the identities of the people in detected in a video. Then, to reduce the entropy of the problem, a step of identity verification exploits the generalization property of linear Support Vector Machine (SVM) classifiers [Vap99] to validate or reject the recognition hypothesis.
A sketch of the system we realized is depicted in Fig. 4.1. We first pre-process each frame applying a face detector (see Sec 3.3), based on a feature selection algorithm that relies on a regularized approach with a sparsity term. Then, a fast coarse face registration is performed by detecting eyes and nose regions, with the same approach used for face detection, and normalizing their position with an appropriate cropping of the region of interest (i.e. the face region). Once faces are detected and coarsely registered, we represent them as Local Binary Patterns (LBP) [AHP06], which were proved to be an efficient descriptor for the face recognition task.
In the training or enrollment phase, starting from a set of training images of size n – which may vary from person to person – we compute an appearance model of the identity, composed of a gallery to be used for face retrieval, and an SVM model employed for face classification.
We need to control the computational cost in order to meet the real time requirement. Although there are many methods for face recognition from video (see Sec. 3.4.4), but they differ from the one we propose in terms of reactivity to the input data. The methods presented by [ZP07] and [HMPL07], for instance, propose two extensions of the LBP to consider the temporal coherence of the video. Despite the efficiency of the descriptors, these approaches introduce a temporal delay of a variable amount of frames in order to fill a temporal buffer from which extract the 3D face patterns. The same temporal delay is introduced by [Agg04], since several frames of the video probe need to be observed before the parameters of the ARMA model can be extracted.
In the method we propose, similarly to [BFO$^+$10], we exploit the temporal component of a video by analyzing each frame at a time, using the information from each incoming frame to refine the query and improve the overall recognition rate. The input video stream associated with a tracked face is analyzed considering a temporal buffer of depth T frames. Each detected face (probe) is associated with a rank of candidate identities with a retrieval approach. The top K candidate identities are then verified by classifying the probe using K One-vs-All SVM classifiers. The obtained confidence values are used to update a score over the temporal buffer and an estimated identity is returned after having considered T frames, to avoid an undefined wait time before obtaining an output from the system.

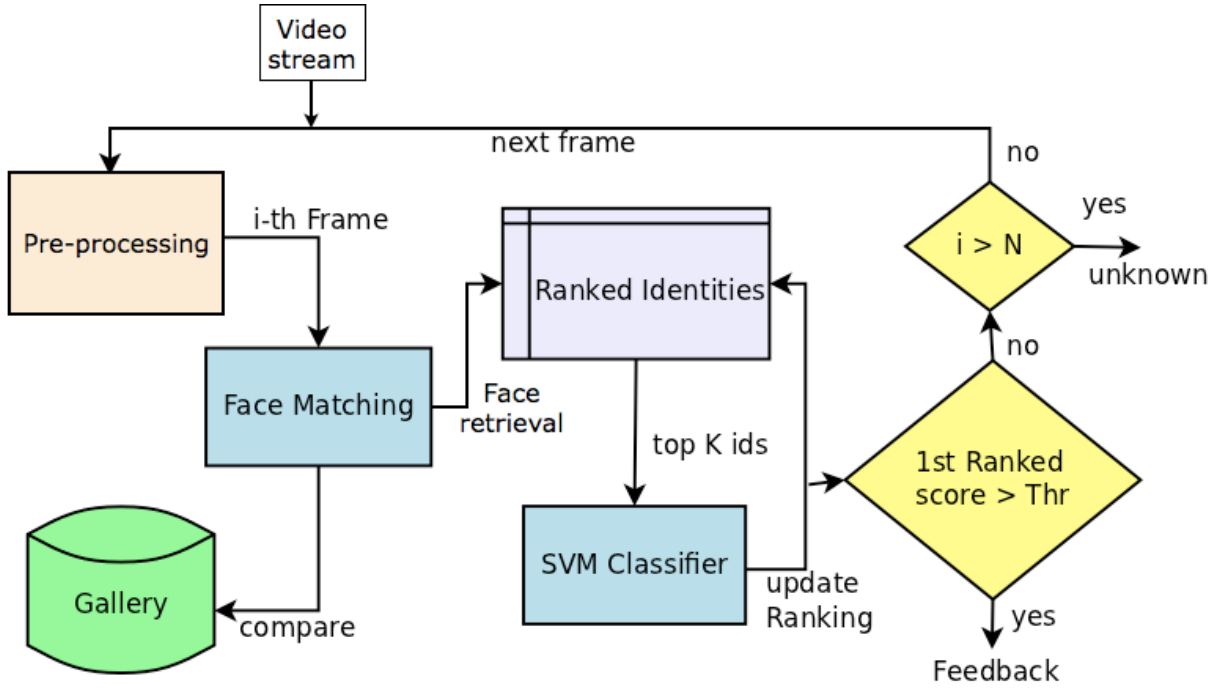


Figure 4.1: Scheme of the proposed method.

In the remainder of this chapter, we will describe in detail each module of the pipeline.

4.2 Face Detection

In this work we adopt a robust real-time face detection method based on a cascade approach [DDMOV09, DMOV09]. First the video frame is segmented with a simple skin detection module, then potential skin regions are analysed using a face detection algorithm tolerant to illumination variations.
The face detection is based on a feature selection plus classification pipeline: the feature selection procedure, applied to a high dimensional feature vector obtained from an over-complete set of rectangle features [VJ04], allows him to automatically build from data an appropriately compact representation of the object of interest (faces in this case). The classification is implemented as a cascade of Support Vector Machine (SVM) classifiers [Vap99] which gives fast detection on the image regions positive to the skin detection test. This results in a fast rejection of non-face area of the image, allowing real-time detection.

Thanks to the data-driven nature of the whole pipeline, the same architecture can be adopted to

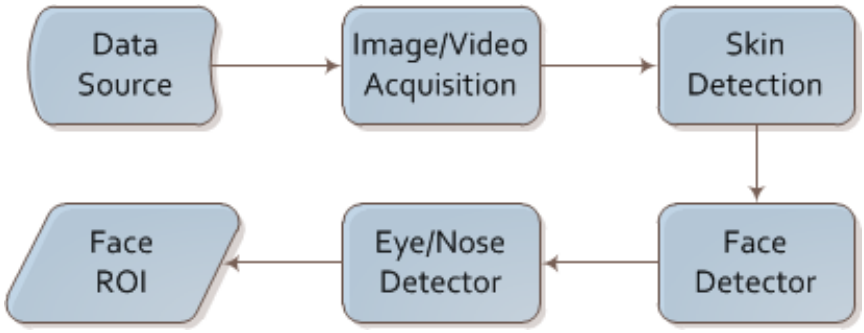


Figure 4.2: Outline of the preprocessing chain used for real-time face detection.

train an eye and a nose classifier. Eyes and nose detection are applied to the sole face region, allowing us to identify the most significant face images for subsequent recognition with a resulting coarse image registration.

4.3 Individual Models

During the training phase, for each subject that is enrolled in the system we need to model his/her appearance starting from a set of n face images $I_1, \dots I_n$, extracted from one or more videos. Since the system deals with uncontrolled settings, the face description needs to be robust to variations. Motivated by their useful properties (as described in Section 3.4.3) we decided to adopt uniform LBP to describe the faces.

More in details, for each identity we need to create the gallery for the retrieval module, and, if n is large enough, to train a One-vs-All SVM classifier for the classification module.

The training phase, described in details in the remainder of this section, produces for each identity $\mathcal{I}$ a model $(\mathcal{G}^{\mathcal{I}}, \mathcal{M}^{\mathcal{I}})$ where $\mathcal{G}^{\mathcal{I}} = \{g_i^{\mathcal{I}}\}_{i=1}^{G}$ is the image gallery, while $\mathcal{M}^{\mathcal{I}} = \{\alpha_i^{\mathcal{I}}, \mathbf{x}_i^{\mathcal{I}}\}_{i=1}^{n_{SV}^{\mathcal{I}}}$ is the model obtained by the linear SVM.

4.3.1 Choosing the Gallery

From the point of view of the face retrieval procedure, the number and the quality of the available labeled images play an important role. Ideally, for each identity the images that make up the gallery should be heterogeneous enough to enclose the appearance variability of the subjects. Additionally the number of images stored should be the same among the different identities to reduce the probability that an identity could be often preferred to the others because of the greater number of samples. To guarantee those conditions, for each identity enrolled in the system, we

automatically select the G (where G is equal for all the identities) most representative samples from a video (or a set of videos) to make up its gallery, adopting the method proposed in [HP04]. Formally, given a training face sequence

$$S_{\mathcal{I}} = \{s_1, s_2, \cdots, s_N\},$$

for the identity $\mathcal{I}$, we want to select the most representative samples

$$\mathcal{G}^{\mathcal{I}} = \{g_i^{\mathcal{I}}\}_{i=1}^{G}$$

to create the gallery for the identity I.
The approach proposed by [HP04] to solve this problem is based on two steps: first reproject the face images in a low-dimensional space in which similar faces are close to each other, then applying the K-means clustering [M$^+$67]. The centroids of the detected cluster will be the selected samples.

The detection of the low-dimensional space is obtained using the Local Linear Embedding (LLE) [SR03], an unsupervised learning algorithm that computes low dimensional, neighborhood preserving embeddings starting from high dimensional data. Assuming that the data are sampled from an underlying manifold, they are mapped into a global coordinate system of lower dimensionality. Considering a matrix X in which each column vector represents the intensity values of a face from a sequence S_I, the LLE algorithm, as described in [HP04], consists of the following steps:

1. Find the nearest neighbors of each point X_i;
2. Compute the weights W_{ij} to reconstruct each data point from its neighbors, minimizing the reconstruction error:
$$\sum_{i=1}^{length(S)} \|X_i - \sum_{j \in neighbors(i)} W_{ij} X_j\|^2$$
3. Compute the embedding Y of dimensionality $d \ll D$, with D dimension of the input data, minimizing:
$$\sum_{i}^{length(S)} \|Y_i - \sum_{j \in neighbors(i)} W_{ij} Y_j\|^2.$$

Once the mapping has been computed, clusters of coherent faces are detected in the new low-dimensional space by K-means. More in details, being G the desired cardinality of the gallery of an identity, x_i the low-dimensional mapped face images and $x_i^{(j)}$ a sample belonging to the j-th cluster, the final detected G centroids c_j minimize the following objective function

$$J = \sum_{j=1}^{G} \sum_{i=1}^{N} \|x_i^{(j)} - c_j\|^2$$

From each cluster, we then extract the example that is closer to its centroid. In the case in which K-means may detect $K < G$ clusters, we extract the missing $K-G$ items repeating the extraction procedure on the non-empty clusters. In Figure 4.3 we report an example of $G = 5$ frames, that are the representatives of the five clusters, selected from a sequence in which a subject is moving without any constraint on the pose.

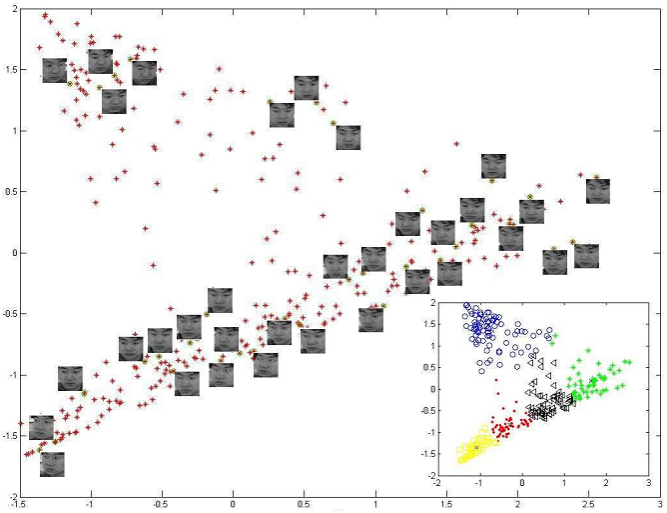


Figure 4.3: An example of low dimensional manifold detected from a set of images. In the right bottom corner, the output of a K-means clustering on the manifold efficiently separates the different appearance of the subject [HP04].

4.3.2 Models for Face Classification

In this work, we cope with the problem of face classification using One-vs-All (OVA) scheme, built training N different binary classifiers, each one trained to distinguish the examples belonging to a single class from the examples in all remaining classes. In [RK04] it has been shown that the OVA scheme achieves the same performances of the exhaustive All-vs-All (AVA) approach, in which $\binom{N}{2}$ binary classifiers are trained, each one separating two classes. It is straightforward that adopting the OVA instead of the AVA scheme reduces the computational load during the training and prediction phase, since the number of classifiers is smaller then in AVA.

Having to train a binary classifier for each identity $\mathcal{I}$, we set up a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n_\mathcal{I}}$ where $\mathbf{x}_i \in \Re^d, d > 0$ are the spatially enhanced histograms extracted from each face, while $y_i \in \{-1, 1\}$ is a label that specifies whether the example is a positive (face of $\mathcal{I}$) or not. The set is used to train and tune a binary linear SVM classifier[1]. SVMs perform pattern recogni-

[1]In our experiments, we adopt the library LIBLINEAR, `http://www.csie.ntu.edu.tw/~cjlin/`

tion between two classes by finding a separation hyperplane characterized by having maximum distance to the closest points in the training data, named support vectors. In the hypothesis of linearly separable data, the optimal separating hyperplane (OSH) for an identity $\mathcal{I}$ is defined as:

$$h_{\mathcal{I}}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} + b$$

where the coefficients α and b are solutions of a quadratic programming problem. The classification result for $\mathbf{x}$ is given by the sign of $h(\mathbf{x})_{\mathcal{I}}$.

4.4 Module for Face Retrieval

In this section we will describe the face retrieval method we adopted in the system, first introducing its basic mechanism on a single query image, then extending it to the video case.

Given a query image q, that is the LBP description of the face region extracted from an input image, we compare it w.r.t. the images stored in the gallery $\mathcal{G}^{\mathcal{I}}$, for each known identity $\mathcal{I}$. The comparison is performed by using a similarity measure between histograms, the normalized histogram intersection [SB91]. Given two LBP histograms I and M

$$H(I, M) = \frac{\sum_{j=1}^{n} \min(I_j, M_j)}{\sum_{j=1}^{n} M_j} \tag{4.1}$$

measures the amount of corresponding LBP labels between two images. The normalization factor enables to obtain a match value between 0 and 1, where 1 means that the two histograms are identical. The gallery images are then ranked according to the value obtained by Equation 4.4 and the first n_m nearest neighbors of q, said r their rank, they are associated to a score $s(r)$, proportional to r. Each identity is then associated to their accumulated similarity score as follows:

$$R(\mathcal{I}) = \sum_{r=1}^{n_m} \delta(r, \mathcal{I}) s(r) \tag{4.2}$$

with

$$\delta(r, \mathcal{I}) = \begin{cases} 1 & \text{if image of rank r} \in \mathcal{I} \\ 0 & \text{otherwise} \end{cases}$$

We now extend the system to consider the temporal coherence in the data, in order to perform a robust face retrieval.

`liblinear/`.

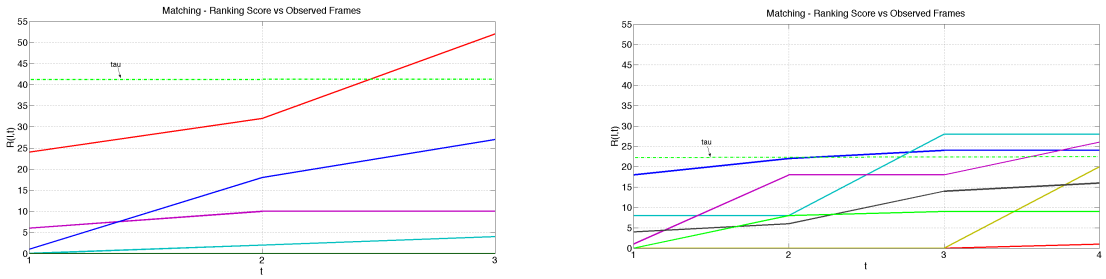


Figure 4.4: Accumulated ranking scores. Left: a simple case, where only one identity grows as time goes by. Right: a noisy case, where various identities compete.

Instead of considering each frame as a new query to the system, it is in fact possible to take advantage of the temporal component of the signal as well as the spatial one.

We can extend Equation 4.2 to consider the temporal component. Then, at time t, each known identity is ranked according to the accumulated similarity score from the first frame $t - T$ up to the query frame at time t, using the following:

$$R(\mathcal{I}, t) = \sum_{k=t-T}^{t} \sum_{r=1}^{n_m} \delta_k(r, \mathcal{I}) s(r) \tag{4.3}$$

where

$$\delta_t(r, \mathcal{I}) = \begin{cases} 1 & \text{if image of rank r} \in \mathcal{I} \\ 0 & \text{otherwise} \end{cases}$$

At time t, the retrieved identity $Id_{Ret}(t)$ is then computed according to the following:

$$Id_{Ret}(t) = \begin{cases} \arg\max_{\mathcal{I}} R(\mathcal{I}, t), & \text{if } \max_{\mathcal{I}} R(\mathcal{I}, t) > \tau \\ unknown, & \text{if } \max_{\mathcal{I}} R(\mathcal{I}, t) \leq \tau \end{cases} \tag{4.4}$$

where τ is a threshold on the similarity score, that allows to adjust the sensitivity of the retrieval module. It is worth noting that Equation 4.4 could be used as an output in cases in which the classification module, introduced in the next section, is not available. This may occur when, for instance, a new identity is enrolled in the system and the user would like to recognize him/her without waiting having to wait for the training of the SVM model.

