



Scuola Dottorale di Ingegneria
Sezione di Ingegneria dell'Elettronica Biomedica,
dell'Elettromagnetismo e delle Telecomunicazioni
XXVII Ciclo della formazione Dottorale

TESI DI DOTTORATO

Web based Face Recognition for Visual
Internet of Things

Candidato: Marco Leo

Docente guida: Prof. Alessandro Neri

Coordinatore del Dottorato

Prof. Alessandro Salvini

Roma, 30 Aprile 2015

A Dissertation submitted to
University "Roma Tre", Rome, Italy
Department of Applied Electronics
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

*A Maria Grazia e a Sofia e Giorgia
che mi sono sempre vicine*

ABSTRACT

ὄψει δ' ἄλλοτε μὲν μιν ἐνωπαδίως ἐσίδεσκεν,
ἄλλοτε δ' ἀγνώσασκε κακὰ χροῖ εἵματ' ἔχοντα.

Ὀδύσσεια - XXIII vv 94-95

ora, cogli occhi, lo ravvisava nel viso,
ora, per le sue misere vesti, non lo riconosceva.

Odissea – Libro XXIII vv 94-95

Penelope non è certa dell'identità dello sconosciuto, lo osserva, a volte le sembra di riconoscerlo, a volte è tratta in inganno dall'aspetto trasandato dello straniero i cui tratti sembrano lontani dall'Ulisse partito da Itaca venti anni prima. La difficoltà di Penelope è la medesima che tutti hanno quotidianamente nella relazione con gli altri.

Quando le persone si incontrano l'attività principale che il cervello umano svolge è quella di associare il volto ad una identità nota. Questo esercizio, normalmente semplice, può risultare molto complesso nel momento in cui i tratti somatici non sono nitidi all'occhio umano, quando il processo di associazione del volto è ostacolato da difficoltà di memoria oppure quando il cervello è ingannato da similarità tra lineamenti di soggetti diversi.

Questa funzione, comune per gli uomini, è molto complessa per i sistemi automatici. Negli ultimi 40 anni sono stati effettuati numerosi studi e sviluppati diversi algoritmi per il riconoscimento dei volti ma l'efficienza dell'uomo nell'associare un volto ad una identità nella varietà delle situazioni della vita quotidiana è ancora lontana. Gli studi sono stati orientati sia a sviluppare tecniche per il riconoscimento automatico di elementi biometrici come l'ovale del viso, la bocca o gli occhi sia per l'associazione di immagini di persone con identità note. Mentre le tecniche per il riconoscimento automatico di elementi biometrici ha avuto un buon successo, come dimostra la diffusa applicazione nei prodotti di consumo come le fotocamere, le soluzioni per l'associazione delle identità sono efficaci solo in condizioni nelle quali il processo di acquisizione delle immagini è ben controllato. Attualmente lo studio degli algoritmi per il riconoscimento automatico dei volti è di interesse per lo sviluppo di strumenti software in grado di modificare i processi di produzione multimediali che sono fondamentalmente basati sulla capacità di un

operatore umano di interpretare una sequenza di immagini. L'applicazione di strumenti automatici per il riconoscimento di volti nelle sequenze multimediali è una modalità operativa che non è ancora completamente consolidata e che non ha trovato applicazione nell'industria settoriale nella quale tuttora si utilizza la capacità umana come strumento di decisione.

Il principale lavoro svolto nel corso del triennio dottorale, il cui percorso è illustrato in questa tesi, è stato quello di studiare un sistema di riconoscimento automatico di volti che consenta l'identificazione di un individuo e la relativa annotazione del contenuto multimediale utilizzando il Web come archivio di immagini. Nell'ambito di questo lavoro di ricerca il Web è stato utilizzato per la realizzazione di un nuovo database di immagini per apprendimento (gallery) dei classificatori con l'obiettivo di proporre elementi innovativi per l'Internet of Things (IoT), nella sua evoluzione identificata come "Visual Internet of Things" (VIoT).

Il primo capitolo illustra lo stato dell'arte delle tecnologie per la rilevazione e il riconoscimento delle identità dei volti presenti nelle immagini partendo dall'analisi della letteratura scientifica più recente, presentando i maggiori archivi di immagini disponibili per la ricerca scientifica ed infine introducendo il concetto di Visual Internet of Things.

Il secondo capitolo descrive gli elementi fondamentali dei processi di riconoscimento facendo riferimento alle tecnologie di pattern recognition che sono alla base degli algoritmi finalizzati al riconoscimento dei volti. Vengono presentati i principali algoritmi sviluppati finora e i principali approcci adottati.

Il terzo capitolo descrive il lavoro di preparazione effettuato sulle sequenze video, i principali problemi affrontati nell'estrazione dei quadri rilevanti ai fini dell'identificazione dei soggetti e i principali risultati ottenuti. Tutte le sequenze video sono state acquisite mediante sistemi di registrazione commerciali da trasmissioni TV e successivamente sono state elaborate al fine di poter ottenere delle sequenze di quadri alle quali applicare algoritmi per la rilevazione dei volti. Si è proceduto ad effettuare registrazioni sia da trasmissioni in HD a 1920x1080i che in SD a 720x576i per poter effettuare a una serie di confronti prestazionali.

Il quarto capitolo descrive le varie fasi seguite nel processo di riconoscimento partendo dalle prime prove effettuate con un insieme ridotto di volti fino alla descrizione di prove effettuate con più algoritmi su un campione esteso di quadri sia in HD che in SD. Nel capitolo è descritto anche il processo seguito per la realizzazione dell'archivio di immagini campione scaricate dal

Web. Vengono presentati I principali risultati ottenuti pubblicati in occasione di alcune conferenze scientifiche.

Nel quinto capitolo, dedicato alle conclusioni, viene presentata una sintesi dei principali risultati, sono messi in luce i maggiori problemi per lo sviluppo di tali soluzioni tecnologiche e sono indicati alcuni temi di approfondimento. Infine viene presentata una breve panoramica sui servizi di prossima generazione in corso di sviluppo e di applicazione da parte di alcune aziende.

ABSTRACT

ὄψει δ' ἄλλοτε μὲν μιν ἐνωπαδίως ἐσίδεσκεν,
ἄλλοτε δ' ἀγνώσασκε κακὰ χροῖ εἵματ' ἔχοντα.

Ὀδύσσεια – XXIII vv 94-95

and now with her eyes she would look full upon his face, and now again
she would fail to know him, for that he had upon him mean raiment.

Odyssey – Book XXIII – lines 94-95

Penelope is not confident of her feeling about the identity of the unknown, she observes him, sometimes she seems recognizing him, sometimes she is misled by the scruffy-looking foreign aspect whose features seem far from the Ulysses who departed from Ithaca twenty years before. The difficulty of Penelope is the same in that all have a daily relationship with others.

When people meet, the main activity that the human brain performs is to associate the face to a known identity. This exercise, usually simple, can be very complicated when the facial features are not clear to the human eye, when the process of association of the face is hampered by poor memory or when the brain is fooled by the similarity between features of different subjects.

This feature, common to humans, is very difficult for automatic systems. In the last 40 years there have been numerous studies and developed several algorithms for face recognition but the efficiency of men to associate a face to an identity in the variety of situations in everyday life is still far away. The studies were directed both to develop techniques for automatic recognition of biometrics as the oval of the face, mouth or eyes and to associate an image of a person with a known identity. While the techniques for automatic recognition of biometrics has been quite successful, as evidenced by the widespread application in consumer products such as cameras, solutions for the association of identity are effective only in conditions in which the process of image acquisition is well controlled. Currently, the study of the algorithms for the automatic recognition of faces is of interest for security applications or for the development of software tools able to modify the multimedia production processes that are fundamentally based on the capability of a human operator to understand a sequence of images. The application of automated tools for the recognition of faces in multimedia sequences is an operative mode that

is not yet fully established and that has not been applied in the industry sector in which still is in use the human capability as decision tool.

The main work carried out during the three-year doctoral, whose path is shown in this thesis, was to study a system of automatic recognition of faces enabling the identification of an individual and the pertaining annotation of multimedia content using the Web as image database. In the context of this research work, the Web was used for the creation of a new database of images for the learning procedures (gallery) of the classifiers with the aim to propose innovative elements for the Internet of Things (IoT), in its evolution identified as "Visual Internet of Things" (VIoT).