4.5 Identity Validation

Although effective descriptors are available, as the case of LBP, it is well assessed that face retrieval generalizes poorly when a single feedback is required (top rank) as it is for face recognition. Low dimensional signals, typical of cameras on mobile devices, may produce noisy representations. Also, the performances worsen when the probe data significantly differs from the gallery and in the presence of a high number of unknown identities. In these cases, we may observe a shrinking effect on the accumulated similarity scores of different identities, i.e. multiple identities are retrieved τ (see Figures 4.4).

Machine learning methods can be used to improve the prediction ability of retrieval when the available training data are rich enough. The Support Vector Machines (SVMs) are well known for their ability to generalize well even in the presence of a relatively small number of samples, and thus appear to be appropriate for our task — where some classes may be poorly represented.

We propose a procedure that refines the frame by frame ranking provided by the face matching module by means of linear SVM classifiers in a One-vs-All schema. We consider a sorting of the scores obtained by the matching phase and select the subset R_t of K top-ranked identities after the matching. Then, for each of them we compute a new score applying the following updating rule

$$R_{Ver}(\mathcal{I},t) = \begin{cases} \alpha R(\mathcal{I},t), \alpha > 1 & \text{if } \forall id \in R_t \setminus \{\mathcal{I}\} \\ & h_{\mathcal{I}}(f_t) > h_{id}(f_t) > 0, \\ \alpha R(\mathcal{I},t), 0 < \alpha < 1, & \text{otherwise.} \end{cases} \tag{4.5}$$

where $R(\mathcal{I},t)$ is the accumulating ranking score for identity $\mathcal{I}$ at frame t (see Equation 4.3), $h_{id}(\cdot)$ is the confidence value obtained classifying the face data f_t at time t using the SVM classifier trained on the identity id and α is a scalar. Then we apply the same heuristic of Equation 4.4, replacing $R(\mathcal{I},t)$ with $R_{Ver}(\mathcal{I},t)$, to return the retrieved identity.

4.6 Experimental assessment

In this section we discuss on the experiments we performed to assess the combined face matching and classification method we propose.

Our aim is to show the robustness of our method as the setting changes, with respect to variability and different quantity of the data. A limit of the other approaches we consider, sole face matching and SVMs, is that they are effective in specific settings, with simple training set the first, thanks to the availability of many data the latter. Our approach instead adapts better to different settings, and thus represents a compromise solution.
We considered two sets of data of different complexities: the MOBO[2] dataset [GS01] and an

[2] http://www.ri.cmu.edu/publication_view.html?pub_id=3904

in-house acquired dataset we called R309.

In the experiments we conducted, we used the $LBP^{2u}_{(16,2)}$ descriptor, K was set equals to 3. The number of considered nearest neighbors n_m was set to 5. Concerning the scores, we tested several approaches and the best results have been obtained using binary weighting system $s(r) = 2^{n_m - r + 1}$.

In the remainder of the section we discuss in details the datasets properties and the experimental evaluations we conducted.

4.6.1 MOBO dataset

Originally proposed for action recognition, the MOBO dataset has been widely adopted as one of the few public datasets available for video-based face recognition. We chose it as the public and free dataset more appropriate to our purposes, since it contains low-resolution videos of quasi-frontal faces (see Fig. 4.5), depicting people moving and walking [GS01]. However, it is

Figure 4.5: Example of extracted frames from videos of two identities in the MOBO dataset.

characterized by small variations in the lighting condition and subject poses.
The number of subjects is 25 and are observed while performing 4 different actions (slow-walk, fast-walk, ball and incline). In our experiments we created the training sets by randomly sampling videos from 2 out of the 4 kind of videos for each identity, using on average 150 frames per individual. The rest of the videos is used for test.

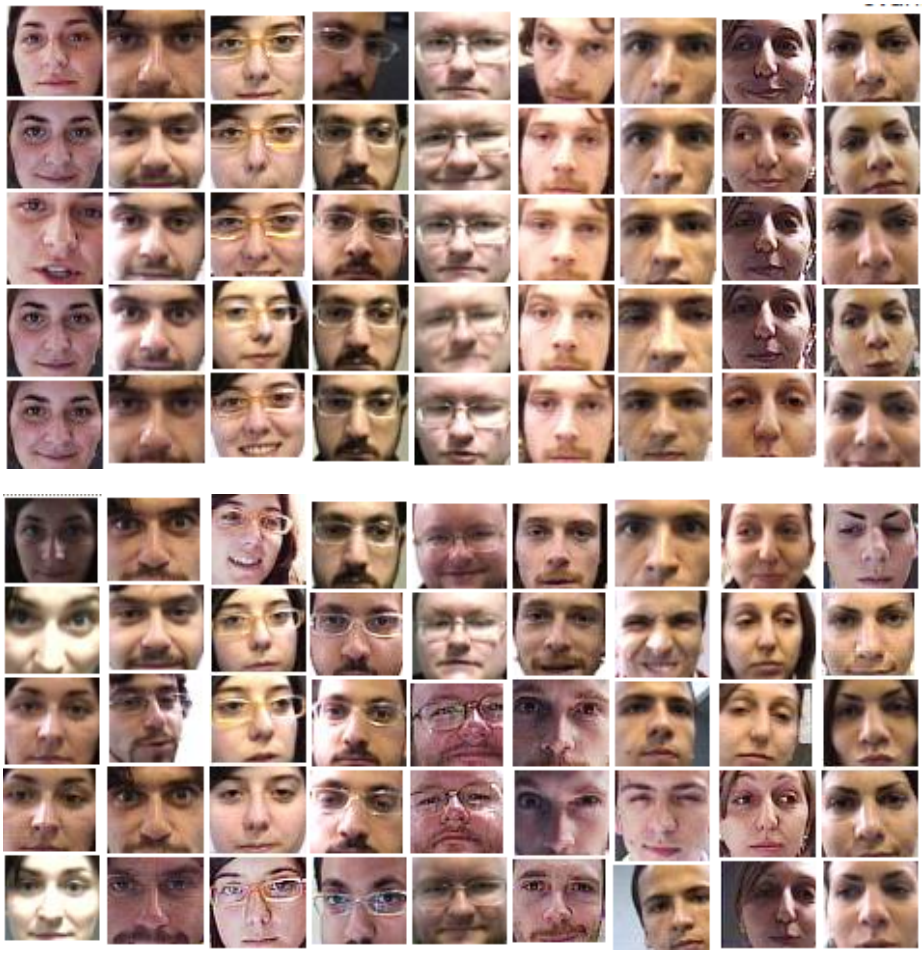


Figure 4.6: Examples of training sets from the R309 dataset (above homogeneous, below heterogeneous).

4.6.2 R309 dataset

The R309 dataset collects indoor videos acquired in unconstrained conditions: the pose and the behavior of the volunteers are entirely unconstrained and very natural. The subjects have been recorded under various lighting conditions and in two different environments (see Fig. 4.7). Currently the R309 dataset consists of 9 identities with an average of 1200 frames for each identity acquired along a period of time of 7 months.

We carried out the analysis considering both *homogeneous* training sets, in which each individual is represented by faces with similar pose and lighting conditions, and *heterogeneous* training sets, characterized by an higher variability of the data. As for the first setting, we considered an average of 74 frames collected from 3 to 5 videos for each identity (see an example in Figure 4.6,

above), while for the second we selected six videos for each identity, 140 frames per identity on average (to show the variability in this case, in Figure 4.6, below, we report the first five images selected by the LLE method – see Sec. 4.3.1). In both settings, the remaining videos not included in the training set have been used to create the test sets.

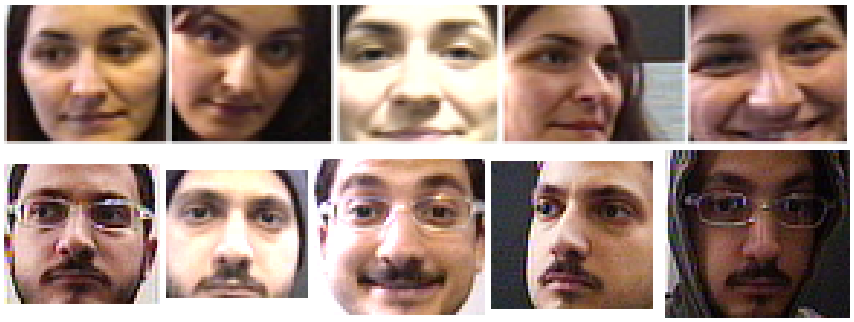


Figure 4.7: Example of data variability of two identities in the R309 dataset.

4.6.3 Experimental analysis

We start off by considering the MOBO dataset. Table 4.1 reports the recognition rates obtained with our approach compared to the performances of other state-of-art methods [HMPL07]. Let us notice that the MOBO dataset is characterized by a low variability of the data, and thus we end up with training and test sets with very similar properties. For this reasons, the performances are high in general, also for simple approaches as plain matching – see Fig. 4.8, where we report the recognition rate obtained training the system on a single type of video for each identity and adopting different combinations of the proposed methods.

From the results we obtained (Table 4.1), we state that our approach performs better than computationally expensive models found in literature, such as ARMA and HMM. It is also worth noting that the proposed time refined face retrieval approach alone outperforms most of the state of the art approaches. The method proposed by [ZP07], instead, achieves better performances, but the nature of the descriptor itself introduces a delay of several frames between the detection of the face and the description of the datum thus making the system less reactive.

With higher data variability, as in the case of R309, considering training and test sets with different properties worsens the performances. As it can be observed in Figure 4.11, where we are considering a homogeneous training set opposed to a generic test set, the recognition rates are consistently lower than in the case of more similar settings (Figure 4.12), due to the lack of information during training. In this second case, it can be noticed that the final classification step allows for better results than the sole matching and that the temporal analysis applied to the score updating consistently helps, especially in more complex cases, as with poor galleries (in

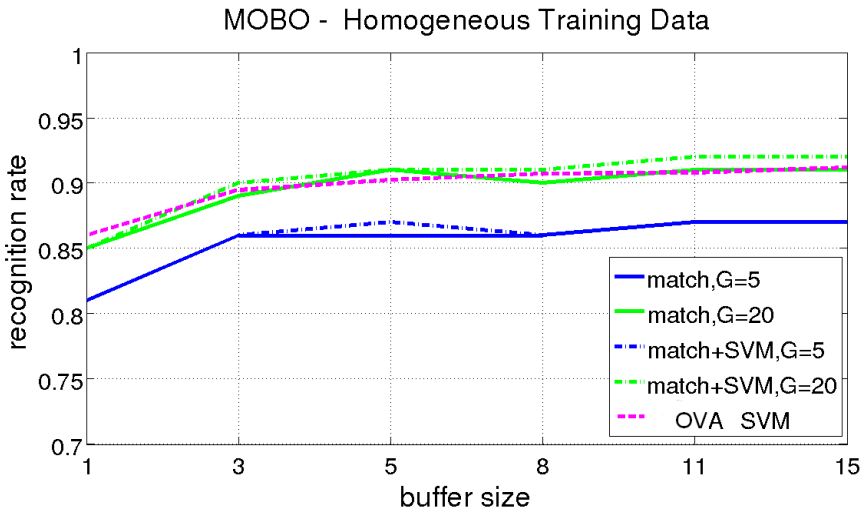


Figure 4.8: Recognition rate on the MOBO dataset varying the number of observed frames with *homogeneous* training data. Each curve represents the performance of the face matching with (dotted lines) and without (full line) the proposed feedback approach, given different gallery cardinalities.

the plot G=5). It can also be observed how we reach the One-vs-All SVMs performances when the gallery is rich enough (G=30 in the experiments).
In Figure 4.10 we show the results obtained testing the system in the extreme case in which one video per identity is available, both for the training of the classifiers and the gallery creation. The overall recognition rate obtained in this case are not competitive. From the outcome of the experiment, we notice that the combination of the retrieval with the classification module produces worse results than using the retrieval module alone. Moreover, when adopting the combined approach, observing more frames, i.e. a larger buffer size, worsen the recognition rate.

The reliance degree on the training set size is analyzed in Figure 4.13. As expected, SVMs suffer from the adoption of small training sets, being unable to appropriately describe the data, while are expected to provide the best results as the size increases. However, in this cases, their computational load increases as well. Let N be the number of individuals enrolled in the system, then if for each individual $\mathcal{I}_i$, $n_{SV}^{\mathcal{I}_i}$ is the number of support vectors of the corresponding SVM model, the number of requested operations is proportional to $\sum_{i=1}^{N} n_{SV}^{\mathcal{I}_i}$. In our case instead this number depends on the size G of the gallery, and on the SVM models but only for the K identities retrieved by the matching, according to $\sum_{i=1}^{N} G + \sum_{j=1}^{K} n_{SV}^{\mathcal{I}_k}$, that is $N \cdot G + \sum_{j=1}^{K} n_{SV}^{\mathcal{I}_k}$. Since, thanks to the matching, it holds in general that $K << N$, our approach is consistently more efficient.

Table 4.1: Comparison among results on the MOBO dataset (in bold the results we computed; all other values are taken from [HMPL07]). Training set: 250 images, G=20, buffer = 15 frames.

METHOD	RECOGNITION RATE
PCA	87,1%
LDA	90,8%
LBP [AHP06]	91,3%
HMM [LC03]	92,3%
ARMA [Agg04]	93,4%
One-vs-all SVM	**90,8%**
VLBP [ZP07]	90,3%
EVLBP + AdaBoost [HMPL07]	97,9%
Retrieval	**92,5%**
Our method	**93,8%**

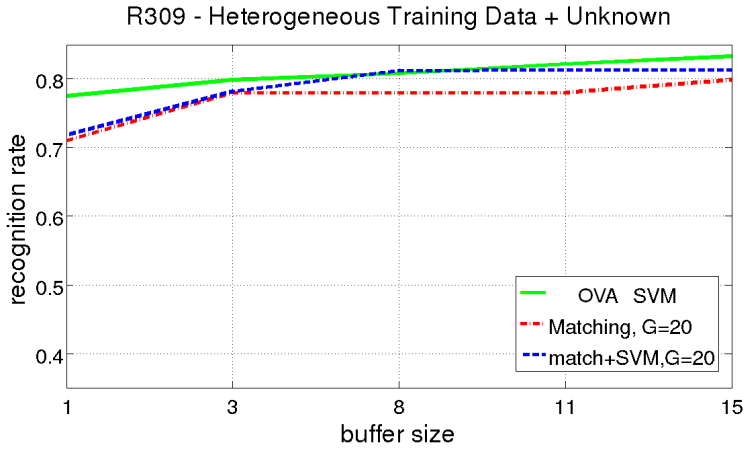


Figure 4.9: Recognition rate on the R309 dataset varying the number of observed frames with *heterogeneous* training data and adding unknown identities in the test set.

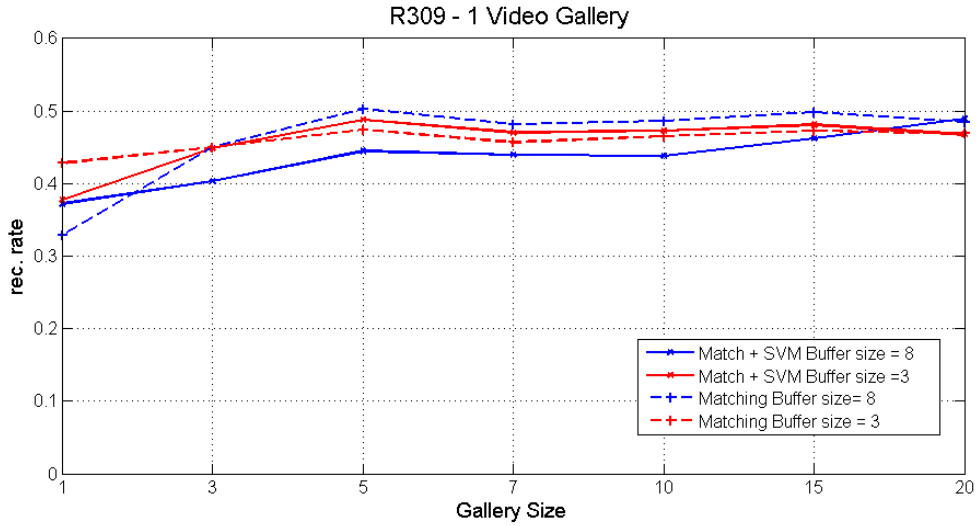


Figure 4.10: Recognition rate on the R309 dataset training the system using one video per identity.

For instance, in the R309 dataset, although the sparsity of the SVM solution allows us to perform, on average, the 37% of comparisons (since for an average training set size of 140×2 we obtain an average number of 100 support vectors), the computational advantage of our approach, with K=2 and N=9, is apparent.
We finally report the performances of our method (compared to the One-vs-All SVMs) with respect to the presence of a 15% on average of unknown identities in the test set, to consider a very likely setting for the application case we have in mind. The plots in Figure 4.9 shows the robustness of our method and the benefit of the temporal analysis that allows for a refinement of the system feedback.

4.7 Discussion

In this chapter we presented a real time face recognition system that combines modules of retrieval and classification. The coherence provided by the temporal evolution of the video data is exploited to refine a recognition score and thus increase the robustness of the system feedback. The approach we propose is a viable solution for settings characterized by variable quality and quantity of data available for each identity to be modeled. Although not limited to, a potential application is within the automatic tools to support impaired people with limited visual capabilities, to help the interactions in familiar or unfamiliar environments by possibly recognizing known individuals.