Chapter one describes the state of the art for the detection and recognition technologies of the identity of faces in images. It begins with an analysis of the most recent scientific literature, presenting the main archives of images available for scientific research and finally introducing the concept of VIoT.

The second chapter describes the basic elements of the recognition process by referring to the pattern recognition technologies that are the basis of algorithms aimed at identification of objects and subsequently the faces. We present the main algorithms developed and the applied approaches.

The third chapter describes the preparation work carried out on video sequences, the main problems faced in extracting the relevant frameworks for identifying the parties and the main results obtained. All video sequences have been acquired through standard registration systems from commercial TV broadcasts and subsequently have been elaborated in order to obtain sequences of pictures on which apply algorithms for the detection and identification of faces. We proceeded to record contents both from HD broadcasts at 1920x1080i and from SD at 720x576i in order to carry out a set of comparisons of performance.

The fourth chapter describes the various steps followed in the recognition process starting from the first tests carried out with a reduced set of faces to the description of the tests carried out with more algorithms on a larger sample of pictures in both HD and SD. The chapter also describes the process followed for the creation of the archive of sample images downloaded from the Web. We present the main results published in some scientific conferences.

The fifth chapter, dedicated to conclusions, contains a summary of the main results and includes a review of the major issues for the development of such technological solutions proposing some topics for additional studies. Finally there is a short overview of the next-generation services in progress of development and implementation by some companies.

RINGRAZIAMENTI

Al termine del percorso di questi tre anni di dottorato desidero ringraziare tutte le persone che a vario titolo mi hanno accompagnato in questo cammino e senza le quali questo lavoro di tesi non sarebbe stato possibile realizzare.

Vorrei ringraziare innanzitutto il Prof. Alessandro Neri per avermi offerto la possibilità di elevare il mio percorso di crescita professionale e per essersi sempre dimostrato disponibile a offrirmi il proprio prezioso contributo teorico e metodologico durante varie fasi del mio lavoro di ricerca. Lo ringrazio per le sue approfondite spiegazioni volte a far sì che potessi approfondire alcuni aspetti dell'ingegneria e per tutte le opportunità di formazione verso le quali mi ha guidato per la mia crescita come dottore di ricerca.

Ringrazio il Prof. Vatalaro per avermi stimolato, insieme al Prof. Neri, nell'intraprendere il percorso di approfondimento della scuola dottorale affinché potessi arricchire e completare il mio percorso di studi.

Ringrazio il Prof. Marco Carli e la Dott.ssa Federica Battisti per il supporto scientifico, per la stima dimostratami, per i suggerimenti ed i consigli che mi hanno fatto da guida in nel tempo trascorso presso COMLAB.

Voglio, inoltre, ringraziare tutti i colleghi e amici dottorandi con cui ho condiviso lezioni, impegni, idee e soddisfazioni.

Ringrazio, infine, immensamente Maria Grazia, Sofia e Giorgia per non avermi mai fatto mancare il proprio sostegno e aver sopportato qualche ora in più di assenza dalla vita familiare. Alcune volte, un po' come Penelope, hanno avuto la pazienza di aspettarmi ma senza di loro nulla sarebbe stato possibile.

ACKNOWLEDGMENTS

At the conclusion of these three years of PhD course I would like to thank all the people who for various reasons was near me during on this journey, without them this work would not have been possible to carry out.

I would first like to thank Prof. Alessandro Neri who gave me the opportunity to improve my professional growth and has always been available to offer me his valuable theoretical and methodological contribution during the various stages of my research work. Thank for his detailed explanations aimed at ensure that I could study in depth some aspects of engineering and for all training opportunities towards which he guided me in my path as a PhD.

Thanks to Prof. Vatalaro for me his stimulus, together with prof. Neri, in undertaking the path of the doctoral school so that I could enhance and complete my studies.

Thanks to Prof. Marco Carli and Dr. Federica Battisti for the scientific support and for the estimation they show in me, for the tips and advices useful for me in the time spent at COMLAB.

I would also like to thank all my colleagues and friends with whom I shared lessons, commitments, ideas and satisfaction.

Lastly, I thank immensely Maria Grazia, Sofia and Giorgia for not having never failed to lend their support and by the fact that they have endured a few extra hours of absence from my family life. Sometimes, a bit as Penelope, they have had the patience to wait for me but without them nothing would be possible.

Table of Contents

ABSTRACT (ITA)	4
ABSTRACT (ENG)	7
RINGRAZIAMENTI	10
ACKNOWLEDGMENTS	11
TABLE OF FIGURES	16
LIST OF TABLES	19
LIST OF ABBREVIATIONS	20
CHAPTER 1	21
1. INTRODUCTION	21
1.1. FIELDS OF APPLICATION	24
1.2. STATE OF ART	25
1.3. THE APPROACHES IN FACE RECOGNITION	26
1.3.1. UNCONSTRAINED FACE RECOGNITION	28
1.4. GALLERY SETS FOR FACE RECOGNITION	31
1.5. GALLERY SETS WITH IMAGES FROM WEB	39
1.6. THE CONCEPT OF VISUAL INTERNET OF THINGS	41
CHAPTER 2	43
2. FACE RECOGNITION PROBLEM	43
2.1. PATTERN RECOGNITION PROCESS	44

2.1.1.	ACQUISITION _____	45
2.1.2.	EDITING _____	45
2.1.3.	FEATURE ANALYSIS _____	46
2.1.4.	CLASSIFICATION _____	46
2.1.5.	ASSESSMENT _____	46
2.2.	BASIC CONCEPTS OF PATTERN RECOGNITION _____	47
2.2.1.	LEARNING MODELS _____	47
2.2.2.	TRAINING _____	48
2.3.	FEATURE EXTRACTION METHODS _____	48
2.4.	TECHNIQUES AND ALGORITHMS FOR FACE DETECTION _____	49
2.4.1.	HOLISTIC APPROACH _____	49
2.4.2.	FEATURE BASED METHODS _____	52
2.4.1.	HYBRID METHODS _____	54
CHAPTER 3	_____	56
3.	THE AUTOMATIC TAGGING OF FACES IN VIDEO SEQUENCES _____	56
3.1.	FACE RECOGNITION IN VIDEOS _____	58
3.2.	THE PARAMETERS OF VIDEO SEQUENCES _____	62
3.2.1.	PIXEL ASPECT RATIO BASICS _____	64
3.2.2.	INTERLACED AND PROGRESSIVE VIDEO _____	65
3.3.	VIDEO SEGMENTATION AND CROPPING _____	67

3.3.1.	PROCESSING OF VIDEO SEQUENCES	67
3.3.1.1.	EDGE CHANGE RATIO	68
3.3.1.2.	HISTOGRAM ANALYSIS	69
3.4.	SCENE DETECTION	69
3.5.	FACE DETECTION IN VIDEOS	71
3.5.1.	QUALITY OF FACES	74
3.5.2.	AUTOMATIC CROPPING AND ZOOMING OF IMAGES	74
3.6.	RESULTS OF VIDEO PRE-PROCESSING ACTIVITIES	76
3.6.1.	FACE DETECTION E TRACKING	76
3.6.2.	ANALYSIS OF PERFORMANCES	78
3.6.2.1.	SEARCH OF THE FACE WITH THE HIGHEST NUMBER OF KEY FEATURES	78
3.6.2.2.	FACE DETECTION RESULTS	80
3.7.	CRITICAL CASES IN FACE DETECTION	83
CHAPTER 4		87
4.	EXPERIMENTATION AND RESULTS EXAMINATION	87
4.1.	THE ADOPTED EXPERIMENTAL MODEL	87
4.1.	ALGORITHMS FOR THE MANIPULATIONS OF IMAGES AND THE EXTRACTION OF THE FEATURES OF THE FACE	88
4.1.1.	EIGENFACES	88
4.1.2.	SIFT	90
4.2.	THE DATABASE OF FACES CREATED USING WEB IMAGES	93

4.3.	SELECTION OF FACES FROM VIDEOS	94
4.4.	FACE RECOGNITION PROCESS	97
4.5.	PERFORMANCE COMPARISON	98
CHAPTER 5		106
5.	CONCLUSIONS AND FUTURE WORKS	106
5.1.	CONCLUSIONS	107
5.2.	FUTURE WORKS	109
6.	REFERENCES	111