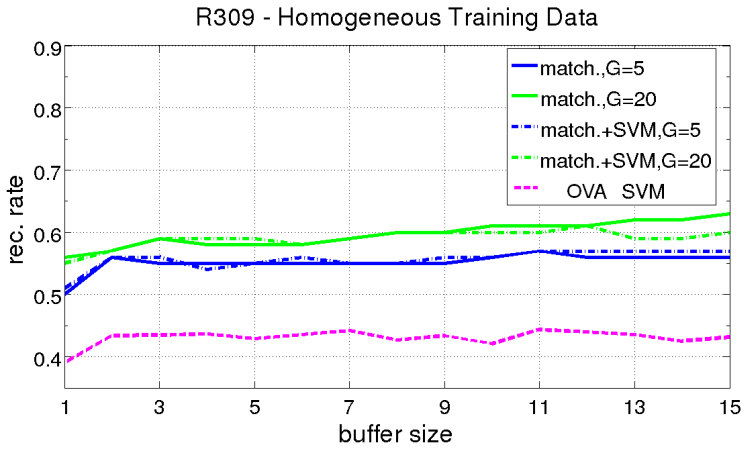


Figure 4.11: Recognition rate on the R309 dataset varying the number of observed frames with *homogeneous* training data. Each curve represents the performance of the face matching with (dotted lines) and without (full line) the proposed feedback approach, given different gallery cardinalities.

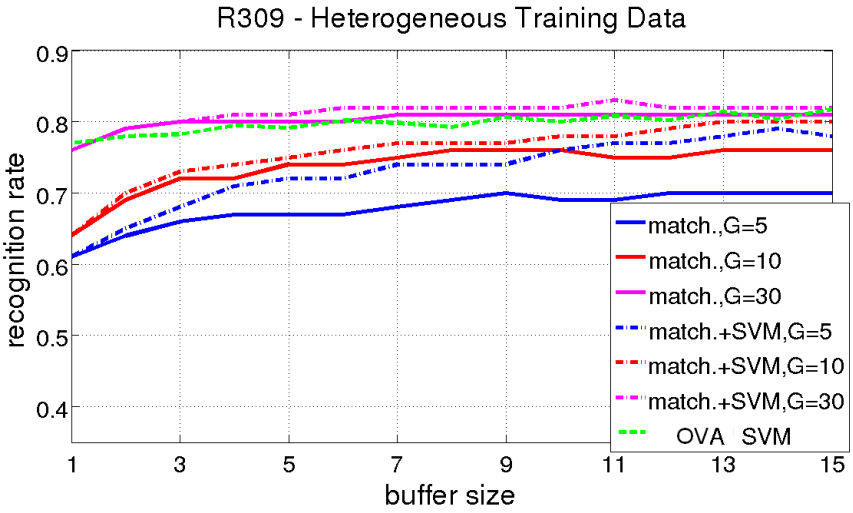


Figure 4.12: Recognition rate on the R309 dataset varying the number of observed frames with heterogeneous training data. Each curve represents the performance of the face matching with (dotted lines) and without (full line) the proposed feedback approach, given different gallery cardinalities.

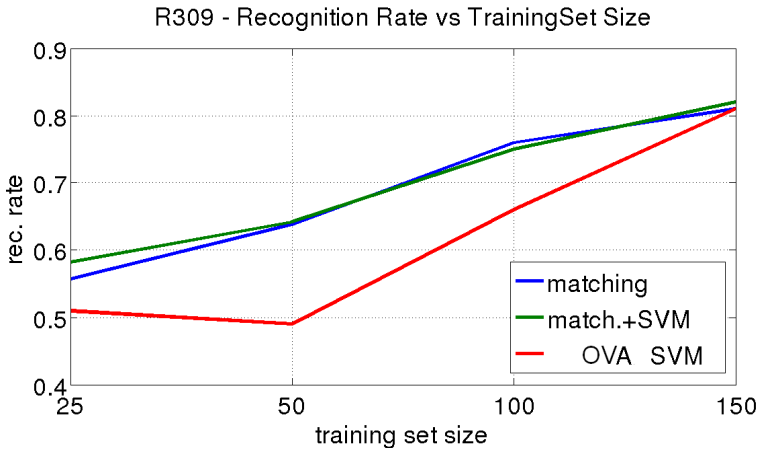


Figure 4.13: Recognition rate on the R309 dataset varying the size of the *heterogeneous* training set, using gallery size G=20 and 15 observed frames.

We tested our method on two different data sets (a benchmark dataset and one acquired in-house) characterized by different complexities and variability concerning lighting conditions and poses constraints. The experimental analysis showed that the method we propose competes with state-of-art approaches and consistently represents a compromise solution able to adapt to settings of variable properties. Also, the temporal analysis for updating the individual's scores helps to improve the reliability of the system, especially when the models are relatively poor. We also showed that the computational load of our method is very slightly afflicted by the growing of the training sets and that it is bounded by the total size of all galleries. The work has been presented in [FNO12].

Chapter 5

Simultaneous feature selection in face recognition

In this chapter we present a novel regularized simultaneous feature selection method and verify its effectiveness in the context of face recognition. We show that with the proposed feature selection method we obtain a very sparse representation of data while at the same increasing the recognition rate. In the remainder of this chapter, after an introduction about the problem of feature selection in Section 5.1, in Section 5.2 we give an overview of the state of the art about feature selection in face recognition. In Section 5.3 we introduce the proposed method for the simultaneous selection of features, while in Section 5.4 we describe the procedure we adopted to evaluate the feature selection performances. Finally in Section 5.5 we describe the results we obtained applying our method to the problem of face recognition on three different datasets.

5.1 Introduction

In the last decade the extraction of image patterns by pooling within (overcomplete) sets of local descriptors has received attention. This procedure has been shown particularly effective for applications to object detection and recognition and image categorization [AHP06, DMOV09, VJ01]. Although local features are characterized by useful properties, their adoption induces the problem of the selection of the right parameters (for instance, the number of samples and the radius of the LBP operator or the Gabor wavelets' parameters [D$^+$85]) for the extraction of representative patterns that convey important information about the problem. Decreeing the parameters of a local descriptor can be considered equivalent to detecting the subset of features – from the set of the whole possible configurations – that maximizes a performance measure, for instance the accuracy of a classifier. The selection of said parameters can be performed empirically by sampling the features space, possibly exploiting some a-priori knowledge about the problem.

Another common approach to the selection problem is to learn the most effective representation of the starting from the data themselves, that represent the available prior on the problem. Adopting formal methods for feature selection [FS95, Tib96], starting from an overcomplete set of features, it is possible to automatically select the most significant subset of features in order to achieve good recognition performances [VJ04, DMOV09], and hence sparsify the data representation.
In [AHP06], that represents the state of the art of face recognition based on LBP, it has been empirically determined that the most efficient way to describe face data with LBP is to divide each face superimposing to it a regular grid. We will show that a regular partition of an image does not represent the best choice when the faces are acquired in uncontrolled settings, where people move freely in the camera field of view and consequently significant registration errors may occur.

To this purpose, we propose an efficient face recognition method, which obtains a very sparse representation of data by means of regularized feature selection. We start from an overcomplete description of face images based on the use of Local Binary Patterns (LBP) [ZHL$^+$05, AHP04]. Then, we apply feature selection to select the most discriminative elements of the face descriptors and improve the computational efficiency of face recognition at run time.

We propose a new formulation of the Group LASSO functional [Bak99] – we call *Multi-Class Group LASSO* – to directly cope with a multi-class setting while capturing the block-structure of LBP. The new formulation allows us to select groups of features that simultaneously discriminate among all the identities.

5.2 State of the art

There are some papers in literature that cope with the problem of selecting features in the face recognition domain starting from local descriptors.
In [LCA$^+$06], starting from an overcomplete LBP description of near infra-red (NIR) face images, the authors adopt an AdaBoost learning procedure to select the most significant regions to improve their methods for face detection and face recognition. A similar approach is adopted in [ZHL$^+$05] but on gray-scale images. The work in [HLW05], instead, introduces a learning method called Jensen-Shannon Boosting (JSBoost) algorithm for feature selection and its effectiveness is demonstrated in the specific case of face recognition using overcomplete LBP descriptions. JSBoost incorporates Hensen-Shannon (JS) divergence in the learning procedure of Adaboost. The motivation for the introduction of the JS divergence [RPTB01] is that it provides a better measure of dissimilarity between samples from two different classes. The features are selected by maximizing the JS divergence, and are adopted to compute the best weak classifiers. The final classifier is then obtained combining the weak classifiers by minimizing the recognition error. In [SBBW05], the authors consider Gabor features and propose to extend AdaBoost

by incorporating mutual information in it, in order to eliminate redundancy in the descriptions. Such redundancy elimination is obtained by excluding the classifiers that carry information already captured by other selected classifiers. A genetic based approach for feature selection is described in [SDB+11]. Here the selection of the region in which compute LBP histograms is made by means of Genetic & Evolutionary Computing (GEC) [ANM06]. Starting from LBP descriptions that are unevenly distributed and computed from overlapping patches, the method achieves the minimization of the number of patches while increasing the recognition accuracy. All the previously cited papers cope with the problem of feature selection adopting methods tailored for binary problems, as for instance AdaBoost is. A typical approaches to convert a multiclass face recognition problem to a binary one is to resort to the intra-personal and extra-personal difference methods proposed by Moghaddam and Pentland [MJP00]. Given two images from the training set, their feature vector difference would be put in the intra-personal difference class if they belong to the same class, otherwise it will be labeled as an extra-personal difference. In this way the feature selection objective is to detect the best set of features that could discriminate if two images belong to the same class or not.
In the case of LBP descriptors, for each pair of images I_A and I_B the feature vector difference is obtained computing the χ^2 distance between corresponding LBP histograms. Formally, for each pair of images a vector x_i is computed and its elements are χ^2 distances:

$$x_i = \left[\chi^2(I_A^1, I_B^1), \chi^2(I_A^2, I_B^2) \cdots \chi^2(I_A^L, I_B^L)\right].$$

The associated label will then be $y_i = -1$ if I_A and I_B belong to the same class, $y_i = 1$ otherwise.

5.3 Regularized multi-class feature selection

In this section we describe the multi-class feature selection method we propose. Let us first formulate the *feature selection problem*: given a training set $(\mathbf{x}_i, y_i)$, with $\mathbf{x}_i \in X \subseteq R^m$ $y_i \in R$, and a dictionary $\mathcal{D} = (\phi_j)_{j=1}^p$ which is a collection of atoms or features, we consider a generalized linear formulation of the input-output relationship

$$\sum_{j=1}^{p} \phi_j(\mathbf{x}_i)\beta_j = y_i \tag{5.1}$$

or, in matrix form $\Phi\beta = \mathbf{y}$, where Φ is the feature matrix defined as $\Phi_{ij} = \phi_j(\mathbf{x}_i)$ (of size $n \times p$, n number of examples and p number of features), $\mathbf{y} \in \{-1, +1\}^{n\times 1}$ is the output vector, and $\beta \in \Re^{p\times 1}$ is the vector of weights. The goal of feature selection is to find a *sparse* β which approximates well the input-output relationship.

Typically, in feature selection applications the dimensions of the dictionary matrix Φ are large, hence algebraic solution to the linear system of Equation 5.1 are unfeasible. Usually, in fact, the

number of features p is significantly larger than the dimension n of the training set, causing the system to be underdetermined; moreover in an overcomplete training set we expect redundant features and correlation between groups of features and such collinearities make the system ill-conditioned. A solution to the system can be found introducing some form of regularization, posing the problem as a penalized least-squares problem. Classical regularization methods, such as the Tikhonov regularization, are formalized as follows:

$$\hat{\beta} = \arg\min_{\beta} |\mathbf{y} - \Phi\beta|_2^2 + \sum_{i=1}^{n} |\beta_i|^2$$

The norm-2 penalty term used in the Tikhonov regularization ensures that the solution $\hat{\beta}$ will be smooth, but does not provide a selection of features. The norm-2 penalty, in fact, does not enforce the sparsity of the solution vector, i.e. to have few component different from zero. For this reason, in recent literature the quadratic penalties have been replaced by sparsity-enforcing ones, such as $L1$-norm [Tib96].

5.3.1 LASSO

The LASSO (Least Absolute Shrinkage and Selection Operator) problem [Tib96] consists of finding a sparse solution to the linear system of Equation 5.1 using an $\ell 1$ penalty as follows:

$$\hat{\beta} = \arg\min_{\beta} |y - \Phi\beta|_2^2 + \tau|\beta|_{\ell 1} \tag{5.2}$$

where the $L1$-norm represents a sparsity enforcing penalty. The parameter τ controls the balance between the misfit and the penalty term. In the case of sparsity enforcing penalty it also regulates the amount of shrinkage that is applied to the estimates $\hat{\beta}$, i.e. the sparsity of the solution.

The use of $L1-$norm penalty makes the dependence of LASSO solutions on $\mathbf{y}$ non linear. Many optimization algorithm has been proposed to solve this non linear problem. A simple iterative algorithm is represented by the following equation:

$$\beta^t = S_\tau[\beta^{t-1} + \Phi^\top(y - \Phi\beta^{t-1})] \;\; t = 0, 1, \cdots \tag{5.3}$$

with β^0 is an arbitrary initial weight vector, and

$$S_\tau(\beta_j) = \begin{cases} \beta_j - \tau sign(\beta_j) & \text{if } |\beta_j| \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

soft-thresholding operator that acts component-wise on the weight vector β^t. The iterative shrinkage-thresholding algorithm (ISTA) of Equation 5.3 has been proven to converge to a minimizer of 5.2 by [DDDM04], although the uniqueness of the solution is not guaranteed because of the non-strict convexity of $L1$-norm.

5.3.2 Group LASSO

Group LASSO was originally proposed in [Bak99] as an extension of LASSO [Tib96] to select groups of features, rather than considering each feature independently. In several application domains, in fact, a prior on the structure of each input datum is available, and may be profitably inserted in the feature selection step. It can be formulated as a regularized minimization problem:

$$\hat{\beta} = \arg\min_{\beta} \left(|\mathbf{y} - \Phi\beta|_2^2 + \tau \sum_{g=1}^{G} |\beta_{I_g}|_2 \right) \tag{5.5}$$

where $g = 1, \ldots, G$ is an index referring to a feature group, while $I_g \subseteq [1...p]$ is the set containing the positions in the input datum of the features of group g. The penalty term enforces the sparsity block-wise, that is the selection of features is performed in groups. To achieve the group selection, the minimization algorithm in Equation 5.3 can be still adopted, but we must first modify the soft-tresholder in Equation 5.4 so that it can perform the thresholding operation block-wise:

$$\tilde{S}_\tau(\beta_{I_g}) = \begin{cases} \beta_{I_g} - \tau\beta_{I_g} & \text{if } |\beta_{I_g}|_{\ell_1} \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

5.3.3 Multi-class Group LASSO

We now re-write the functional (5.5) to deal with the simultaneous selection of groups of meaningful features for a multi-class problem. Let us consider N binary problems $\Phi\beta_c = \mathbf{y}_c$, $c = 1 \ldots N$. Similarly to above, Φ is the features matrix of size $n \times p$, $\mathbf{y}_c \in \{-1, +1\}^{n \times 1}$ is the output vector, containing 1 in correspondence of examples of class c and -1 elsewhere. Finally, $\beta_c \in \Re^{p \times 1}$ is the weight vector for the class c. We formulate the Multi-Class Group LASSO (MC-GrpLASSO) as follows:

$$\hat{B} = \arg\min_{B} \left(|Y - \Phi B|_2^2 + \tau \sum_{g=1}^{G} |B_{I_g}|_2 \right). \tag{5.7}$$

Y is a $n \times N$ matrix composed as $[\mathbf{y}_1\mathbf{y}_2 \ldots \mathbf{y}_N]$, and, similarly, B is a $p \times N$ such that $B = [\beta_1\beta_2 \ldots \beta_N]$. The simultaneous selection of block of features is obtained modifying once again

the soft-thresholder in order to perform the thresholding on the submatrices of B detected by the group index g:

$$\tilde{S}_\tau(B_{I_g}) = \begin{cases} B_{I_g} - \tau B_{I_g} & \text{if } |B_{I_g}|_F \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{5.8}$$

where in place of the norm one, we now use the Frobenius norm on the matrix B_{I_g} that corresponds to the weight vectors of all the N classes for the group g. Modifying the thresholder in this way, we are able to simultaneously discard the blocks of features that are not relevant for all the classes at the same time.

5.4 The proposed pipeline

In this section we describe our approach to select and assess features for face recognition. We assume we have in input a training set $\{\mathbf{x}_i\}_{i=1}^n$ of n face images of N different identities. We first represent each image according to an appropriate dictionary of features based on Local Binary Patterns (LBP). Then, we apply MC-GrpLASSO to obtain a set of groups of features meaningful for all the N classes. Finally, we train a multi-class SVM classifier on the training set of images represented with the selected groups only. This allows us to obtain, at run time, a very fast representation and classification of each face image. We adopt an overcomplete dictionary of Local Binary Pattern (LBP) descriptors, whose use in face recognition is well known in literature. Similarly to [ZHL+05], we consider a grid of overlapping regions having different scale and aspect ratio (faces images are all rescaled to the size 40×40 pixels). We then extract uniform 8-bits LBPs [OPM02] and quantize their values with 59-bins histograms, accordingly to the original work. The final overcomplete description consists of 841 LBPs, thus each training image $\mathbf{x}_i$ is described by $841 \times 59 = 49,619$ features and the feature matrix Φ will have 49619 columns. Figure 3.9 describes the obtained feature vector in a simplified situation, with no overlapping regions.

We specialize the functional (5.7) to our case by grouping the features according to the image region they are associated to (see Figure 5.1). The main advantages of adopting a feature selection method that takes into account the structure of the features are two. On one hand, selecting groups of features simultaneously produce more significant representations of the data, while on the other hand there is the computational advantage deriving from the need to only compute one set of features for all the classes. To find a solution to our regularization problem we adopt an algorithm based on proximal methods [MRS+10], that have been demonstrated to provide effective optimization procedures to solve many regularized algorithms with convex non differentiable penalties. The regularization parameter τ of Equation 5.7 can be tuned to obtain solutions with a different level of sparsity. Finally we analyze the obtained solution B and select the regions

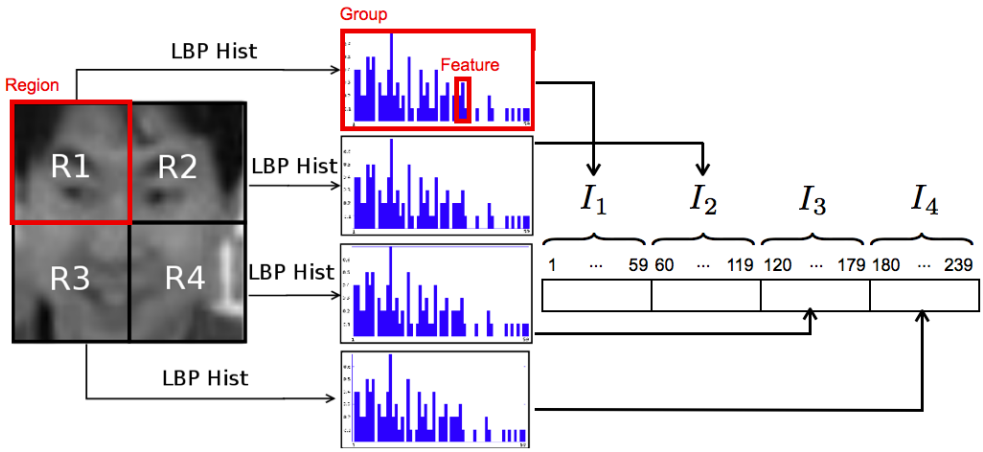
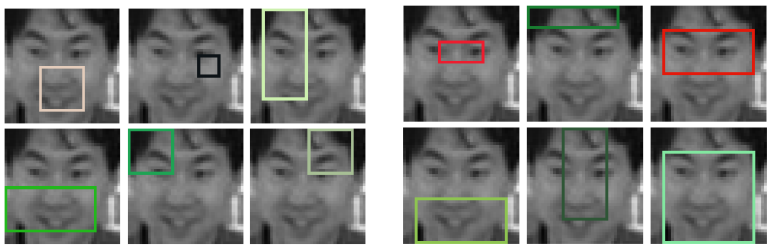


Figure 5.1: A visualization of our face descriptors, a concatenation of LBPs.



(a) (b)

Figure 5.2: MOBO dataset: subset of 6 selected features using a) LASSO and b) MC GrpLASSO.

associated to the rows of B with no zero values: in this way the features we select are meaningful for all the N classes.

To give a qualitative impression of the kind of features produced by our method, Fig. 5.2 reports two examples of a subsets of selected features obtained with MC-GrpLASSO in comparison with the ones obtained with LASSO. The features selected by MC-GrpLASSO (Figure 5.2.a) appear to be more meaningful than the corresponding set obtained by LASSO (Fig. 5.2.b).

Once we have found a sparse and meaningful representation for our multi-class problem, we train a multi-class classifier base on SVM and a *Winner-Takes-All* strategy [JW03].

Notice that it would have been possible to classify data directly via the solution of the MC-GrpLASSO, but the performance obtained with SVMs are usually higher because of the well known shrinkage effect of the weights of LASSO methods [CT07]. Also linear binary SVM can be implemented very efficiently and are the preferred choice for real-time classification systems.

We then train N One-vs-All binary SVM classifiers on the set of LBPs restricted to the regions

(or groups) selected. For the binary classifier of class c, the negatives are randomly sampled from all the classes j, $j = \ldots N$, $j \neq c$.

Treating the multiclass feature selection problem as a set of N individual feature selection problems leads to the detection of a projection operator $\pi_c : \Re^p \rightarrow \Re^{q_c}$, with q_c is the cardinality of $support(\beta_c)$, for each class $c = 1 \cdots N$. For a new datum $\mathbf{x}$, the resulting global classifier is then formulated as a combination of the N discriminant functions $h_1 \cdots , h_N$ of the binary classifier as

$$h(\mathbf{x}) = \arg \max_{c=1\ldots N} \{h_c(\pi_c(\mathbf{x}))\}.$$

For each example, thus, to obtain a result from the global classifier we need to compute N projections, one for each class in the problem, which brings a substantial computational overhead.

The method we propose, instead, performing a simultaneous feature selection, detects a unique projection $\pi(\mathbf{x})$ that will be shared by all the N classes. Hence, computing just one projection in place of N, we obtain a global classifier as follows:

$$h(\mathbf{x}) = \arg \max_{c=1\ldots N} \{h_c(\pi(\mathbf{x}))\}$$

with a resulting computational cost reduced by a factor N.

5.5 Experimental analysis

In this section we present the experimental evaluations of the proposed pipeline, with particular emphasis on the feature selection problem. The analysis is performed on three different datasets: two public benchmarks – *MOBO* [GS01] (see Section 4.6.1) and *Choke Point* [WCM+11] and a in-house dataset, *R309* (Section 4.6.2).

The *Choke Point* dataset is characterized by a high variability of image quality, appearance and lighting conditions. It includes videos of 29 subjects walking through two portals, acquired using an array of three cameras (Figure 5.5).

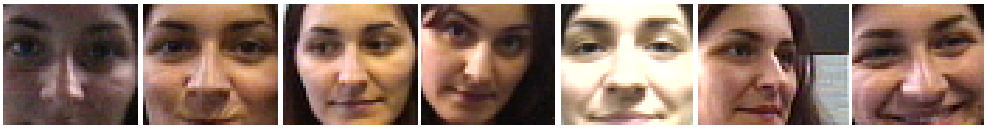


Figure 5.3: *R309* dataset: images of a same subject with different pose and lighting conditions.

For each method we compare the results obtained with our approach to multi-class feature selection and the ones obtained by reformulating the multi-class problem as a binary problem (as in [MJP00] for AdaBoost and LASSO binary feature selections respectively). According to this

Figure 5.4: Example of extracted frames from videos of two identities in the MOBO dataset.

Figure 5.5: Example of extracted frames for one identity from Choke Point dataset. Notice the low quality of the images and the lighting variability.

formulation the positive examples are obtained by computing the region-based χ^2 similarity of image pairs belonging to the same class, while the negative examples are obtained in the same way from pairs of inter-class images.

In the reminder of the section we discuss the experiments carried out. Each dataset is divided in training, validation, and test, in the proportions specified for each dataset. The training set is used for both feature selection and classifiers training. The validation set is used for tuning the parameters of the classifiers (that is, the regularized parameter of each binary SVM). The results reported in the tables have been obtained on the test set.

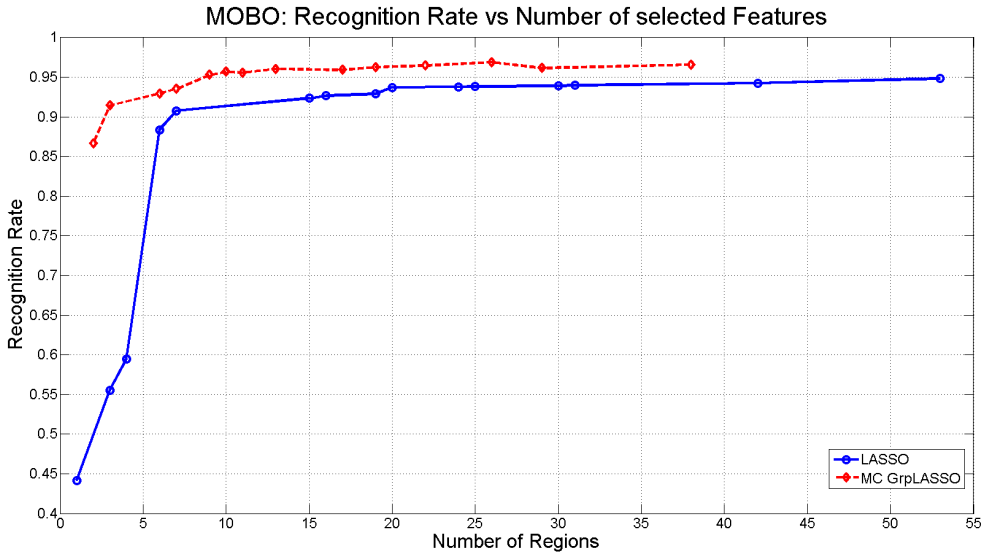


Figure 5.6: MOBO dataset: Recognition rate at different number of features using LASSO and MC Group LASSO methods

MOBO dataset. For this set of data we consider a training set of 50 examples per identity, extracted from a pool of 5 videos ($\#\text{training set} = 1000$). The validation set is collected following the same strategy of 5 other videos. The remaining videos are used to build the test set, made of 200 examples per subject ($\#\text{test set} = 4000$).

Figure 5.6 shows the recognition performance of MC-GrpLASSO at various sparsity degrees of the solution on the validation set. It is apparent how, with at least 20 features, the recognition rate is quite stable. This suggests that by adding features we simply obtain a more redundant description. Table 5.1 reports the recognition performances we obtained using a fixed grid of 7×7

Table 5.1: Comparison on the MOBO dataset

METHOD	# Features	TPR	FPR
fixed grid LBP [AHP04]	49	95,6%	0,2%
LBP + LASSO	25	94,1%	0,3%
LBP + AdaBoost	25	90,1%	0,5%
LBP + MC-GrpLASSO	**26**	**96,9%**	**0,1%**
LPB + MC-GrpLASSO	13	96,1%	0,2%

regions as a baseline method [AHP04]. The goal is to achieve comparable or better performances with a sparser representation. MC-GrpLASSO outperforms the baseline method and is also

sensibly better than LASSO and AdaBoost with an image description of approximately the same size. It is also worth to point out that the boosting in the performances, although less significant, persists using a set of features of cardinality around $\frac{1}{4}$ of those proposed in [AHP04].

Choke Point dataset. For these experiments we adopt the experimental protocol of [WCM$^+$11] to compare our performances with the results reported in the paper. In particular, in [WCM$^+$11] the authors compare the use of Multi-Region Histograms (MRH) [SL09] and LBP [AHP04] for face description, coupled with Mutual Subspace Method (MSM) [YFM98]. Tab. 5.2 summarized our comparative analysis. MC-GrpLASSO performance is well above the other methods. In Fig. 5.7 we also show the stability of the solution for a different number of selected groups.

Table 5.2: Comparison among results on the *Choke Point* dataset.

Recognition Method	Subset Selection Method	Rec. Rate
MRH + MSM	Asym_shrp [ARACGM08]	75.4%
	Gabor_asym [SLL09]	84.0%
	DFFS [BK05]	83.4%
	Proposed in [WCM$^+$11]	86.7%
LBP + MSM	Asym_shrp [ARACGM08]	70.5%
	Gabor_asym [SLL09]	74.5%
	DFFS [BK05]	74.6%
	Proposed in [WCM$^+$11]	75.8%
LBP + LASSO	–	93.1%
LBP + MC-GrpLASSO	–	**96.9%**

R309 dataset. We conclude our experimental analysis on the *R309* dataset. Tab. 5.3 shows how the overall performances on this dataset are significantly lower than for the other datasets. This is due to a consistent variability on the subjects appearances (see an example on Fig. 5.3), also considering that the acquisitions have been made on a temporal span of months.

In Fig. 5.8 we report the recognition rate obtained with different sets of features using the LASSO and MC-GrpLASSO method. It can be seen that MC-GrpLASSO provides better recognition performances compared to LASSO, even if in some cases (e.g. when the number of selected features is around 20) the two methods perform similarly.

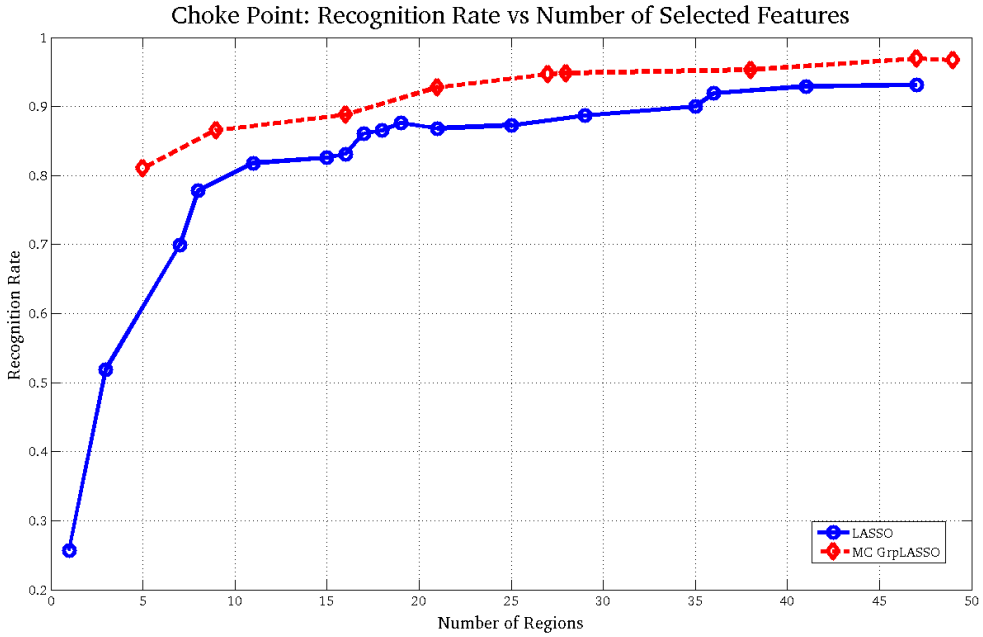


Figure 5.7: *Choke Point* dataset: Recognition rate at different number of features using LASSO and MC Group LASSO methods

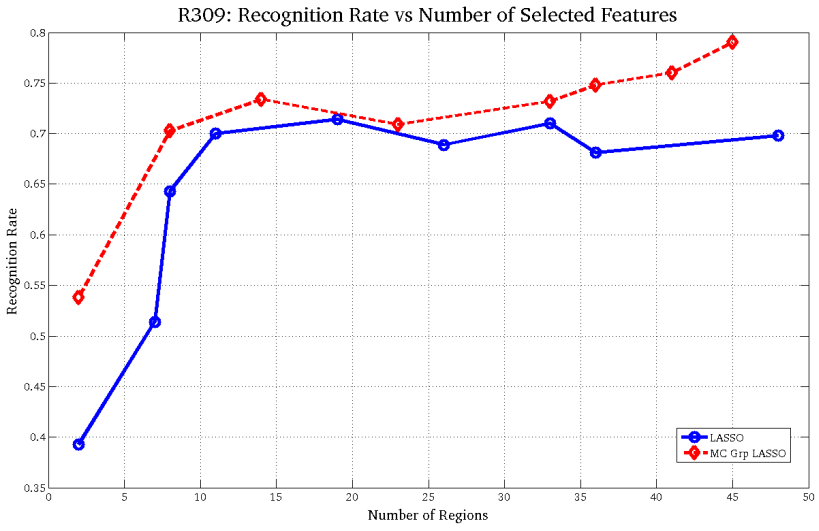


Figure 5.8: *R309* dataset: Recognition rate at different number of features using LASSO and MC Group LASSO methods

Table 5.3: Comparison among results on the R309 dataset

METHOD	# Features	TPR	FPR
LBP + MC-GrpLASSO	**45**	**79,6%**	**2,6%**
LBP + MC-GrpLASSO	14	73,4%	3,3%
$F.G.LBP^u$	49	71,6%	3,5%
LBP+LASSO	19	71,4%	3,6%

5.6 Discussion

In this chapter we presented a novel method for the simultaneous selection of features in multi-class problems. We proposed an extension of the Group-LASSO method , the *Multi-Class Group LASSO* to directly cope with a multi-class setting and exploiting the block-structure of features. Unlike previous approaches found in the literature, the new formulation allows us to simultaneously select groups of features for all the classes in the problem. We verified the effectiveness of the proposed method in the context of face recognition. We presented an extensive experimental analysis, both on benchmark datasets and on a complex set of data acquired in-house. The experimental analysis showed that our method outperforms other state-of-art approaches on all the three datasets. Starting from an overcomplete description of face images made of Local Binary Patterns, we apply a step of feature selection to select the discriminative information for the problem, and obtain a lower dimensional description. In a future work, we will integrate the proposed method as a module of the recognition system presented in Chapter 4, aiming to further reduce the computational time and improve its recognition performances.