TABLE OF FIGURES

Figure 1 – Identification Card containing the Bertillon Measurements	22
Figure 2 - Photo Composition Template U.S. Passport and International Travel in US	28
Figure 3 - Face recognition typical pipeline	29
Figure 4 – Sample of colour version of FERET	33
Figure 5 - Sample of UMIST database	33
Figure 6 - Sample of Yale Face Database	34
Figure 7 - Sample of AT&T database	34
Figure 8 - Samples of BANCA database	35
Figure 9 - FRGC sample (from www.clementcreusot.com/phd/)	36
Figure 10 - Sample of LFW database	37
Figure 11 - Sample of BioID database	38
Figure 12 - The basic stages involved in the design of a classification system	45
Figure 13 - Holistic basis images from a) Principal Component Analysis b) Linear Discriminant Analysis	51
Figure 14 - Taxonomy of the face recognition methods based on PCA	51
Figure 15 - Match results for one of the test images	53
Figure 16 - Square pixels and 4:3 frame aspect ratio (A), nonsquare pixels and 4:3 frame aspect ratio (B), and nonsquare pixels displayed on a square-pixel monitor (C)	65
Figure 17 - For interlaced video, entire upper field (odd-numbered lines) is drawn to screen first, from top to bottom, in one pass (A). Next, entire lower field (even-numbered lines) is drawn to screen, from top to bottom, in one pass (B). For noninterlaced video	66
Figure 18 – Distortion produced by interlacing over a frame acquired from broadcast TV	67
Figure 19 – Example of Edge Detection	69
Figure 20 - Diagram of the events identified by the algorithm for scene change detection	70
Figure 21 - Representation by Integral Image	72
Figure 22 – Classifier Cascade	73
Figure 23 - Scene with the detection of two faces	75
Figure 24 - Result of the automatic scaling of faces in which two images of different sizes initials have been reported both to the size of 180x200 pixels without loss of proportions	76

Figure 25 - Flow chart of the Detection and Tracking algorithm 77

Figure 26 - Algorithm for finding the best frame in a video sequence..... 79

Figure 27 – Case of application of KLT algorithm: On the right the frame with a face with a number of features with overlap of features 80

Figure 28 - Statistics on total number of events 81

Figure 29 - Statistics on total number of events TG#1..... 81

Figure 30 - Statistics on total number of events TG#1..... 82

Figure 31 – Samples of frames acquired in the latest phase of work. On the left a sample in HD and on the right a sample in SD. 82

Figure 32 - Map of matching face with biometric elements not detectable..... 83

Figure 33 - Map of the hue in which a face in the foreground has a uniform colour from the background 84

Figure 34 - Case of a figure with a background with regular geometric elements 84

Figure 35 - Case of detection of a face almost totally occluded 85

Figure 36 - Case of a correct detection of a face is not occluded 85

Figure 37 - The case of a false negative in the face of partial occlusion 86

Figure 38 - The case of a false positive detected a piece of furniture 86

Figure 39 - Block diagram of the proposed detection and tracking process..... 88

Figure 40 – Euclidean distance of a face collected from Web and two images captured from videos from the Eigenspace of the same subject 90

Figure 41 – The SIFT descriptor is a spatial histogram of the image gradient..... 91

Figure 42 – Comparison of images by means SIFT of the same subject..... 92

Figure 43 – A subset of annotated faces DB downloaded from Web 94

Figure 44 - Face detection, zooming and cropping..... 94

Figure 45 - Example of region containing face..... 95

Figure 46 - Regions containing faces extracted from the same news stream at different times. As can be noticed in different frames, the same character image differs from pose, expression, and lighting conditions. 95

Figure 47- Face DB extracted from the recorded TV streams..... 96

Figure 48 – Recognition process implemented in experimentation 97

Figure 49 – Best candidate calculating the Euclidean distance..... 99

Figure 50 - Comparison between still to still and video to still images using SIFT 102

Figure 51 - ROC curve using SD test set 104

Figure 52 - ROC curve using HD test set 104

LIST OF TABLES

Table 1 - Main databases for face recognition	39
Table 2 – Enrolment table provided by NIST	59
Table 3 - Video features for the datasets foreseen from NIST	60
Table 4 - Digital television standards	64
Table 5 – Performances in application of KLT algorithm	80
Table 6 – Total number of events	81
Table 7 - Number of events TG #1	81
Table 8 - Number of events TG #2	82
Table 9 – Eigenfaces results	100
Table 10 – SIFT results	100
Table 11 – Results of comparison SD/HD	103
Table 12 – Global score combining Eigenfaces and SIFT	103

LIST OF ABBREVIATIONS

2DPCA	Two Dimensional PCA
ATSC	Advanced Television Standards Committee
AUC	Area Under Curve
BANCA	Biometric Access Control for Networked and e-commerce Applications
DAR	Display Aspect Ratio
DARPA	Defense Advanced Research Products Agency
DoG	Difference of Gaussians
DVB-T	Digital Video Broadcasting - Terrestrial
ECR	Edge Change Ratio
FERET	Face Recognition Technology
FIVE	Face in Video Evaluation
FRGC	Face Recognition Grand Challenge
FRVT	Face Recognition Vendor Test
HD	High Definition
HIT	Human Intelligence Tasks
HMM	Hidden Markov Models
HOG	Histogram of Oriented Gradients
IoT	Internet of Things
ISDB-T	Integrated Services Digital Broadcasting
KLDA	Kernel LDA
KLT	Kanade-Lucas-Tomasi
KPCA	Kernel PCA
LBP	Local Binary Patterns
LDA	Linear Discriminant Analysis
LFW	Labeled Faces in the Wild
NIST	National Institute of Standards and Technology
PAR	Pixel Aspect Ratio
PCA	Principal Component Analysis
ROC	Receiver Operating Characteristic
S2S	Still-to-Still
S2V	Still-to-Video
SAR	Storage Aspect Ratio
SD	Standard Definition
SIFT	Scale Invariant Feature Transform
SSS	Small Sample Size
SURF	Speeded Up Robust Features
SVM	Support Vector Machines
V2S	Video-to-Still
V2V	Video-to-video
VIoT	Visual Internet of Things
YTF	YouTube Faces

CHAPTER 1

1. INTRODUCTION

Facial recognition technology has emerged as an attractive solution to address many current service scenarios for identification and the verification of identity in the field of security, banking, social or legal.

The face processing research has been applied in many commercial applications with significant examples in face detection, nevertheless the ability to go beyond the detecting of faces and to associate them the identity of the persons pictured can have an even larger number of potential applications. Among these, for example, the security systems that, in many applications, are already able to detect and automatically track human objects but that could improve their capabilities by reporting the identities by searching specific characters in a group or in a stream.

The face recognition was the subject of a great deal of research even before the use of computers. A great contribution was done by Alphonse Bertillon in order to recognize criminals who were repeatedly arrested without having to resort to large collections of portraits [1]. Bertillon was a French criminologist who first developed, in 1879, an anthropometric method of physical measurements of body parts, especially components of the head and face, in order to produce a detailed description of an individual and obtain a “spoken portrait” (Figure 1).


DEPARTMENT OF POLICE SERVICE				NEW HAVEN, CONN.			
Name: Edward Blasco.		Aliases: Wm. Braen.		Color: Wh.			
Crime: Trespass on R.R. Cars.		Disposition: City Court, Nov. 27, 1933, suspended sentence		Date of Arrest: Nov. 23d, 1933.			
BERTILLON MEASUREMENTS							
Height	5' 10"	Head length	10 1/2"	L. Foot	10 1/2"	Chin to Ears	Age 17 Born in
Stretch	5' 10"	Head Width	10 1/2"	L. Mid. F.	10 1/2"	Periph. Z.	Apparent Age 17.
Trunk	5' 10"	Check Width	10 1/2"	L. Lh. F.	10 1/2"	Prent.	Nativity ?
Eng. Height	6' 1"	R. Ear	10 1/2"	L. Fore A.	10 1/2"	Prent.	Occupation NONE.
Remarks incident to Measurement		Height in shoes.					
							
DESCRIPTIVE							
Right index finger to be impressed IMMEDIATELY after Signature is written		Inclin.	Stidge	Build	Stature	Teeth	Complexion
		Height	Base	Build	Stature	Up. ft.,	Med. dk.
		Width	Length	Projection	Brush	are bad.	Weight 137.
		Perul.	Perul.	Chin	Build	Slim.	
Measured at Police Headquarters, New Haven, Conn., Date Nov. 23d, 1933.							
Prisoner's Signature		Edward Blasco					
Remarks: Blasco being the one and two others were arrested at Cedar Hill 2 1/2 yards if they had reached where the electric wires are, would have been killed							

Figure 1 – Identification Card containing the Bertillon Measurements

Tracking and recognizing face objects is a standard task for humans that act in the daily environment using the natural capability of capture and process data from the surrounding environment and the capabilities of the brain. Instead the automatic systems require complex settings and training processes in order to reach acceptable performances in many applications. The research in this field is active on several sides and different approaches have been proposed in order to face the several problems of the automatic systems. As example of the difficulties, the appearance of the face objects can completely change due to lighting and viewing variations and hence the illumination or pose introduces problems for wide applications of this type of applications with a sufficient level of trustworthiness.

The human faces are frequently present in multimedia contents such as news, films or home video and the design of an indexing process able to perform automatic face recognition is a challenging issue in order to generate segments of coherent video contents for browsing, skimming or summarization. As other techniques as speech recognition, natural language processing and other understanding processes, the face processing can be a powerful tool for automatic indexing and retrieval of digital multimedia contents.

In the past the problem of the face recognition has been associated mainly with the face detection task. This approach has focused the attention of the researchers on the features of the digital images so several algorithms have been designed in order detect the biometrics elements such as mouth, eyes or nose. The detection application has found an important

moment in 2001 when Viola and Jones proposed an efficient algorithm for accurate automatic detection of faces and a widespread use after 2005 when such technology was introduced into consumer-level digital cameras becoming today a standard feature on most digital cameras in order to assist in setting parameters such as focus, exposure and colour balance.

Recently the detection task has been disjoined by the problem of the recognition and in many applications, like image and video indexing and retrieval or video scene classification/news summarization, the tools for recognition can improve the current applications by adding automatic tags in order to reduce the time of the human work on trivial tasks and improving the performances in processing of contents.

The problem of face recognition is close to standard pattern classification or machine learning problem where, given a set of face labelled images with the person's identity (the gallery set) and an unlabelled set of face images of the same group of people (the probe set), a matcher must be able to make the pair between the images of the probe and the gallery sets.

Usually the problem of recognition is divided in three steps:

- the face is detected, selected and located in a video or in an image (face detection);
- a collection of characteristic elements, known as a feature vector, is extracted from each detected face;
- a classifier is trained to assign to each feature vector a label with a person's identity by using an archive of known characters.

The classifiers are mathematical functions which, given a feature vector, return a value, or a set of values, as the measurement of correspondence to a subject's identity.

Currently the real applications of automatic face recognition are not till reliable in many service scenarios and there are examples of face recognition systems not performing up to expectations when deployed in commercial applications [2].

In this thesis the focus is on the capabilities of automatic systems to perform the face recognition on multimedia videos using the Web as knowledge base.

The target is the application of the concept of Internet of Things (IoT) also to image processing. IoT is basically the capability of objects to be source and destination of information. As well as the physical objects, also the multimedia contents can be elements able to produce information. The application of the Computer Vision tools for the face

recognition, or identification methods, turns the multimedia contents into a sort of virtual object able to produce information and apply the concept of Visual Internet of Things [3].

1.1. FIELDS OF APPLICATION

Many applications implement the automatic face recognition for example for security in application for mobile device authentication, identity card duplication or verification. Face recognition technology also has the potential to improve the law enforcement in surveillance, forensic identification or video annotation. The capability of these systems has not yet been fulfilled due to the essential difficulties in application of matching algorithms with low-quality face images and videos mainly in unconstrained environments. The face recognition in video annotation or in law enforcement scenarios foresees a workflow where the “humans in the loop” are the elements able to perform the matching and where the frames of interest are extracted in order to tag the candidate faces manually. Till now the researchers have focused the attention on still-image matching problem mainly using large face databases such as driver’s licenses or mug shot photos.

The National Institute of Standards and Technology (NIST) sponsored in 2006 the Face Recognition Vendor Test (FRVT) which is one of the most important large scale independent synopsis of the state-of-the-art for face recognition systems [4]. The previous tests in the series were the FERET, FRVT 2000, and FRVT 2002. The interest of NIST for face recognition is growing in order extent the application also in more challenging scenarios including face recognition in videos. Additionally, the increase of IP cameras, the large use of video portal, the ubiquity of CCTV and the videocameras on smartphones, in conjunction with emerging forensic and security scenarios, improves the need of automated face recognition; there are a lot of examples of criminal or terrorist incidents where a mature face recognition technology for unconstrained face matching in videos could have been valuable to law enforcement agencies. The FRVT 2006 measured the progress of facial recognition systems including commercial systems using Windows or Linux based algorithms. The acquired data comprised a large standard dataset of “full frontal” pictures provided to NIST by the U.S. State Department using non-conforming pixel resolutions and lighting angles of 36,000 pictures of persons applying for non-immigrant visas at U.S. consulates in Mexico. The tests evaluated 4 dimensions of facial recognition: high resolution still imagery, 3D facial scans,

multi-sample still facial imagery, and pre-processing algorithms that compensate for pose and illumination.

An important field of application is also the video processing for image tagging in service scenarios where the tasks are video-to-video, still-to-video or video-to-still face matching. Some professional companies offer video annotation services for monitoring of the presence of specific characters or for commercial breaks.

The market of video annotation is basically based on crowdsourcing paradigms. Amazon Mechanical Turk (MTurk) is a website used for “crowdsourcing” (retrieving information) from a large number of human participants (workers) [5]. A worker on MTurk can do simple Human Intelligence Tasks (HITs) for a “requester.” Mechanical Turk provides a way for responses on the exact same task to be collected from many different individuals, so a “crowd” is effectively working together to form one final response or result. The workers must annotate the start, end boundaries and the label of all occurrences of activities in videos. Individual filtering checks the consistence in the answers of each tagger with the characteristic of dataset in order to identify and remove nonserious taggers. Collaborative filtering checks the agreement in annotations among taggers. The filtering techniques detect and remove non-serious taggers and finally the majority voting applied by the Mechanical Turk generates a single final annotation set.

1.2. STATE OF ART

In the last decades, the increased availability of multimedia content has increased to the need of designing fast and efficient systems for analysing video and audio contents. Additionally the need for automatic, unsupervised, information extraction is of primary importance for exploring large databases or for collecting in a fast and efficient way statistics on multimedia content.

To this aim, the face detection and face recognition are of great support in this task. The first attempts to find in an image, or more in general in a video sequence, the presence of faces while the latter tries to match whom the identified face belongs to. Both tasks have to face with several problems strictly linked to the nature of the multimedia content such as quality, partial visibility of faces, illumination, pose and perspective. In case of video analysis, the goal becomes even more challenging since those parameters can vary over time.

Face identification and face recognition have been widely investigated in literature and several approaches have been proposed based on Eigenfaces, Neural Networks, Hidden Markov Models (HMM), geometrical feature matching, template matching, line edge map and thermal image based technique as detailed in [6]. Moreover, the possibility of recognizing characters in TV series and movies has been recently widely investigated. In [7][8][9], the authors propose different approaches for automatically inserting textual data in TV series and movies with the aim of unsupervised labelling of specific characters appearing in the scenes. In [10], Discrete Wavelet Transform and Support Vector Machine are used for detecting and recognizing subjects in a video sequence. In [11] the problem of identifying faces automatically collected from TV streams is presented. The approach is based on an unsupervised definition of a cast-specific metric, adapted to characters appearing in videos. This technique allows not only faces labelling but also to recognize and to cluster the characters. The authors in [12] present a person track based approach rather than a face track one. The aim is to label all characters in TV contents by probabilistically fusing info on face recognition, clothing appearance, speaker recognition and contextual constraints.

In [13] the entity extraction from transcripts and video-caption recognition were integrated with temporal information to boost the overall accuracy of retrieval. In [14] some video shots related to a named individual were found by exploring various information sources from video data, such as names appearing in transcripts, face information and, most importantly, the temporal alignment of names and faces. In [15] the faces have been labelled with their corresponding names using supervised learning methods such as Support Vector Machines (SVM) and multiple instance learning. In order to avoid high variations of detected faces in video, Zhai et al. in [16], instead of using detected faces, used the “body” as extended face region (e.g. the neck) in support of comparison of two faces in detecting anchor-men.