Chapter 6

Text Spotting Via Inference on the Extremal Region Tree

In this chapter we introduce a novel method for blob classification for text detection. We present a Bayesian framework that takes into account the tree structures of blobs (extremal regions) obtained by local binarization to filter out text candidate elements. After giving an overview about the problem of text detection in Section 6.1, in Section 6.2 we introduce the state of the art on the problem. In Section 6.3 we describe in details the method we propose while in Section 6.5 we describe the experiments and obtained results.

6.1 Introduction

Text detection or localization is the process of determining the presence and the location of text.

The past few years have seen increasing interest among the research community in algorithms capable to quickly detect image areas containing text. By concentrating OCR resources on such "promising" areas, good computational efficiencies can be achieved. This is particularly important for applications involving resource-constrained mobile platforms (including Augmented Reality, robotics, and assistive technology).

The problem of text detection in uncontrolled scenes is still an open problem. The localization of text, in fact, shares difficulties common to all the problems of object recognition, together with other peculiar characteristics, such as the appearance variability of the characters due to the use of different fonts, orientation and styles. Other conditions that increase the difficulty of the task are represented by blurring, low contrast in the image, textured background, clutter, different lighting conditions, shadows, noise (Figure 6.1).

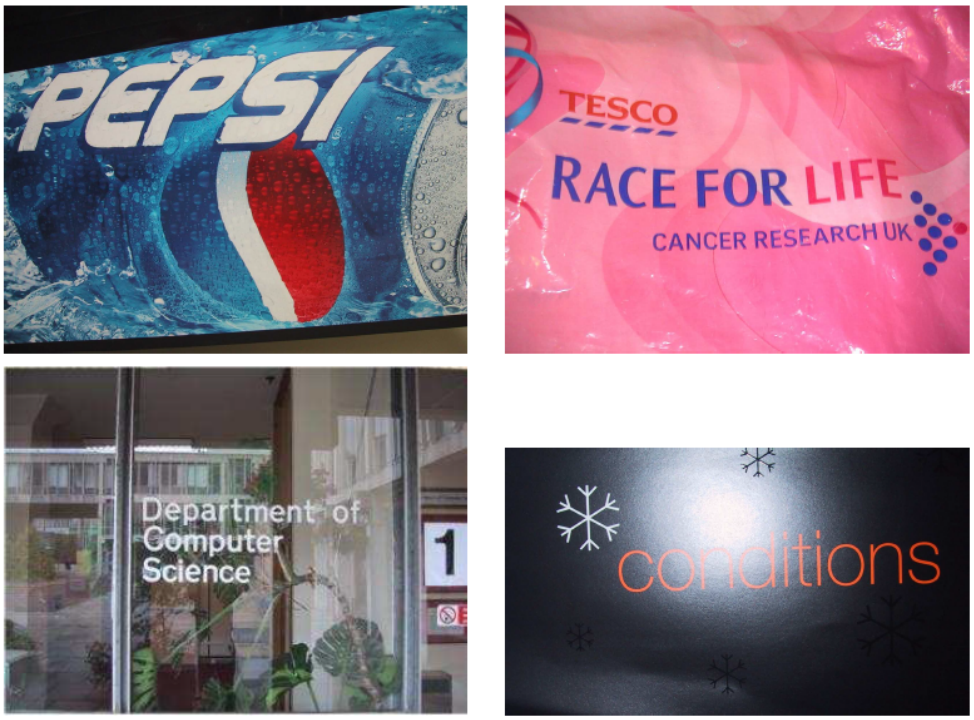


Figure 6.1: Example of text appearance variability. Different scales and rotation together with highlights and reflections represent challenges for the text detection (images from ICDAR datasets).

An exhaustive approach to text detection is computationally expensive. Given an image of N pixels, an exhaustive search would have to analyze the 2^N subsets of pixels.
Text detection methods deal with the complexity of the problem by reducing the number of subsets adopting two different approaches. One approach is to adopt sliding windows in order to reduce the search to a subset of image rectangles ([LLL$^+$11, CY04]); another approach, instead, consists of searching individual characters by grouping pixels into regions, under the assumption that pixels belonging to the same character share similar properties, e.g. color or intensity ([NM12, NM11a]).
Methods based on sliding windows reduce the number of tests to cN, with c being a constant that can assume very small values – in the case of methods with fixed scale and rotation – or large values when considering different parameters for the sliding window (scale, aspect ratio, rotation, etc.) [NM12]. Hence taking into account different appearance of the text increases the complexity of the problem.
Concerning region-based (or blob-based) methods, their advantage is that the complexity usually does not depend on the properties of the text, but on the other side their are sensitive to clutter and occlusions.

Figure 6.2: Example of Text Detection. Ideally each word should be included in a bounding box.

6.2 State of the Art

In this section we give an overview of the state of the art for text detection. The different methods are grouped according to the approach they use for the selection of candidate text regions in the images.

6.2.1 Region-based methods

Region-based methods usually adopt a bottom-up approach by grouping small components into successively larger components. Region properties between text and background, such as color, edge and connected component are often used to extract candidate text objects. Pattern recognition and/or geometrical analysis is usually used to merge the text components and filter out non-text components, also taking into account their spatial arrangement in order mark the text regions in the image. In [OSA94] presented a method based in which the scene image is first segmented into regions using an adaptive thresholding method, and successively each component is tested observing the gray-level difference between adjacent regions, in order to detect character candidate regions. This approach has several limitations about the alignment of the text (upright and not connected) and monochrome color. The method proposed in by [ZKJ95], text detection in color image is performed. The image is preprocessed applying a color reduction technique, that quantizes the color space using the peaks in the RGB histogram. The idea behind

this approach is based on the assumption that text regions cluster together in the new color space and that they occupy a significant portion of the image. Each component is successively filtered using spatial heuristics. In [Kim96], instead, the image is segmented using color clustering in the RGB space. After a filtering preprocessing, horizontal text lines and segments are extracted using iterating the analysis of the projection profile of each component. In the post-processing stage, the text segments are then merged using some heuristics. In [cSDB98] pixel that have similar gray levels are merged together to form a component. The candidate regions are then verified using size, area, fill factor and contrast. Cluster-based templates, instead, are used in [KJ00] in order to filter out non-character components. The template-based approach ease the difficulty of defining heuristics for the detection of text.
[HSYS01], under the assumption that every character is depicted using a single color, they perform an image segmentation selecting different representative colors from the color histogram. The result of this segmentation approach is a set of binary images, from which strings fo text are extracted using a multi-stage relaxation [HSY$^+$97]. The filtering phase is applied after merging all the binary images. The likelihood of a string of characters is defined according to the alignment of the single characters in the string and other geometric properties of the components. The filtering technique proposed in this work make it possible to detect shadowed and curved strings. [KS08] propose a static text detection algorithm to extract text from videos. The method is based on the spatio-temporally consistent properties of the static text, mainly focusing on the color. Some simplifying assumptions are made: the text color is preserved in consecutive frames, the text is characterized by uniform color and the boundaries of the text can be preserved because of the high contrast against the background. The problem of text localization is solved in [SNDC07] by detecting character strokes. The character segmentation is obtained thresholding the stroke-width.Related to [SNDC07], [EOW10] introduce the stroke width transform (SWT) and an efficient algorithm for its computation, producing a new image feature. Unlike [SNDC07], the proposed SWT is densely estimated in every pixel and the scope is non-local. [Liu08] propose a framework that uses the Niblack algorithm [Nib86] to threshold images and groups components into regions using geometry features. Using an intensity histogram based filter and an inner distance based shape filter text candidates are selected, while filtering out false positives (see Figure 6.3. In [BSN$^+$08], the candidate text blocks are extracted using a multi-scale Harris-corner based method. The position similarity and color similarity of Harris corners are used to generate boundaries of text objects. [ZZ10a] use histogram of oriented gradients [DT05] to capture the similarity of stroke edges. The range of gradient orientation is divided into 4 bin and the computed histograms are compared to detect candidate text regions. Graph spectrum is used to capture relationship among candidate blocks and cluster them into groups. In [SPT11] a K-means clustering in the Fourier-Laplacian space is used to identify candidate text regions. The skeleton of each connected component is then used to separate the different text strings from each other. In the final stage, the straightness of the text strings and edge density are used to filter out the false positive. [Kim05] propose to extract text using intensity information from natural scene images. The proposed method combines two image segmentation algorithms. The first algorithm is composed of gray value stretching and binarization using the

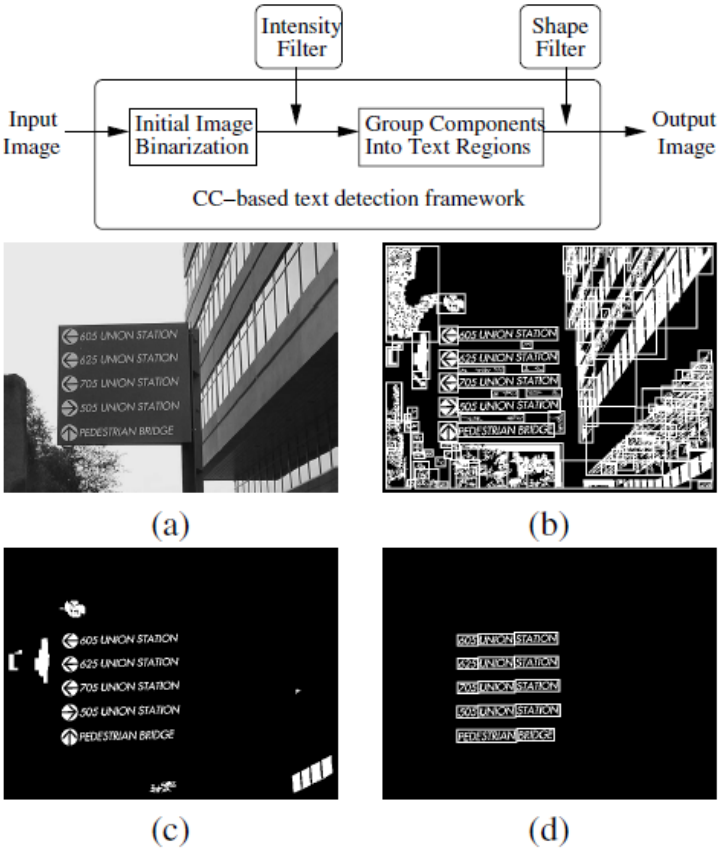


Figure 6.3: The algorithmic pipeline (top) and the processing output (a). Output of (b) the framework only, (c) the framework combined with the intensity filter, and (d) the framework using both filters. [Liu08]

average intensity of the image, while the second method is instead a Split and Merge approach. More recently, [NM11b, NM11a] proposed to detect characters as Maximally Stable Extremal Regions (MSERs) [MCUP04] and perform text recognition using the segmentation obtained by the MSER detector. [CTS+11] propose edge-enhanced MSER, combining Canny edges with MSER. Further, to obtain more robust result, they efficiently generate a stroke width transform image of the detected regions using the distance transform. The filtering is then performed using the geometric and stroke width information. In [NM12] the character detection problem is posed as an efficient sequential selection from the set of Extremal Regions (ERs). In a first stage of classification, the probability of each ER being a character is computed with $O(1)$ complexity and the candidate regions are selected selecting the ERs with locally maximal probability. A second stage, using more computationally expensive features, is responsible of the final filtering of the false positives.

(a)

Figure 6.4: An example of different output obtained by [BSN+08] (a) Detected Harris corner. (b) Position and color similarity. (c) Position similarity

A different category of region-based methods is represented by the approaches that achieve the detection of candidate text area using edge features. Edge-based methods are based on the assumption that there is high contrast between the text and the background and are mainly used to extract text from videos. The edges of the text boundary are detected and using some heuristic the non-text regions are filtered out.
[HK00] present a morphological approach to extract text from images, combining the RGB components of an input image in order to obtain an intensity image. After the color conversion, edges are extracted using a morphological gradient operator. A binary edge image is obtained thresholding the edge detection output, and an adaptive thresholding is performed for each candidate region in the intensity image. The method presented in [LS06] use multiscale edge detector to detect the text regions. They compute the edge strength, density, and orientation variance to form the multiscale edge detector. [SHT08] introduced the use of edge straightness for the elimination of edges that do not belong to text in video frames. Text block are identified combining heuristic rules and edge analysis. The method proposed in [AGP10] realizes a multi-resolution two-stage system for text detection in video images. Text lines are coarsely detected from the edge map. In the second stage, the detection result is refined using an SVM classifiers trained on features extracted using a novel LBP descriptor, edge LBP (eLBP), that capture the local edge distribution.

6.2.2 Sliding Window-Based Methods

Window-based approaches are built on the observation that text regions in images are characterized by distinctive properties that distinguish them from the background. The use of machine learning methods in this class of approaches is common, usually adopting ad-hoc features to verify the detected text objects.

In [CY04] the authors use AdaBoost to learn a strong classifier for detecting text in unconstrained city scenes, adopting features characterized by low entropy of the positive training examples in order to obtain similar responses of the classifier for any text input. A multilayer feed-forward

network is adopted in [LW02] to detect text in images and video frames. In the case of video, the system exploits the temporal redundancy to refine the candidate text lines. [TCYZ05] compute the average intensity and statistics of the number of edges from training samples, training an AdaBoost classifier to detect text candidate blocks and text boundaries are then matched using deformable templates. In [HP09] instead the authors utilize Mean Difference feature, Standard Deviation, HOG feature, to train a Complexity AdaBoost classifier to extract text in scene images. The approach in [WHL09] extracts gray scale contrast and edge orientation histogram features; then an SVM classifier is used to verify detected text objects. [LLL+11] adopts Modest AdaBoost on Classification and Regression Tree trained on a large number of features.

6.3 The Proposed Method

In this section we present a blob-based approach to text spotting, and particularly with the use of extremal regions. Extremal regions (ER [MCUP04]) are connected regions of binarized pixels, that is, pixels all with brightness below (or above) a certain threshold. As shown in [MCUP04], ERs for properly chosen thresholds can indeed capture the contour of text characters even for images taken "in the wild", possibly in difficult lighting conditions. Unfortunately, choosing a correct binarization threshold in these situations is difficult. Unlike document images taken under controlled light, images in the wild do not normally have a distinct, bimodal histogram that can be used for threshold selection. For example, in Fig. 6.5 (a), we show a piece of text under non-uniform illumination. Although there exists a threshold that would correctly binarize the word, finding this threshold is quite difficult, and standard algorithms (e.g. Otsu's algorithm, see Fig. 6.5 (b)) would fail in this situation.

We introduce a new strategy for blob detection and classification that exploits the simplicity and power of extremal region analysis without the need for early binarization. The idea stems from the realization that extremal regions form a tree structure, where each level in the tree corresponds to a binarization threshold (see Fig. 6.5 (c)). This tree structure imposes a strong prior on the labels that can be assigned to extremal regions. Specifically, labeling an extremal region for a certain threshold as 'text' implicitly assumes that all extremal regions in the sub-tree also inherit the 'text' label. Rather than committing to a certain threshold, we hypothesize a labeling for all extremal regions (all nodes in the tree) by maximizing a function that takes into account both the appearance of the extremal regions, and the tree-induced label inheritance constraint. We set this problem in a Bayesian framework and derive a simple solution based on dynamic programming.

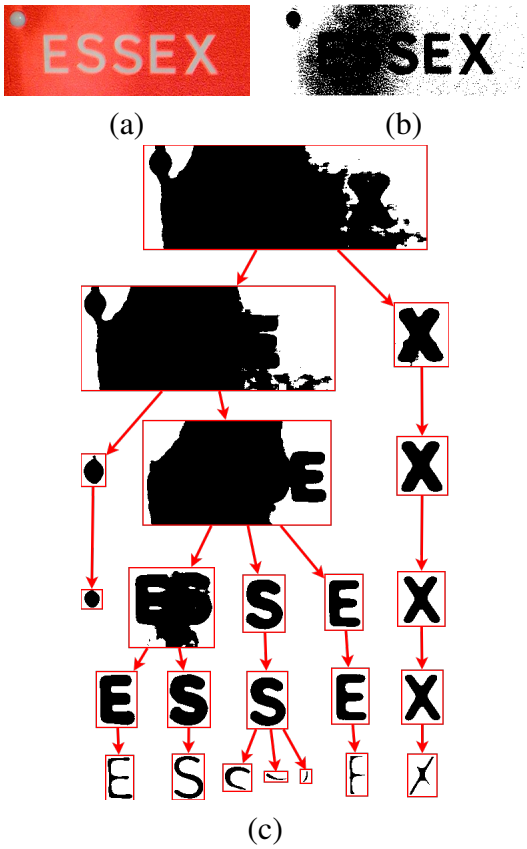


Figure 6.5: (a) A piece of text under non-uniform illumination. (b) Binarization using Otsu's algorithm. (c) The extremal region tree for this image.

6.4 ER Classification on Trees

A *sublevel set* is the set of points in the image whose brightness values I are less than or equal to a certain threshold value T. An *extremal region*[1] (ER) is a connected component of points in a sublevel set [MCUP02]. The shape and number of ERs depend on the threshold T. As T grows, extremal regions may expand, two or more extremal region may join to form a super-extremal region, and new extremal regions may be generated (see Fig. 6.5 (c)). Thus, the threshold T induces a tree structure on the extremal regions it defines. In an extremal region tree (ERT), each node is associated with an ER for some threshold T, and the ER associated with a node contains the ERs associated with all the node's descendants. The ERT's root is the ER for $T = 255$ (i.e., the whole image). The ERT's leaves comprise ERs at $T = 0$ as well as ERs that have been generated at a certain value of T (i.e., all of their points are larger then or equal to T). Nodes

[1]Note that the definition of extremal regions and of extremal region trees given here is quite different from the original definition in [LP90].

with multiple children represents situations in which multiple ERs have been merged into one ER.