1.3. THE APPROACHES IN FACE RECOGNITION

In general the solution to face recognition problem is the identification of a specific individual, rather than just detecting the presence of a human face which usually is called face detection. The term “face recognition” refers mainly to the following issues:

- Face Identification: Given a picture of a face, an automatic system must decide which person among a set of people the picture represents and, eventually, if any.
- Face Verification: Given two pictures, each of which containing a face, an automatic system has to decide whether the two people pictured represent the same individual (e.g., verify that the person pictured in one image is the same as the person pictured in the other).

There are two general approaches for applying automated face recognition applications:

- constrained;
- unconstrained.

In constrained face recognition the images are captured in controlled environments where the subjects are collaborative and where they know that are being photographed. Instead, in unconstrained case, the face recognition involves matching images that were captured in an uncontrolled environment without the explicit collaboration or knowledge of the subjects.

The constrained face recognition procedures are typically used by organizations such as the Department of Motor Vehicles where driver's licenses that requires photos or by the U.S. Department of State for passports and visa documents. Till now the research has focused on constrained approach because the face recognition accuracy improves a lot when the facial pose, facial expression and image illumination are perfectly set. This type of experimentation has had many initiatives that have had successful outcomes. For example the U.S. Passport and International Travel in USA provides the following requisites as photos composition template for the documents:

- photo must be 2 inches by 2 inches;
- the height of the head (top of hair to bottom of chin) should measure 1 inch to 1 3/8 inches (25 mm - 35 mm);
- make sure the eye height is between 1 1/8 inches to 1 3/8 inches (28 mm – 35 mm) from the bottom of the photo.

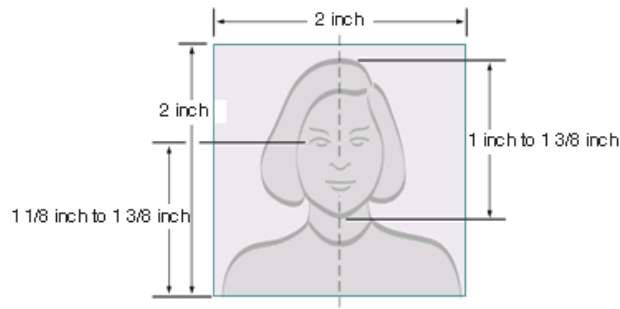


Figure 2 - Photo Composition Template U.S. Passport and International Travel in US

Unconstrained face recognition, where there is no expectation for subject cooperation, can be considered as the next generation of face recognition technology. Unconstrained face recognition captures an image, often with a surveillance camera or in a TV show such as in a live environment as at a sporting event, at the airport or at a political event. The application of this paradigm is challenging by the fact that ideal facial images are extremely hard to capture when subjects are moving in a crowd. Despite the fact that surveillance video or TV scene can be able to capture hundreds of frames of a subject's face, there may be only few frames that are surely usable. In this work were carried out some experiments also with scenes containing several people at the same time.

1.3.1. UNCONSTRAINED FACE RECOGNITION

An important instance in face recognition is the experimental setup both of the gallery and the probe sets. The training is performed with galleries by using n_i images each of N subjects and at test time, when an image is submitted to the matcher, the task is to determine which of the subjects in the gallery corresponds to the probe image according a set of rules (Figure 3).

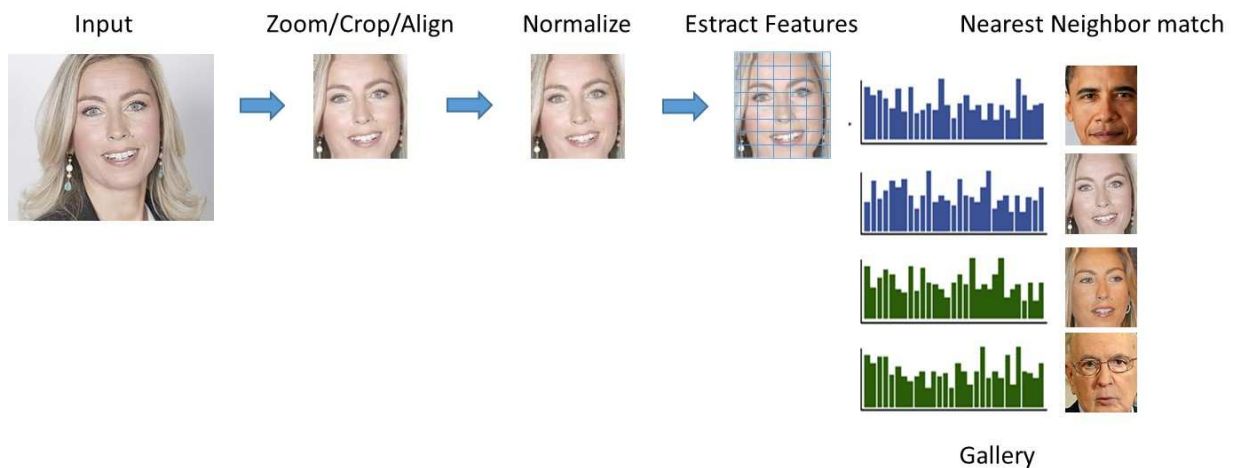


Figure 3 - Face recognition typical pipeline

The limitations of this formulation of object recognition are the following two assumptions:

- exist only a fixed number of object classes known at training time and for each class all samples must be provided at training time;
- in many cases the whole recognition process must be re-trained for every set of identities each time there is a new sample to add and all the training samples must be processed again.

In order to overcome these limitations of the training process, the target of the recognition can be simplified and, given two images, in some cases it can be reduced to the problem of determining whether the images are of the same object class (matched pair) or not (mismatched pair). An arbitrary object can be considered belonging to a set of classes and associated to one image or to another due to factors such as viewpoint, background and occlusions.

In addition some images, used as probe, may be of classes not included in the training set so it is necessary to set correctly the approach in consideration of which an image can be associated to an element of the gallery or discarded. The large amount of intra-class variability makes the problem of visual identification of never seen objects particularly difficult.

As the object recognition, the research that has originated the face recognition, two issues that have arisen are:

- What happens when the scale recognition as well as the number of classes increases?
- How to generalize in order to include new categories and quickly upgrade the training from a small number of examples?

Additionally one of the core difficulties in object recognition is the large amount of intra-class variation in appearance due to factors such as lighting, background, and perspective projection of the 3D objects.

Solving the face identification problem requires addressing each of these issues. The design of a recognition framework requires management of many parameters from the identities not included in the training up to the scaling problems when a gallery grows a lot in terms of magnitude. This is a particularly difficult issue in face recognition where faces share very similar structure and where the variation among the sets of categories can be small due to factors such as head pose, background, occlusion and facial expression.

The application of face recognition has now good performances in case of use constrained galleries but an important progress should be made on the unconstrained face verification task that will also have wider applicability in improving general object recognition.

Usually the face processing is approached by using supervised learning in order to infer a possible identity from a labelled training dataset. These data consist of a set of reference samples where each element is a *pair* consisting of an input object, usually a vector, and an output value. For face recognition, the supervision is in the form labelled face images with the identity of the person in the image, or pairs of face images that are labelled as two images of the same person or two images of two different persons. In face recognition the labelled data are often in the form of face images labelled with pose or the location of facial features such as corners of eyes, nose and mouth or training images containing specific facial features. The obtaining of the labelling of data usually is a manually intensive work instead, on the contrary, it requires rather less effort to obtain many unlabelled face images without identity or pose information. For instance, such images could be obtained by running a face detector over many images and tuning the detector in order to produce a low number of false positives (e.g., high precision, low recall). The images can be addressed as unlabelled faces or partially labelled data, as they have been identified as face images but have no other annotations.

In the context of this work, the target is on making use labelled images of characters acquired in unconstrained conditions by making the downloading of them from Web and using them to identify unlabelled images captured in multimedia videos. The potential application of unconstrained faces as gallery sets and the large availability of images on Web can produce

a more intensive application of face recognition techniques also in workflows where today only the humans are able to carry out efficiently the task of identification.

1.4. GALLERY SETS FOR FACE RECOGNITION

In the years, many face databases have been created under different controlled conditions in order to study specific parameters of the face recognition problem such as, for example, variables as position, pose, lighting and background texture and camera quality. There are a lot of applications for face recognition where the target is the control of a specific set of parameters of acquired image while there are many other where the image searcher has little or no control over the database parameters.

There are several databases available to researchers for face recognition, which are designed for specific purpose and range in size scope and format. The images in many of them were acquired in order to evaluate algorithms or to study specific features face. The acquisition of a face database over a specific period of time and on particular population has advantages, in some areas of research, to give the experimenter the direct control over the parameters of variability in the database to study as, for example, the demographic diversity.

In order to approach more general face recognition problems, in which faces are drawn from a very broad distribution, one may wish to train and test face recognition algorithms on highly diverse sets of faces. Nevertheless it is possible to manipulate a large number of variables in a lab in order to set up a database. There are two problems in this approach:

- the processing requires a lot of effort;
- the settings of the distributions of various parameters must be correctly calibrated in order to make more suitable the database.

The main questions in settings the database can be: How the resolution effects on the quality of image? How many subjects wear glasses or sunglasses? What percentage have beards or moustaches? How many are smiling? How many backgrounds contain cars, signals, sea, people or crowd? The pictures of the same character in the gallery have all the same characteristics?

One possible solution to this problem can be the fulfilment of a collection of pictures of the same character in different live situations in order to have a “natural” distribution of faces. Surely the canonical distribution of faces captured using the constrained approach does not

reproduce a natural distribution of faces that can be useful in different possible application domains and, such database can be an important tool in studying the unconstrained pair matching problem.

While some databases, such as the Caltech [17], present highly diverse image sets, many of these databases are not designed for face recognition but originally rather for face detection. These types of databases were constructed to empirically evaluate recognition algorithms on specific domains. Furthermore the experimental results are reported using different test sets and the application of simple comparing methods can lead, in some cases, to not clear results because there are still a few issues that need to be carefully considered in performance evaluation even when the methods use the same test set. One issue is that each team of researchers apply different interpretations of what a “successful detection” is and another one is that sometimes different training sets are used, particularly, for appearance-based methods.

The Face Recognition Technology (FERET) program database is a large database of facial images, divided into development and sequestered portions and is the “de facto” a standard evaluation methodology [18]. The FERET program begun in 1993 up to 1998 and has been made available to researchers for testing face recognition algorithms. It was sponsored by the Department of Defense's Counterdrug Technology Development Program through the Defense Advanced Research Products Agency (DARPA). The FERET database consists of images taken in different frontal views and in left and right profiles and its program has been founded on the following guidelines:

- sponsoring of research on facial recognition;
- collection and distribution of the database that consists of 14.126 faces images of 1199 individuals;
- comparative evaluation of different algorithms for the face recognition.

Currently FERET is maintained by National Institute of Standards and Technology in USA. The final version consists of 14051 eight-bit grayscale images of human heads with views ranging from frontal to left and right profiles. There is also a colour version (Figure 4).



Figure 4 – Sample of colour version of FERET

The FERET database has been widely used to assess the strengths and weaknesses of different face recognition approaches, before the start of this program there was no set of images available to accurately evaluate or compare facial recognition algorithms. The FERET database made it possible for researchers to develop algorithms on a common database of constrained faces and to report results in the literature using it. The final part of the FERET program was the definition evaluations conditions that allow the comparison of the abilities of facial recognition algorithms using the FERET database. Each face image in the database has a size 256×384 pixels.

Other two databases which are usually used in the face recognition are UMIST and YALE. The UMIST Face Database consists of 564 images of 20 people. Each covers a range of poses from profile to frontal views. Subjects cover a range of race/sex/appearance [20]. The images of each subject cover a range of poses from right profile to frontal views and have a resolution of 220x220 pixels in all in PGM format, in 256 shades of grey (Figure 5).



Figure 5 - Sample of UMIST database

The Yale Face Database (size 6.4MB) contains 165 grayscale images in GIF format of 15 individuals. The database contains 11 images per subject one per different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised and wink. Each face image in the database has a size 640x480 pixels (Figure 6).



Figure 6 - Sample of Yale Face Database

The face database from AT&T Cambridge Laboratories (formerly known as the Olivetti Research Laboratory database - ORL), contains a set of face images taken between April 1992 and April 1994 at the lab and consists of 10 different images for 40 distinct subjects [21]. The images were taken at different times, varying the lighting, facial expressions, and facial details (glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position with tolerance for some side movements. Each face image in the database has a size 112x92 pixels (Figure 7).



Figure 7 - Sample of AT&T database

The XM2VTSDB multimodal database from the European ACTS projects was developed for access control experiments using multimodal inputs. It contains four recordings of 295 subjects taken over a period of four months. Each recording contains a speaking head shot and a rotating head shot. Sets of data taken from this database are available including high quality colour images, 32 KHz 16-bit sound files, video sequences and a 3d Model [22].

The BANCA (Biometric Access Control for Networked and e-commerce Applications) database is a large multi-modal database developed for training and testing multi-modal verification systems [23]. The BANCA database was captured in four European languages in two modalities (face and voice). The subjects were recorded in three different scenarios, controlled, degraded and adverse over 12 different sessions spanning three months. High quality digital camera was used to record the images for the controlled and adverse

conditions and the images of the degraded condition were taken with a low quality web cam. In total 208 people were captured, half men and half women. All the images are stored in colour PNG at resolution 720x576 pixels (Figure 8).



Figure 8 - Samples of BANCA database

The AR database was collected at the Computer Vision Center in Barcelona and contains over 3,276 colour images of 126 people (70 males and 56 females) in frontal view [24]. This database is designed for face recognition experiments under several mixing factors, such as facial expressions, illumination conditions, and occlusions. All the faces appear with different facial expression (neutral, smile, anger, and scream), illumination (left light source, right light source, and sources from both sides), and occlusion (wearing sunglasses or scarf). The images were taken during two sessions separated by two weeks. All the images were taken by the same camera setup under tightly controlled conditions of illumination and pose. The resulting RGB colour images are 768 × 576 pixels in size.

Usually in these databases all faces are taken from a straight-on frontal pose, with facial features such as eyes in the same position within the image, neutral facial expression, similar lighting condition, and lack of any occluding objects such as hat wear or glasses. This lack of variation from factors such as pose, lighting, expression and background characterizes many of the standard data sets traditionally used to study face recognition. The implicit assumption made for all these data sets is that the control over the image acquisition process is fully supervised. The control of image acquisition allows to mitigate problems coming from aspects such as lighting and background and to instruct the person being photographed to maintain a particular pose and expression. This assumption fits well for some specific applications such as in security domains where the target is the verification the identity of

the person pictured in a passport photo. For many other applications these assumptions no longer holds and the violation of these hypothesis can lead to rapidly degraded performance. While many applications requires the full control of the parameters in acquisition of the images, in order to study the unconstrained case the faces in the database must be acquired from sources with little or no control over parameters as pose o illumination reproducing the conditions usually experienced by people in everyday life. These databases have as basic requirement the exhibition of “natural” variability such as pose, lighting, focus, basic resolution, facial expression, age, gender, race, accessories, make-up, occlusions, background and mainly the photographic quality.

One important database is Face Recognition Grand Challenge (FRGC) that, with the participation also of NIST, ran from May 2004 to March 2006 [25]. The main target of FRGC was the innovation in face recognition in order to support the efforts of the U.S. Government in this field. Within the FRGC program, in addition to database, were developed also new face recognition techniques and prototype systems to increase the performances involving researchers in companies, academia and research institutions. The data in FRGC database consists of 50,000 recordings divided in training and validation subsets. The training set is designed for training algorithms instead the validation partition is for the assessment of performances laboratory setting and consists of data from 4,003 subject sessions (Figure 9).

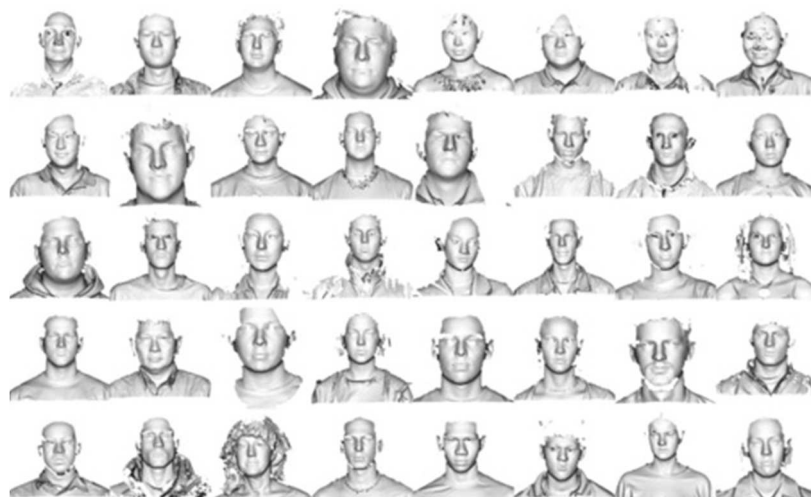


Figure 9 - FRGC sample (from www.clementcreusot.com/phd/)

Recently some other public domain databases have been designed to study unconstrained face recognition exist such as the Labeled Faces in the Wild (LFW) [26], BioID [27] or YouTube Faces (YTF) [28]. These have been published mainly in order to challenge computer vision researchers with large-scale unconstrained face recognition in still images and videos.