Thus, an ERT represents a structured collection of all ERs in an image. Intuitively, we expect that some of these ERs may be sufficient to identify semantic-rich "blobs". However, selecting the "optimal" ERs has proven difficult, as discussed earlier. Neumann and Matas [NM12] recently proposed a different approach: extract geometric features from *all* extremal regions, assign a score to each region based on a trained classifier, and then select regions based on a "global" criterion. More precisely, their algorithm selects ERs whose score represents a local maximum in the ERT path from a leaf to the root. This criterion, which has produced promising results [NM12], is intuitively appealing; however, it lacks a solid theoretical justification. In this paper we propose a new Bayesian approach for ER selection that is shown to produce substantially improved results. We also introduce a much simpler, greedy approach (dubbed "Winner Take All") which, though suboptimal, is shown to also produce good results.

6.4.1 Notation

In the following, we enumerate the N nodes of the extremal region tree using a single index variable i. The N_i children of node i are indicated by the set

$$\mathbf{c}(i) = \{c_1(i), c_2(i), \ldots, c_{N_i}(i)\}$$

The set of all descendants of node i is denoted by $\mathbf{d}(i)$. The appearance of an ER is captured by the feature vector I^i; the set of all feature vectors (one per ER) is denoted by $\mathbf{I}$.

Each extremal region can be labeled as 'text' or 'not text'. A given image can be hand-labeled, for the purpose of algorithm training and verification, by drawing regions containing text. At runtime, the system needs to be able to identify regions containing text, and success is declared when these regions are consistent with hand-labeling according to some suitable criterion. Note however that no one-to-one correspondence between hand-drawn regions and ER may be found in general. Thus, a criterion is needed to induce a label on each ER based on hand-drawn regions. The criterion we adopted defines an ER as 'text' when at least the 90% of its pixels are within a hand-labeled region.

We will assume that ER labeling satisfies the following property:

> If a node i is labeled as 'text', all of its children should also be labeled as 'text'.

This requirement derives from the fact that children of an ER are always contained within the ER. This also implies that labeling a node i as 'text' induces a similar labeling on all the nodes in the subtree with root in i. A subtree whose root is labeled as 'text' but whose root's parent is labeled as 'not text' will be called a *maximal text subtree*.

We will denote the label assigned to the i-th node (ER) by the binary random variable v^i. The event v^i=1, meaning "the i-th node is 'text'", will be sometime expressed by the shorthand 1^i. The random vector $\mathbf{v}^{\mathbf{c}(i)}$ represents the labels of the children of node i:

$$\mathbf{v}^{\mathbf{c}(i)} = \{v^{c_1(i)}, v^{c_2(i)}, \ldots, v^{c_{N_i}(i)}\}$$

The set of all variables $\mathbf{v}^i$ is denoted by $\mathbf{v}$.

6.4.2 Algorithm 1: Inference on ERT

The goal of this algorithm can be stated as follows:

> *Find the label assignment $\bar{\mathbf{v}}$ that maximizes the posterior probability of assignment given all the observables:*
>
> $$\bar{\mathbf{v}} = \arg\max_{\mathbf{v}} P(\mathbf{v}|\mathbf{I})$$

Once a global labeling $\bar{\mathbf{v}}$ has been assigned, the algorithm returns the maximal text subtrees' roots.

In order to make the problem tractable, we use Bayes' theorem to transform $P(\mathbf{v}|\mathbf{I})$ into a quantity that is amenable to factorization:

$$P(\mathbf{v}|\mathbf{I}) \propto P(\mathbf{I}|\mathbf{v})P(\mathbf{v}) \tag{6.1}$$

We will assume that the appearance of the ERs in the tree are conditional independent given the labeling:

$$P(\mathbf{I}|\mathbf{v}) = \prod_{i=1}^{N} P(I^i|v^i) \tag{6.2}$$

This is a common simplifying hypothesis which, however, is unlikely to hold true for parent - single child node pairs (since the parent is likely to be very similar to its single child). Partly for this reason, we prune the ERT as discussed more in detail in Sec. 4.6.

We assume that a classifier is trained to produce a likelihood value $P(I^i|v^i)$ of observing the feature I^i in an ER labeled as v^i. Given the tree structure of the variables, we can write:

$$P(\mathbf{v}) = \prod_{i=1}^{N} P(v^i|\mathbf{v}^{\mathbf{d}(i)}) \tag{6.3}$$

In this expression, we assume that $P(v^i|\mathbf{v}^{\mathbf{d}(i)})$ for the leaves (which have no descendants) is set equal to a prior probability value $P(v^i)$ that is kept constant for all leaves.

v^i	$\mathbf{v}^{\mathbf{c}(i)}$	$P(v^i\|\mathbf{v}^{\mathbf{c}(i)})$
0	$\mathbf{1}$	$1-\alpha$
1	$\mathbf{1}$	α
other		0

Table 6.1: The simple expression of $P(v^i|\mathbf{v}^{\mathbf{c}(i)})$.

As customary with similar tree structures, we assume that

$$P(v^i|\mathbf{v}^{\mathbf{d}(i)}) = P(v^i|\mathbf{v}^{\mathbf{c}(i)}) \tag{6.4}$$

and thus our problem can be restated as follows:

$$\bar{\mathbf{v}} = \arg\max_{\mathbf{v}} \prod_{i=1}^{N} P(I^i|v^i)P(v^i|\mathbf{v}^{\mathbf{c}(i)}) \tag{6.5}$$

Term $P(v^i|\mathbf{v}^{\mathbf{c}(i)})$ in (6.5) deserves special attention. It is clear that, given the labeling constrains stated in Sec. 6.4.1, a child node cannot be 'not text' when is its parent is 'text', and thus $P(v^i = 1|\mathbf{v}^{\mathbf{c}(i)})$=0 as soon as any $v^{c_j(i)}$ is equal to 0. Otherwise stated:

$$P(\mathbf{1}^{\mathbf{c}(i)}|1^i) = 1 \tag{6.6}$$

where $\mathbf{1}^{\mathbf{c}(i)}$ indicates that all $v^{c_j(i)}$ are equal to 1. Thus, the only term that remains to be specified to fully define $P(v^i|\mathbf{v}^{\mathbf{c}(i)})$ is $P(1^i|\mathbf{1}^{\mathbf{c}(i)})$. We will assume that this quantity is constant throughout the tree; we will denote this constant by α. Tab. 6.1 highlights the resulting simple expression of $P(v^i|\mathbf{v}^{\mathbf{c}(i)})$.

Intuitively, our Bayesian inference algorithm uses the "context" to reduce the risk of isolated classification errors. For example, an internal node that, based on the appearance of its associated ER, could, by itself, be misclassified as 'text', will probably end up being correctly classified as 'non text' if many of its descendants do not look like 'text'. At the same time, though, the presence of a single node that, by itself, would be misclassified as 'non text' with high confidence (large $P(I^i|0^i)$), may result in all its ancestors being classified as 'non text'. This represents the main weakness of the proposed algorithm.

6.4.2.1 Linear Maximization Algorithm

We now show how to compute the maximization in Euation (6.5). We first rewrite it as:

$$\max_{v^1} P(I^1|v^1) \cdot \max_{\mathbf{v}^{\mathbf{d}(1)}} P(v^1|\mathbf{v}^{\mathbf{c}(1)}) \prod_{j \in \mathbf{d}(1)} P(I^j|v^j) P(v^j|\mathbf{v}^{\mathbf{c}(j)}) \tag{6.7}$$

$$= \max_{v^1} P(I^1|v^1) \cdot L_1(v^1)$$

where v^1 is the root of the tree and we defined

$$L_i(v^i) = \max_{\mathbf{v}^{\mathbf{d}(i)}} P(v^i|\mathbf{v}^{\mathbf{c}(i)}) \prod_{j \in \mathbf{d}(i)} P(I^j|v^j) P(v^j|\mathbf{v}^{\mathbf{c}(j)}) \tag{6.8}$$

$$= \max_{\mathbf{v}^{\mathbf{c}(i)}} P(v^i|\mathbf{v}^{\mathbf{c}(i)}) \max_{\mathbf{v}^{\mathbf{d}(i)/\mathbf{c}(i)}} \prod_{j \in \mathbf{d}(i)} P(I^j|v^j) P(v^j|\mathbf{v}^{\mathbf{c}(j)})$$

The following recursion holds:

$$L_i(v^i) = \max_{\mathbf{v}^{\mathbf{c}(i)}} P(v^i|\mathbf{v}^{\mathbf{c}(i)}) \prod_{j \in \mathbf{c}(i)} P(I^j|v^j) L_j(v^j) \tag{6.9}$$

which makes it possible to compute the maximum-a-posterior label assignment starting from the tree's leaves.

At first sight, it would seem that maximizing (6.9) requires consideration of all 2^{N_i} vectors of binary combinations (where N_i is the number of children of if the i-th node). In fact, thanks to the particular form of $P(v^i|\mathbf{v}^{\mathbf{c}(i)})$ (see Tab. 6.1), only a number of tests proportional to $N(i)$ is necessary. Indeed:

$$L_i(1) = \alpha \prod_{j \in \mathbf{c}(i)} P(I^j|1^j) L_j(1^j) \tag{6.10}$$

$$L_i(0) = \max \left((1-\alpha) \prod_{j \in \mathbf{c}(i)} P(I^j|1^j) L_j(1^j), \right. \tag{6.11}$$

$$\left. \max_{\mathbf{v}^{\mathbf{c}(i)} \neq \mathbf{1}^{\mathbf{c}(i)}} \prod_{j \in \mathbf{c}(i)} P(I^j|v^j) L_j(v^j) \right)$$

Let $\bar{v}^j = \arg\max_{v^j} P(I^j|v^j)L_j(v^j)$ and $\bar{j} = \arg\max_{j \in \mathbf{c}(i)} P(I^j|0^j)L_j(0^j)$. Then

$$\max_{\mathbf{v}^{\mathbf{c}(i)} \neq \mathbf{1}^{\mathbf{c}(i)}} \prod_{j \in \mathbf{c}(i)} P(I^j|v^j)L_j(v^j) =$$

$$\begin{cases} \prod_{j \in \mathbf{c}(i)} P(I^j|\bar{v}^j)L_j(\bar{v}^j) \text{ if not all } \bar{v}^j = 1 \\ \prod_{j \in \mathbf{c}(i), j \neq \bar{j}} P(I^j|1^j)L_j(1^j) \cdot P(I^{\bar{j}}|0^{\bar{j}})L_{\bar{j}}(0^{\bar{j}}) \text{ otherwise} \end{cases}$$

Note that computation of $L_i(0)$ and $L_i(1)$ has linear cost in N_i.

6.4.3 Algorithm 2: Winner-Take-All (WTA)

This algorithm is very simple. Each ER is tested using the classifier, tuned to a constant threshold θ. (In other words, the i-th ER is classified as 'text' if $P(I^i|1^i) > \theta$.) Then, for each ER classified as 'text', all of its descendants are automatically labeled as 'text'. Intuitively, this algorithm assumes that classification at higher level of the tree is more reliable, and uses the constraint $P(\mathbf{1}^{\mathbf{c}}(i)|1^i) = 1$ to condition all of the descendants of an ER classified as 'text'. This algorithm is very efficient to implement, as it simply amounts to exploring the tree depth-first starting from the root, classifying each encountered node, and neglecting the whole subtree of any node that was classified as 'text'.

6.4.4 Algorithm 3: Neumann-Matas (NM)

In this algorithm, the probability $P(I^i|1^i)$ is tracked for the path from each leaf to the root. A node i is marked as 'text' if $P(I^i|1^i)$ is a local maximum in the path, and this local maximum exceeds a threshold and the difference between local maximum and local minimum also exceeds a threshold [NM12].

6.4.5 Features

Here we describe the features we adopted to describe the appearance of the ERs detected in the images. To better exploit the ERT, we need scale invariant features that could describe ERs that encompass more than one character. Hence starting from the set of features proposed by [NM12], designed for the detection of characters, we extend it adopting the features proposed in [SSC11] and other shape descriptors.

The features we adopted from [NM12]:

- aspect ratio - the ratio between the width and the height of the bounding box;
- region compactness - $\frac{\sqrt{a}}{p}$ where a is the area of the region and p is the perimeter;
- number of holes - $1 - \eta$, with η Euler number of the region;
- horizontal crossings c_i - the number of transitions between pixels belonging ($p \in r$) and not belonging ($p \notin r$) to the region in the row $i - th$ of the region r;
- convex hull ratio;
- hole area ratio;
- number of outer boundary inflexion points.

[SSC11] proposes a set of moment based features for the detection of blurry text in low resolution images taken in the wild, that are suitable for the recognition of group of characters. Following Sanketi et al., said (x_c, y_c) the geometric center of a blob and BA its area, we compute the following set of features:

- X Moment = $\sum_{(i,j) \in blob} (x_c - x_{ij})$
- Y Moment = $\sum_{(i,j) \in blob} (y_c - y_{ij})$
- X Moment Second (XMS) = $\sum_i (\sum_j (y_c - y_{ij})^2) dA_i$, where $dA_i =$ sum of the ith row
- Y Moment Second (YMS) = $\sum_j (\sum_i (x_c - x_{ij})^2) dA_j$, where $dA_j =$ sum of the jth column
- X Moment Radius Gyration = $\frac{XMS}{BA}$
- Y Moment Radius Gyration = $\frac{YMS}{BA}$
- X Moment Two (XMT) = $\sum_i \sum_j (y_c - y_{ij})^2$
- Y Moment Two (YMT) = $\sum_j \sum_i (x_c - x_{ij})^2$
- Perimeter

We also include the following features for shape description:

- Contour Roughness [ZQJX07];
- Edge Symmetry [ZZ10a];
- Correlation [ZDO09];

- Mean and variance of blob width: for each row of the ER bitmap, we compute the longest chain of pixels equal to 1, then we compute mean and variances of those quantities;
- Number of bays: number of connected components generated by the difference between the ER convex hull and the ER bitmap.

6.5 Experiments and Results

In this section we present the features adopted for the description of the blobs and we discuss the algorithmic choices we made. We will illustrate the obtained results, comparing our classification method with the WTA and NM algorithms introduced in Section 6.4.3 and 6.4.4 respectively.

6.5.1 Computing the Likelihood Terms

The likelihood terms $P(I^i|v^i)$ are computed from the output of an SVM classifier with Radial Basis Function (RBF) kernel trained on the features described in Section6.4.5. A classifier generally produces a number that, after suitable normalization, can be treated as a class-posterior probability $P(v^i|I^i)$ [NMC12]. In our experiments we apply Platt Scaling and Isotonic Regression [Pla99] to map SVM outputs on $[-\inf; +\inf]$ to posterior probabilities on $[0; 1]$. We then convert the obtained posterior to a quantity that is proportional to a class-conditional likelihood by dividing $P(v^i|I^i)$ by the prior term $P(v^i)$, that is left as an external parameter of the algorithm.

6.5.2 Experiments on the ICDAR dataset

We tested our ER classification algorithm on the ICDAR 2011 Robust Reading dataset, which contains 230 training images and 255 test images. Training was performed on 8560 'text' and 10,360 'non-text' ERs (where the 'text' and 'non text' labels were inferred from the ground truth labels from the dataset) (Figure 6.8).

Note that a 'text' ER may actually contain a character, a group of characters, or only part of a character (Figure 6.6). All ERs in the test images were classified as 'text' or 'not text'. We used the following metric to assess the quality of classification. For a given choice of parameter, we computed, for each image n, the union of all ERs classified as 'text' (S_c^n) as well as the union of all ERs labeled as 'text' (S_l^n). We defined an overall measure of Precision and Recall as follows:

$$\text{Precision} = \frac{\sum_n |S_l^n \cap S_c^n|}{\sum_n |S_c^n|} \text{ , Recall} = \frac{\sum_n |S_l^n \cap S_c^n|}{\sum_n |S_l^n|}$$

Figure 6.6: Example of "text" ERs in the training set

where $|S|$ is the the of the set of points S. In order to account for both dark text on light background and light text on dark background, we run the algorithm twice, with opposite polarity of the ERs, and take the union of the the ERs with both polarities that were classified as 'text' as our result.

The complexity of Bayesian inference is proportional to the number of nodes in the tree. Tree reduction can be considered to reduce the cost of the algorithm. As hinted in Sec. 6.4.2, a well-thought tree reduction strategy can also lead to better compliance with the conditional independence assumption (6.2). We have considered three different approaches to tree reduction. The first approach is through binning of the quantization levels. For example, rather than considering all integer thresholds T between 0 and 255, one could consider only multiples of some base value $m_q > 1$. The second strategy is to impose a minimum "lifetime" to an ER, where by lifetime we define the maximal length of the single-child chain associated with this evolving ER. If a single-child chain has length shorter than a certain threshold m_s, we simply attach the children of the last node in the chain as children to the parent of the first node. The third considered strategy is to remove a single-child node when the incremental ratio of the areas of the ERs associated to the node's parent and to the node is smaller than a threshold m_a. Intuitively, this node is redundant since it is probably very similar (in shape) to its parent.