LFW, in particular, provides images galleries under different training paradigms such as image-restricted and unrestricted. In image-restricted setting only binary "matched" or "mismatched" labels are given for each couple of images. Instead, under the unrestricted setting, the identity information of the person appearing in each image is also available allowing for, potentially, additional image pairs. The main contribution of LFW is the provision of an important set of unconstrained face images with a large range of the variation in many parameters. The data sets contain more than 13,000 images of faces collected from Web. Each face is labelled with the name of the person and, among them, in 1680 cases the people pictured have two or more different photos (Figure 10). All the faces contained in the database were detected using the Viola-Jones face detector.



Figure 10 - Sample of LFW database

The BioID Face Database has been recorded and is published in order to provide to researchers the capability to compare the quality of their face detection algorithms. This database contains 1521 grey level images with a resolution of 384x286 pixels (Figure 11). Each image shows a frontal view of the face of one out of 23 different test persons. In BioID the captured images are in realistic context with significant variability in pose, lighting, and expression. Also the backgrounds include realistic scenarios such as office or home settings.



Figure 11 - Sample of BioID database

This table shows a list non exhaustive of the face main databases for face recognition available for the researchers [26].

Database	# of people	Total images	Highlights
AR Face Database, Purdue University, USA	126	4000	Frontal pose, expression, illumination, occlusions, eye glasses, scarves
AT&T Database (formerly ORL Database)	40	400	Variation of time, lighting, facial expression, eye glasses
BioID Face Database	23	1521	Real world conditions, grey scale, background, lighting, expression, eye positions given
Caltech Faces	27	450	Lighting, expression, background
Caltech 10000 Web Faces	≈10000	10000	Wide variability, facial features annotated
CAS-PEAL Face Database	1040	99594	Very large, expression, accessories, lighting, simultaneous capture of multiple poses, Chinese
Cohn-Kanade AU-Coded Facial Expression Database	100	500 sequences	Dynamic sequences of facial expressions
EQUINOX HID Face Database	N/A	N/A	Non-visible light modalities
Face Video Database of the Max Planck Institute for Biological Cybernetics	N/A	246 video sequences	6 simultaneous viewpoints, carefully synchronized, video data
Facial Actions and Expressions	24	≈7000	Expression, colour, grayscale
Face Recognition Grand Challenge Databases	>466	>50000 images and 3D scans	Vary large, lighting, expression, background, 3D, sequences
FERET Database, Colour	1199	14126	Colour images, changes in appearance through time, controlled pose variation, facial expression
Georgia Tech Face Database	50	750	Expression, illumination, scale, orientation
Indian Face Database	40	>440	Frontal, Indian subjects

Japanese Female Facial Expression (JAFFE) Database	10	213	Rated for emotional content, female, Japanese
MIT-CBCL Face Recognition Database	10	>2000	Synthetic images from 3D models, illumination, pose, background
M2VTS Multimodel Face Database (Release 1.00)	37	185	Large pose changes, speaking subjects, eye glasses, time change
M2VTS, Extended, Univ. Of Surrey, UK	295	1180 videos	Rotating head, speaking subjects, 3D models, high quality images
NIST Mugshot ID	1573	3248	Front and side views
NLPR Face Database	≈22	450	Lighting, expression, backgrounds
PIE database CMU	68	41368	Very large database, pose, illumination, expression
Psychological Image Collections at Stirling (PICS)	N/A	N/A	Targeted at psychology experiments
UCD Colour Face Image Database for Face Detection	≈299	299	Targeted at detection applications, highly varied , colour
UMIST Face Database	20	564	Pose, gender, race, grayscale
University of Essex, UK	395	7900	Racial diversity, eye glasses, beards, college age
University of Oulu Physics-Based Face Database	125	>2000	Highly varied illumination, eye glasses
VALID Database	106	530	Highly variable office conditions
WidTIMIT Database	43	Multiple videos per person	Video, audio, reading, head rotation
Yale Face Database	15	165	Expressions, eye glasses, lighting
Yale Face Database B	10	5760	Pose, illumination

Table 1 - Main databases for face recognition

1.5. GALLERY SETS WITH IMAGES FROM WEB

The continuous growth of large-scale image archives on the Web makes available a virtually unlimited supply of images that can have an important impact on multimedia research and mainly on face recognition. A simple Web search allows to end user to select among innumerable Web images on any topic, from celebrity images including portraits, posters, movies, snapshots and news images up to the images of ordinary persons. The research of

images can be performed using several parameters such as resolution, colour, type, time or even proprietary rights. The set of possible results can be valued potentially infinite for the end user that is not able to examine all the outcomes of the research in a reasonable time. The search engines propose to the user on the top of the results page what, at the moment of the research, appear as the most relevant regarding the key inserted in the search field and according the algorithms and the criteria adopted to define the ranking of the key itself. Usually, in fact, the queries related to celebrities on well know characters constantly rank the highest positions among all the image queries revealing the intensive user interest for this type of images. These images provide an opportunity in order to create large scale training datasets in order to innovate the field of face recognition and machine learning research.

The collection and labelling of the faces of characters from generic images searched on Web is a challenging task because the results are collected using mainly the criteria of relevance of each specific search engine. These images are retrieved by means tags and indexing algorithms that assign relevance for each search key. The surrounding text of a Web image or the contents of the robots.txt file often comprises words and phrases far from a standard grammar structure or that are related a specific contexts non in connection with the character in the figure. Therefore often it is difficult to apply natural language processing techniques to extract celebrity names and estimate the likelihood of a celebrity appearing in the image. Furthermore the face of famous characters when are published on the Web exhibit large visual variations among them due to pose, makeup, expression and occlusion caused by sunglasses or fancy hairstyles or even caused by the constrains imposed by the tools for Web publishing. Frequently the same picture is published with different size in terms of pixels or using specific file format such as png instead of jpg. In other cases the tag describing the image can refer to a characters related to another character not shown in the picture or to the same in another age of life. All these problems are only the a part of the more general problem of the information retrieval in a case where the “information noise” introduces great difficulty in association of names with faces for visual analysis and where there is large uncertainty on the quality of the captions for general Web images. The information retrieval systems have always had to deal with the problem that keywords are a limited capability to express correctly an information meaning. The automatic information retrieval systems start in the 1960s when were designed to search repositories of newspaper

articles, scientific papers, patents, legal abstracts and other document collections in response to keyword queries. Often the list of keywords can be short and inexpressive and can suffer of the problem of synonymy, that is the opportunity to refer to the same thing in multiple ways, as well as the polysemy that is the semantic relation for which can be found multiple meanings for the same term such as, for example, the words jaguar or tiger that can produce as results images primarily about automobiles or computers as Apple and only later animals. For a long time the information retrieval was the area of interest of librarians, patent attorneys and some other people whose jobs consisted of searching in collections of documents. In these scenarios people were trained in how to formulate effective queries, and the documents they were searching tended to be written by professionals, using a controlled style and vocabulary. With the advent of the Web, where everyone can be at the same time author and searcher, the problems of the information retrieval is improved in scale and complexity with Internet and the search of images has followed this trend. In perspective, the Web can be considered the large repository of images able to provide a continuous updated version of the pictures that can be processed and adapted for face recognition purpose.

1.6. THE CONCEPT OF VISUAL INTERNET OF THINGS

The capability to use images to retrieve information from other information contents is an important feature that will improve its importance with the evolution of the Web and the widespread availability of cameras, sensors and social networks and that has affinity with the concept of IoT. The IoT refers to uniquely identifiable objects (things) and their virtual representations in an Internet-like structure that is the idea of a network world connected by things which consist of data, connections and information processing. The Visual IoT (VIoT) is a concept derived from the IoT and that is focused on the combination of the visual and image processing technologies with IoT itself [29].