We experimented with various values for the thresholds m_q, m_s and m_a and found that the choice $m_q = 2$, $m_s = 3$ and $m_a = 0.1$ gave good results while reducing the average number N of ERs per tree from 3360 to 246 (Table 6.2). Fig. 6.7 shows the ROC curves for the three algorithms described in Sec. 6.4. The ROC curve for the WTA and the N-M algorithms are obtained by varying the associated threshold. For Algorithm 1 (inference on ERT), the ROC curve was obtained modifying the value of the prior $P(1)$. The outcome of our experiments shows that our Bayesian inference algorithm performs better than greedy methods, such as WTA and N-M algorithms.

The conditional prior probability of a node being "text" given that all of its children are labeled

Pruning method	avg. N
no pruning	3360
$m_a = .1$	1680
$m_q = 2$	1670
$m_q = 2, m_a = 0.1, m_s = 3$	246

Table 6.2: Average Number of nodes in the ERTs after performing different combinations of pruning.

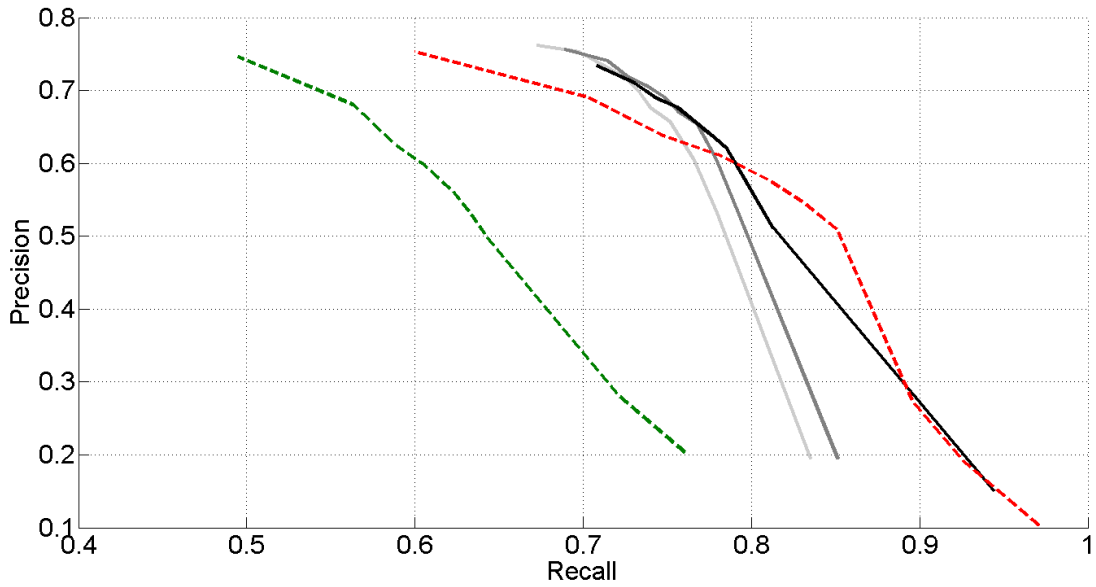


Figure 6.7: Precision-recall curves for three different ER classification algorithms. Dashed green line: Algorithm 3 (N-M); Dashed red line: Algorithm 2 (WTA); Solid line: Algorithm 1 (inference on ERT). Black: α=1; medium gray: α=0.9; light gray: α=0.8.

a "text" can be learned from a set of training ERTs.

The measurement of the value of $P(\mathbf{1}^{\mathbf{c}(i)}|1^i)$ w.r.t. the number of children of the nodes in the tree depicted in Figure 6.9) suggests the option of adopting of an adaptive strategy for the conditional prior, varying its value according to the number of children of the nodes. In Figure 6.10 we

Figure 6.8: Left Column: Example of images from ICDAR dataset. Right Column: output of the proposed Bayesian algorithm.

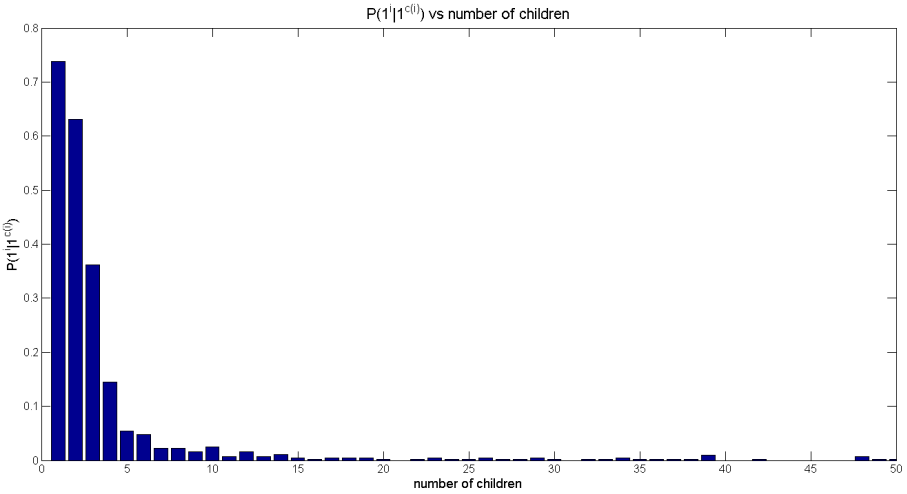


Figure 6.9: The value of $P(\mathbf{1}^{\mathbf{c}(i)}|1^i)$ measured w.r.t. the number of children of the nodes of the ERT.

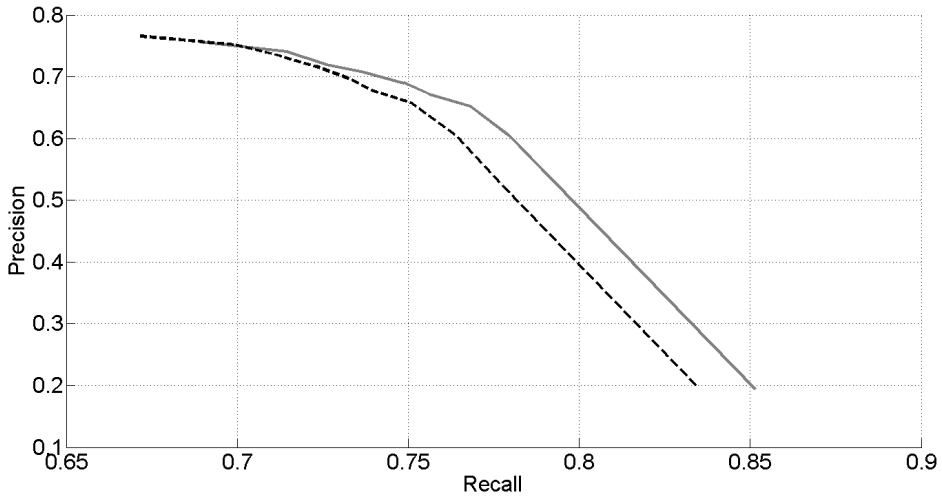


Figure 6.10: Precision-recall curves for Algorithm 1. Solid line: α=0.9; dashed: adaptive α.

report a comparison between Algorithm 1 with $\alpha = 0.9$ and the same algorithm with an adaptive α. The results of the experiment show a general adaptive approach seems to perform worse than a fixed value approach.

6.6 Discussion

We have introduced an algorithm to jointly binarize an image and classify the binarized blobs (extremal regions). This algorithm constitutes the first step of a complete text spotting pipeline, whereby the blobs classified as 'text' by our algorithms are further concatenated into words, a process that enables removal of misclassified blobs. Unlike the traditional "early commitment" approach, where an image is first binarized and then the resulting blobs are classified, our algorithm casts the problem of binarization and classification jointly in a Bayesian framework. The tree structure of extremal regions and the presence of a strong inheritance constraint on this tree lead to efficient maximization, further improved through a principled strategy for tree reduction. The experimental results show that our technique performs substantially better than a previously proposed approach that also considered the extremal regions tree.

Chapter 7

Conclusions

In this thesis we addressed two challenging research problems for the computer vision community and of interest for visually impaired individuals: face recognition and text detection. The methods we presented in this thesis are all characterized by a contained computational cost, since we had in mind prospective real-time applications to be used on mobile devices by visually impaired users.

We proposed a real-time face recognition system that, combining the face retrieval and face classification paradigms, mitigates the drawbacks of the two approaches when considered separately. The system we developed exploits the temporal coherence of video data combining retrieval and classification in a feedback loop to refine a recognition score and thus increase the robustness of the system feedback. The experimental analysis we conducted on two different datasets showed that our method competes with state-of-art approaches and is able to adapt to settings of variable properties. We also demonstrated that adding the temporal component can effectively improve the recognition performances.

We introduced a novel method for the simultaneous selection of structured features in multiclass problems. Starting off from Group-LASSO, a well-known method for structured feature selection and founded on a well grounded theory, we reformulated the functional so that it could take into account multiclass problems, the *Multi-Class Group LASSO*. Simultaneous selection was obtained by adapting the proximal methods framework. The effectiveness of the proposed method has been assessed in the context of face recognition on three different datasets characterized by various degrees of variability. Starting from an overcomplete description of face images using LBP descriptors, we applied our feature selection method to select the most significant subset of features common to all the identities. We showed that our method selects very sparse representations outperforming other state-of-the-art methods.

The results we obtained are very promising and in a future work we plan to integrate the Multi-Class Group LASSO method in our face recognition system. Exploiting the simultaneous se-

lection of features we will reduce the dimensionality of the face descriptions, increasing the efficiency of the system by avoiding the computation of redundant features, while possibly improving recognition rate.

Another direction for future work is the application of Multi-Class Group LASSO to the EVLBP descriptors, an extension of LBP to explicitly include the temporal component present in videos and described in Chapter 3.

The realization of a mobile platform implementation and the study of an effective user interface for blind users could represent an extension of our real-time face recognition system.

Concerning the problem of text detection, we introduced an algorithm to binarize an image and classify the extremal regions extracted from it to filter out non text elements. We casted the problem in a Bayesian framework, transforming the problem of binarization and classification of an image into a likelihood maximization of the labeling of the tree of the extremal regions. We exploited the tree structure and imposed some inheritance constraints that allowed us to efficiently perform said maximization. We also improved both the efficiency and the detection performances by the introduction of strategies for the pruning of the tree. We showed that our method performs better than two other greedy methods.

The algorithm we presented constitutes the first step of a text detection pipeline. A direction for future work is the development of an end-to-end system in which the blobs classified as 'text' by our algorithm are further concatenated into words, possibly exploiting the topology of the tree to implement a robust concatenation strategy.

This Ph.D. started from the project "Study and development of low cost video-based systems to support visually impaired user" [1] funded by Fondazione Carige [2] and has successively been sponsored by Regione Liguria [3].

[1] http://slipguru.disi.unige.it/index.php?option=com_jresearch&task=show&view=project&id=6&Itemid=12

[2] http://www.fondazionecarige.it/

[3] http://www.regione.liguria.it/

Bibliography

[Agg04] Gaurav Aggarwal. A system identification approach for video-based face recognition. In *Proceedings of International Conference on Pattern Recognition*, pages 175–178, 2004.

[AGP10] Marios Anthimopoulos, Basilis Gatos, and Ioannis Pratikakis. A two-stage scheme for text detection in video images. *Image Vision Comput.*, 28(9):1413–1426, September 2010.

[AHP04] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. *Computer Vision-ECCV 2004*, pages 469–481, 2004.

[AHP06] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12):2037 –2041, dec. 2006.

[ANM06] Ajith Abraham, Nadia Nedjah, and Luiza Mourelle. Evolutionary computation: from genetic algorithms to genetic programming. In Nadia Nedjah, Luiza Mourelle, and Ajith Abraham, editors, *Genetic Systems Programming*, volume 13 of *Studies in Computational Intelligence*, pages 1–20. Springer Berlin / Heidelberg, 2006. 10.1007/3-540-32498-4_1.

[ARACGM08] Enrique Argones Rúa, José Luis Alba Castro, and Carmen García Mateo. Biometrics and identity management. chapter Quality-Based Score Normalization and Frame Selection for Video-Based Person Authentication, pages 1–9. 2008.

[Bak99] S. Bakin. Adaptive regression and model selection in data mining problems. *thesis edition*, 1999.

[BFO+10] Luigi Balduzzi, Giovanni Fusco, Francesca Odone, S. Dini, Marco Mesiti, Augusto Destrero, and Alberto Lovato. Low-cost face biometry for visually impaired users. In *IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications*, 2010.

[BHK97] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):711–720, 1997.

[BK05] Hyeon Bae and Sungshin Kim. Real-time face detection and recognition using hybrid-information extracted from face space and facial features. *IMAVIS*, 23(13):1181 – 1191, 2005.

[BLGT06] Manuele Bicego, Andrea Lagorio, Enrico Grosso, and Massimo Tistarelli. On the use of sift features for face authentication. In *In: Conf. on Computer Vision and Pattern Recognition Workshop (CVPRW). (2006*, page 35. IEEE Computer Society, 2006.

[BM11] H. Bagherinia and R. Manduchi. Robust real-time detection of multi-color markers on a cell phone. *Journal of Real-Time Image Processing*, pages 1–17, 2011.

[BP93] R. Brunelli and T. Poggio. Face recognition: features versus templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(10):1042 –1052, oct 1993.

[BSN+08] Hongliang Bai, Jun Sun, Satoshi Naoi, Yutaka Katsuyama, Yoshinobu Hotta, and Katsuhito Fujimoto. Video caption duration extraction. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[CCC+11] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, Tao Wang, D.J. Wu, and A.Y. Ng. Text detection and character recognition in scene images with unsupervised feature learning. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 440 –445, sept. 2011.

[CKZ02] R. Chellappa, V. Kruger, and Shaohua Zhou. Probabilistic recognition of human faces from video. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–41 – I–44 vol.1, 2002.

[CLK+00] L.F. Chen, H.Y.M. Liao, M.T. Ko, J.C. Lin, and G.J. Yu. A new lda-based face recognition system which can solve the small sample size problem. *Pattern recognition*, 33(10):1713–1726, 2000.

[CM07] James Coughlan and Roberto M. Functional assessment of a camera phone-based wayfinding system operated by blind users. In *the Int. IEEE-BAIS Symposium on Research on Assistive Technology*, 2007.

[cSDB98] Jae chang Shim, C. Dorai, and R. Bolle. Automatic text extraction from video for content-based annotation and retrieval. In *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, volume 1, pages 618 –620 vol.1, aug 1998.

[CT07] E. Candes and T. Tao. The Dantzig selector: statistical estimation when $p$is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[CT10] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2567 –2573, june 2010.

[CTS+11] Huizhong Chen, Sam S Tsai, Georg Schroth, David M Chen, Radek Grzeszczuk, and Bernd Girod. Robust text detection in natural images with edge-enhanced maximally stable extremal regions. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 2609–2612. IEEE, 2011.

[CY04] Xiangrong Chen and A.L. Yuille. Detecting and reading text in natural scenes. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–366 – II–373 Vol.2, june-2 july 2004.

[CZY92] Y.Q. Cheng, Y.M. Zhuang, and J.Y. Yang. Optimal fisher discriminant analysis using the rank decomposition. *Pattern recognition*, 25(1):101–111, 1992.

[D+85] John G Daugman et al. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Optical Society of America, Journal, A: Optics and Image Science*, 2:1160–1169, 1985.

[DCC96] H. Demirel, T.J. Clarke, and P.Y.K. Cheung. Adaptive automatic facial feature segmentation. In *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pages 277 –282, oct 1996.

[DDDM04] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on pure and applied mathematics*, 57(11):1413–1457, 2004.

[DDMOV09] A. Destrero, C. De Mol, F. Odone, and A. Verri. A sparsity-enforcing method for learning face features. *Image Processing, IEEE Transactions on*, 18(1):188 –201, jan. 2009.

[DHG+10] Weihong Deng, Jiani Hu, Jun Guo, Weidong Cai, and Dagan Feng. Emulating biological strategies for uncontrolled face recognition. *Pattern Recognition*, 43(6):2210 – 2223, 2010.

[DMOV09] Augusto Destrero, Christine Mol, Francesca Odone, and Alessandro Verri. A regularized framework for feature selection in face detection and authentication. *International Journal of Computer Vision*, 83:164–177, 2009.

[DT05] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[EC97] Kamran Etemad and Rama Chellappa. Discriminant analysis for recognition of human face images. *Journal of Optical Society of America A*, 14:1724–1733, 1997.

[EOW10] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2963 –2970, june 2010.

[FHT01] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer Series in Statistics, 2001.

[FNO12] G. Fusco, N. Noceti, and F. Odone. Combining retrieval and classification for real-time face recognition. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pages 276 –281, sept. 2012.

[FS95] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Eur. Conf. on Comp. Learning Theory*, 1995.

[Fuk90] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Academic Pr, 1990.

[GCPC95] H.P. Graf, T. Chen, E. Petajan, and E. Cosatto. Locating faces and facial parts. In *Proc. First Int'l Workshop Automatic Face and Gesture Recognition*, pages 41–46, 1995.

[GKP09] Lakshmi Gade, Sreekar Krishna, and Sethuraman Panchanathan. Person localization using a wearable camera towards enhancing social interactions for individuals with visual impairment. In *Proceedings of the 1st ACM SIGMM international workshop on Media studies and implementations that help improving access to disabled users*, MSIADU '09, pages 53–62, New York, NY, USA, 2009. ACM.

[GMP00] S. Gong, S.J. McKenna, and A. Psarrou. *Dynamic Vision: From Images to Face Reco*. Image Processing. Imperial College Press, 2000.

[GS01] Ralph Gross and Jianbo Shi. The cmu motion of body (mobo) database. Technical report, 2001.

[HFLJ06] M. A. Hersh, René Farcy, Roger Leroux, and Alain Jucha. Electronic travel aids and electronic orientation aids for blind people: technical, rehabilitation and everyday life points of view. In *In M. A. Hersh (Ed.), CVHI 2006: Conference on Assistive Technology for Vision and Hearing Impairment*, 2006.

[HK00] Yassin MY Hasan and Lina J Karam. Morphological text extraction from images. *Image Processing, IEEE Transactions on*, 9(11):1978–1983, 2000.

[HL01] Erik Hjelmas and Boon Kee Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83(3):236 – 274, 2001.

[HLC10] Chao-Kuei Hsieh, Shang-Hong Lai, and Yung-Chang Chen. An optical flow-based approach to robust face recognition under expression variations. *Image Processing, IEEE Transactions on*, 19(1):233 –240, jan. 2010.

[HLW05] Xiangsheng Huang, S.Z. Li, and Yangsheng Wang. Jensen-shannon boosting learning for object recognition. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 144 – 149 vol. 2, june 2005.

[HMPL07] Abdenour Hadid, Matti Matti Pietikäinen, and Stan Z. Li. Learning personal specific facial dynamics for face recognition from videos. In *Proceedings of the 3rd international conference on Analysis and modeling of faces and gestures*, AMFG'07, pages 1–15, Berlin, Heidelberg, 2007. Springer-Verlag.

[HP04] A. Hadid and M. Pietikainen. Selecting models from videos for appearance-based face recognition. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 1, pages 304–308. IEEE, 2004.

[HP09] S.M. Hanif and L. Prevost. Text detection and localization in complex scene images using constrained adaboost algorithm. In *Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on*, pages 1 –5, july 2009.

[HSY+97] H. Hase, T. Shinokawa, M. Yoneda, M. Sakai, and H. Maruyama. Character string extraction by multi-stage relaxation. In *Document Analysis and Recognition, 1997., Proceedings of the Fourth International Conference on*, volume 1, pages 298 –302 vol.1, aug 1997.

[HSYS01] Hiroyuki Hase, Toshiyuki Shinokawa, Masaaki Yoneda, and Ching Y. Suen. Character string extraction from color documents. *Pattern Recognition*, 34(7):1349 – 1365, 2001.

[HY91] Z.Q. Hong and J.Y. Yang. Optimal discriminant plane for a small number of samples and design method of classifier on the plane. *pattern recognition*, 24(4):317–324, 1991.

[ICS09] V. Ivanchenko, J. Coughlan, and H. Shen. Staying in the crosswalk: A system for guiding visually impaired pedestrians at traffic intersections. 2009.

[ICS10] Volodymyr Ivanchenko, James Coughlan, and Huiying Shen. Real-time walk light detection with a mobile phone. In Klaus Miesenberger, Joachim Klaus, Wolfgang Zagler, and Arthur Karshmer, editors, *Computers Helping People with Special Needs*, volume 6180 of *Lecture Notes in Computer Science*, pages 229–234. Springer Berlin Heidelberg, 2010.

[JA09] Rabia Jafri and Hamid R Arabnia. A survey of face recognition techniques. *journal of Information Processing Systems*, 5, 2009.

[Jol05] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2005.

[JV03] M.J. Jones and P. Viola. Face recognition using boosted local features. 2003.

[JW03] Brijnesh J. Jain and Fritz Wysotzki. A competitive winner-takes-all architecture for classification and pattern recognition of structures. In *Int. Conf. on Graph based representations in patt. rec.*, pages 259–270, 2003.

[Kan73] Takeo Kanade. Picture processing system by computer complex and recognition of human faces. In *doctoral dissertation, Kyoto University*. November 1973.

[KHR10] K.M. Kramer, D.S. Hedin, and D.J. Rolkosky. Smartphone based face recognition tool for the blind. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 4538 –4541, 31 2010-sept. 4 2010.

[Kim96] Hae-Kwang Kim. Efficient automatic text location method and content-based indexing and structuring of video database. *Journal of Visual Communication and Image Representation*, 7(4):336 – 344, 1996.

[Kim05] Kwangbaek Kim. Intelligent immigration control system by using passport recognition and face verification. In *Proceedings of the Second international conference on Advances in neural networks - Volume Part II*, ISNN'05, pages 147–156, Berlin, Heidelberg, 2005. Springer-Verlag.

[KJ00] E Y Kim and K Jung. Automatic text region extraction using cluster-based templates. In *In: 4th Internat. Conf. on Advances in Pattern Recognition and Digital Techniques ICAPRDT*, pages 418–421, 2000.

[KLBP05] Sreekar Krishna, Greg Little, John Black, and Sethuraman Panchanathan. A wearable face recognition system for individuals with visual impairments. In *Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility*, Assets '05, pages 106–113, New York, NY, USA, 2005. ACM.

[KS08] Daehyun Kim and Kwanghoon Sohn. Static text region detection in video sequences using color and orientation consistencies. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1 –4, dec. 2008.

[LC03] Xiaoming Liu and Tsuhan Chen. Video-based face recognition using adaptive hidden markov models. pages 340–345, 2003.

[LCA+06] Stan Z. Li, Rufeng Chu, Meng Ao, Lun Zhang, and Ran He. Highly accurate and fast face recognition using near infrared images. In *in Proc. IAPR Int. Conf. Biometrics*, pages 151–158, 2006.

[LGP+05] Yixiong Liang, Weiguo Gong, Yingjun Pan, Weihong Li, and Zhenjiang Hu. Gabor features-based classification using svm for face recognition. In Jun Wang, Xiao-Feng Liao, and Zhang Yi, editors, *Advances in Neural Networks – ISNN 2005*, volume 3497 of *Lecture Notes in Computer Science*, pages 118–123. Springer Berlin Heidelberg, 2005.

[Liu08] Xu Liu. A camera phone based currency reader for the visually impaired. In *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*, Assets '08, pages 305–306, New York, NY, USA, 2008. ACM.

[LLL+11] Jung-Jin Lee, Pyoung-Hean Lee, Seong-Whan Lee, A. Yuille, and C. Koch. Adaboost for text detection in natural scene. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 429 –434, sept. 2011.

[Low04] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[LP90] L.M. Lifshitz and S.M. Pizer. A multiresolution hierarchical approach to image segmentation based on intensity extrema. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6):529–540, 1990.

[LS06] X. Liu and J. Samarabandu. Multiscale edge-based text extraction from complex images. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 1721 –1724, july 2006.

[LW02] Rainer Lienhart and Axel Wernicke. Localizing and segmenting text in images and videos, 2002.

[LWF05] J.N.K. Liu, Meng Wang, and Bo Feng. ibotguard: an internet-based intelligent robot security system using invariant face recognition against intruder. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 35(1):97 –105, feb. 2005.

[M+67] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA, 1967.

[MCUP02] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *In British Machine Vision Conference*, pages 384–393, 2002.

[MCUP04] Jiri Matas, Ondrej Chum, Martin Urban, and Tomáš Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004.

[MD09] Federico Matta and Jean-Luc Dugelay. Person recognition using facial video information: A state of the art. *J. Vis. Lang. Comput.*, 20(3):180–187, June 2009.

[MHLM05] M.A. Mattar, A.R. Hanson, and E.G. Learned-Miller. Sign classification using local and meta-features. In *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, page 26, june 2005.

[MJP00] Baback Moghaddam, Tony Jebara, and Alex Pentland. Bayesian face recognition. *Pattern Recognition*, 33(11):1771–1782, 2000.

[MKB10] Roberto Manduchi, Sri Kurniawan, and Homayoun Bagherinia. Blind guidance using mobile computer vision: a usability study. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*, ASSETS '10, pages 241–242, New York, NY, USA, 2010. ACM.

[MLK+06] James R. Marston, Jack M. Loomis, Roberta L. Klatzky, Reginald G. Golledge, and Ethan L. Smith. Evaluation of spatial displays for navigation without sight. *ACM Trans. Appl. Percept.*, 3(2):110–124, April 2006.

[Moo04] Hyeonjoon Moon. Biometrics person authentication using projection-based face recognition system in verification scenario. *Biometric Authentication*, pages 227–247, 2004.

[MP97] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):696–710, 1997.

[MRS+10] Sofia Mosci, Lorenzo Rosasco, Matteo Santoro, Alessandro Verri, and Silvia Villa. Solving structured sparsity regularization with proximal methods. In *ECML*, 2010.

[MW08] Ethan Meyers and Lior Wolf. Using biologically inspired features for face processing. *International Journal of Computer Vision*, 76:93–104, 2008.

[Nib86] W. Niblack. *An introduction to digital image processing*. Prentice-Hall International, 1986.

[Nix85] M. Nixon. Eye spacing measurement for facial recognition. In *SPIE Proceedings*, volume 575, pages 279–285, 1985.

[NM98] C. Nastar and M. Mitschke. Real-time face recognition using feature combination. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 312 –317, apr 1998.

[NM11a] L. Neumann and J. Matas. Text localization in real-world images using efficiently pruned exhaustive search. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 687 –691, sept. 2011.

[NM11b] Lukas Neumann and Jiri Matas. A method for text localization and recognition in real-world images. In *Proceedings of the 10th Asian conference on Computer vision - Volume Part III*, ACCV'10, pages 770–783, Berlin, Heidelberg, 2011. Springer-Verlag.

[NM12] L. Neumann and J. Matas. Real-time scene text localization and recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3538 –3545, june 2012.

[NMC12] Alexandru Niculescu-Mizil and Rich Caruana. Obtaining calibrated probabilities from boosting. *CoRR*, abs/1207.1403, 2012.

[OPH96] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

[OPM02] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[OSA94] J. Ohya, A. Shio, and S. Akamatsu. Recognizing characters in scene images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(2):214 –220, feb 1994.

[Pla99] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press, 1999.

[PMRR00] P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(10):1090–1104, October 2000.

[PMW10] V. Pradeep, G. Medioni, and J. Weiland. Robot vision for the visually impaired. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 15 –22, june 2010.

[RB00] David A. Ross and Bruce B. Blasch. Wearable interfaces for orientation and wayfinding. In *Proceedings of the fourth international ACM conference on Assistive technologies*, Assets '00, pages 193–200, New York, NY, USA, 2000. ACM.

[Rei93] D. Reisfeld. *Generalized Symmetry Transforms: Attentional Mechanisms and Face Recognition*. Tel Aviv University., 1993.

[RK04] Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *J. Mach. Learn. Res.*, 5:101–141, December 2004.

[RP+99] M. Riesenhuber, T. Poggio, et al. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2:1019–1025, 1999.

[RPTB01] Yossi Rubner, Jan Puzicha, Carlo Tomasi, and Joachim M Buhmann. Empirical evaluation of dissimilarity measures for color and texture. *Computer Vision and Image Understanding*, 84(1):25 – 43, 2001.

[SB91] M.J. Swain and D.H. Ballard. Color indexing. *International journal of computer vision*, 7(1):11–32, 1991.

[SBBW05] Linlin Shen, Li Bai, Daniel Bardsley, and Yangsheng Wang. Gabor feature selection for face recognition using improved adaboost learning. In *Proceedings of International Workshop on Biometric Recognition System, in conjuction with ICCV'05*, 2005.

[SDB+11] Joseph Shelton, Gerry Dozier, Kelvin Bryant, Joshua Adams, Khary Popplewell, Tamirat Abegaz, Kamilah Purrington, Damon L. Woodard, and Karl Ricanek. Genetic based lbp feature extraction and selection for facial recognition. In *Proceedings of the 49th Annual Southeast Regional Conference*, ACM-SE '11, pages 197–200, New York, NY, USA, 2011. ACM.

[SHT08] P. Shivakumara, Weihua Huang, and Chew Lim Tan. Efficient video text detection using edge features. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1 –4, dec. 2008.

[SL09] Conrad Sanderson and Brian C. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *Int. Conf.on Adv. in Biometrics*, pages 199–208, 2009.

[SLL09] Jitao Sang, Zhen Lei, and Stan Z. Li. Face image quality evaluation for iso/iec standards 19794-5 and 29794-5. In *Int. Conf. on Advances in Biometrics*, pages 229–238, 2009.

[SNDC07] Krishna Subramanian, Prem Natarajan, Michael Decerbo, and David Castañòn. Character-stroke detection for text-localization and extraction. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 1, pages 33–37. IEEE, 2007.

[SPT11] P. Shivakumara, Trung Quy Phan, and Chew Lim Tan. A laplacian approach to multi-oriented text detection in video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(2):412 –419, feb. 2011.

[SR03] L.K. Saul and S.T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *The Journal of Machine Learning Research*, 4:119–155, 2003.

[SSC11] P. Sanketi, H. Shen, and J.M. Coughlan. Localizing blurry and low-resolution text in natural images. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 503–510. IEEE, 2011.

[Suk00] Gita Sukthankar. Face recognition: A critical look at biologically-inspired approaches. Technical report, 2000.

[SW96] D.L. Swets and J.J. Weng. Using discriminant eigenfeatures for image retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(8):831 –836, aug 1996.

[SWH+05] P. Silapachote, J. Weinman, A. Hanson, M.A. Mattar, and R. Weiss. Automatic sign detection and recognition in natural scenes. In *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, page 27, june 2005.

[TBGL86] Q. Tian, M. Barbero, Z.H. Gu, and S.H. Lee. Image classification by the foley-sammon transform. *Optical Engineering*, 25(7):257834–257834, 1986.

[TBR+05] B.S. Tjan, P.J. Beckmann, R. Roy, N. Giudice, and G.E. Legge. Digital sign system for indoor wayfinding for the visually impaired. In *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, page 30, june 2005.

[TCYZ05] Zhuowen Tu, Xiangrong Chen, Alan L Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of Computer Vision*, 63(2):113–140, 2005.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 1996.

[TP91] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[TT07] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *Analysis and Modeling of Faces and Gestures*, pages 168–182, 2007.

[Vap99] V. Vapnik. *The nature of statistical learning theory*. springer, 1999.

[VJ01] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.

[VJ04] Paul Viola and Michael Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2004.

[WBB11] K. Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1457–1464. IEEE, 2011.

[WCB10] T. Winlock, E. Christiansen, and S. Belongie. Toward real-time grocery detection for the visually impaired. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 49 –56, june 2010.

[WCM+11] Yongkang Wong, Shaokang Chen, S. Mau, C. Sanderson, and B.C. Lovell. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 74 –81, june 2011.

[WFKvdM97] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. In *Image Processing, 1997. Proceedings., International Conference on*, volume 1, pages 129 –132 vol.1, oct 1997.

[WHL09] Xiufei Wang, Lei Huang, and Changping Liu. A new block partitioned text feature for text verification. In *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, pages 366–370. IEEE, 2009.

[WL+05] Bruce N Walker, Jeffrey Lindsay, et al. Navigation performance in a virtual environment with bonephones. In *Proc. of the Int'l Conf. on Auditory Display (ICAD2005)*, pages 260–263, 2005.

[WW10] Yong Wang and Yi Wu. Face recognition using intrinsicfaces. *Pattern Recognition*, 43(10):3580 – 3590, 2010.

[WYG+09] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Yi Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210 –227, feb. 2009.

[YFM98] O. Yamaguchi, K.. Fukui, and K.-i. Maeda. Face recognition using temporal image sequence. In *Int. Conf. on Autom. Face and Gesture Recognition*, pages 318–323, 1998.

[YHC92] A.L. Yuille, P.W. Hallinan, and D.S. Cohen. Feature extraction from faces using deformable templates. *International journal of computer vision*, 8(2):99–111, 1992.

[YM04] D. Yuan and R. Manduchi. A tool for range sensing and environment discovery for the blind. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, pages 39–39. IEEE, 2004.

[ZCK98] W. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 336–341. IEEE, 1998.

[ZCPR03] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *Acm Computing Surveys (CSUR)*, 35(4):399–458, 2003.

[ZDO09] Luca Zini, Augusto Destrero, and Francesca Odone. A classification architecture based on connected components for text detection in unconstrained environments. In *Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, AVSS '09, pages 176–181, Washington, DC, USA, 2009. IEEE Computer Society.

[ZG09] Xiaozheng Zhang and Yongsheng Gao. Face recognition across pose: A review. *Pattern Recognition*, 42(11):2876 – 2896, 2009.

[ZHL+05] Guangcheng Zhang, Xiangsheng Huang, Stan Li, Yangsheng Wang, and Xihong Wu. Boosting local binary pattern (lbp)-based face recognition. In Stan Li, Jianhuang Lai, Tieniu Tan, Guocan Feng, and Yunhong Wang, editors, *Advances in Biometric Person Authentication*, volume 3338 of *Lecture Notes in Computer Science*, pages 179–186. Springer Berlin / Heidelberg, 2005. 10.1007/978-3-540-30548-4_21.

[ZKJ95] Yu Zhong, K. Karu, and A.K. Jain. Locating text in complex color images. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 146 –149 vol.1, aug 1995.

[ZP07] Guoying Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):915 –928, june 2007.

[ZQJX07] Kai-hua Zhu, Fei-hu Qi, Ren-jie Jiang, and Li Xu. Automatic character detection and segmentation in natural scene images. *Journal of Zhejiang University SCIENCE A*, 8:63–71, 2007.

[ZZ10a] Cha Zhang and Zhengyou Zhang. A survey of recent advances in face detection. *Microsoft Research, June*, 2010.

[ZZ10b] Cha Zhang and Zhengyou Zhang. A survey of recent advances in face detection. *Microsoft Research, June*, 2010.