The IoT addresses the current evolutive trend of the information technology and identifies a world of intercommunicating “things”. Most of the Internet of things systems utilize the RFID or other non-contact wireless technology as their sensor achieving successes in many applications. Although there are many factors in favour of the use of RFID, these labels must

be attached on every object for recognition and this solution can not be implemented in some situations.

The VIoT is proposed to improve the scenario foresees by IoT and to provide a visual method to access the object labels. The IoT refers to the “physical” identifiable objects and their virtual representations in an Internet-like structure [29] addressing the concept of network world connected by things, which consists of sense, connection and intelligent information processing. As a special kind of IoT, VIoT focus on combination of the visual related technologies to traditional IoT, which consisted of cameras, information transmission net and intelligent image or video analysis algorithms. With the help of visual cameras, the VIoT can get the object meaning via image processing of the scene, attach it a visual label to the object and, then, return the label itself to the information network. Compared with IoT, the biggest advantage of VIoT is its ability to extract labels from the appearance of objects without the need of RFID or other explicit labels attached on the objects.

The data exchanged by this aggregate of "connected objects" can build environments able to have the awareness of the context in dialogue both among themselves and with humans. In a recent report, the consulting firm Gartner has predicted that in 2020 the "intelligent" objects will be 26 billion [31] and that they will be not only real but also virtual instances of real objects or, in addition, information elements derived or extracted from other information or content. The instance of an object on the Web will be an abstract element and will represent the connection point that will enable producers and consumers of information to remain hooked to the object itself by enabling both the manipulation that the development of new services.

Based on biometrics, computer vision and pattern recognition, face recognition is one of the most appropriate applications of VIoT going beyond the current applications where all face recognition technologies are only tools, as visual sensors like cameras or videos, while face feature are an inherent label of human beings. However, it is not easy to put face recognition on VIoT framework or make it run well, since many face recognition technologies could only work well under constrained situations like low scale, tiny illumination discrimination and approximately frontal posture.

CHAPTER 2

2. FACE RECOGNITION PROBLEM

The first phase in face recognition is to understand how set the process so the target and the problem boundaries are well identified. In order to reach this target it is important to realize that face recognition belongs to a class of problems in the pattern recognition field. Therefore the first level is to recognize or, in a more formal way, to classify an object according to certain rules and constraints. In the case of face recognition the main objective of course is to classify human faces. Another important aspect in face processing is to distinguish two types of procedures:

- identification;
- verification.

For the identification the system aims to identify a given input through a comparison process of known individuals taken from a database [32], instead the main goal of a system of recognition is to confirm or deny a claimed identity.

Usually the verification case is a simplified version of face recognition problem, often the processing time is reduced and the above all database is acquired in constrained conditions. The identification problem, on the other hand, requires a much larger database and in some cases a long processing time when used to recognize a great number of people or a character in a large dataset.

Although both processes are different in many aspects they have similarities in terms of technique and algorithms that can be solved the same way as problems in pattern recognition domain. Consequently is important understanding the two main concepts related to pattern recognition that are:

- features;
- classifiers.

Features are the measurements extracted using some algorithm from an object and used to make inference on particular characteristics or properties so it is possible proceed to a classification in

order to distinguish different types of objects. In the general case of pattern recognition, the measurements used for the classification could be, for example, the mean value and the standard deviation. When a set of features are organized together they comprise a "feature vector" [33]. In mathematical terms, the feature vectors laying within a space which delimits the problem boundary identifies uniquely a single pattern (object). Adopting this approach the problem of space can be referred as the object feature space because it contains all the particular characteristics for the object under evaluation.

The classifiers can be defined as processes deployed according definite rules and constraints in order to divide the object feature space into decision regions. These sets of all regions represent all the classes that the classifier is able to identify. As a result of this approach a classifier for face identification must be able to create an abstract decision line in the feature space which separates those faces resembled to the model being identified from other models in the database of faces. The decision line can be defined as a decision threshold and constitutes the classifier whose role is the division of the feature space into regions each of them corresponding to a class, if is a feature vector, matching an unknown pattern. If the result of a test falls in the region of a class, the classifier attributes the pattern to the class itself despite this not necessarily entails that the decision is correct and, when a wrong classification occurs, a misclassification follows.

2.1. PATTERN RECOGNITION PROCESS

The face recognition belongs to a specific class of problems in the pattern recognition field. The pattern recognition can be defined as the assignment of patterns within a number of categories or classes following the extraction of significant features from a selected area of an image. According the "Pattern Recognition" book [33] a typical pattern recognition process is composed by five basic stages to which was added the step of editing as shown in Figure 12. If an unknown sample image is provided in input to a pattern recognition system and there is a set of predefined classes at the output, the task of the process is the assignment of it to the class the with acceptable classification error rate.

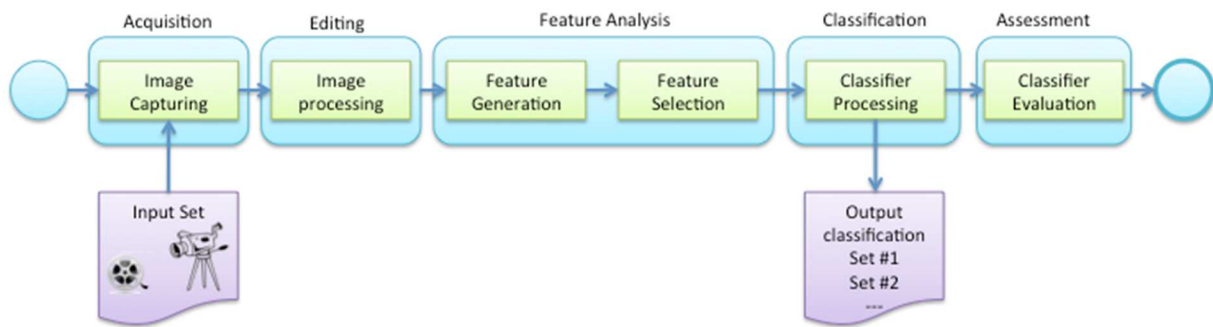


Figure 12 - The basic stages involved in the design of a classification system

2.1.1. ACQUISITION

This step involves the devices, camera or the devices playing videos, able to capture the images. The performances of precision the facial recognition systems collide often with the quality of the acquired images because the resolution of each graphic element within a frame in many cases is not sufficient to produce suitable results. In this step, the automatic pattern recognition systems could also perform an assessment of the quality of the acquired images. For example, in interactive acquisition, the portrait could be acquired more times in order to obtain an image that meets the constraints of the process of recognition.

2.1.2. EDITING

This phase usually comprises processing operations on the image aiming at improve the representation of the patterns. It can include filtering, noise removal, segmentation or image normalization according the nature of pattern recognition task. For example, in face recognition, the images can be registered or live so, as outcome of this phase, the processor has to make sure that a face appear in the image. Additionally the pattern samples can contain some “noise” that can be reduced before classification phase where the term “noise” is usually referred to any property of the pattern that can deceive a pattern recognition system.

In addition, before comparison of patterns, it could be necessary firstly to detect specific elements to recognize and to perform some adjustment such as the adaptation to the same scale of the test images in the classifier or to perform rotations or zooming in order to put in the correct position the object such as the head positioned vertically. This step is quite simple when working on controlled portraits, because each image only contains one face, but can become very difficult

when the task is the extraction of a multitude of faces from a video coming from a PTZ camera or from a video on broadcast TV.

Moreover some recognition tasks can require also the segmentation of the content or segmentation in sub-regions in order, for instance, to segment faces in an image to create meaningful patterns for the feature extraction step.

Finally the normalization is the operation to scale the features of data to fit within a specified range in order to provide a representation of patterns useful for the next steps of the process.

2.1.3. FEATURE ANALYSIS

The pattern recognition analysis requires the application of algorithms to apply mathematical transformations in order to compare images. These transformations are aimed at highlighting the characteristic features of an image such as frequencies, directions or contours. The feature generation performs an automatic or manual selection of all features which describe the object. These features are extracted and selected in order to allow the reduction of the number of the features of patterns and preserving as much as possible the other important information. This step is important because the computational resources are limited and their selection can save the characteristic elements discarding all the redundant and useless features.

2.1.4. CLASSIFICATION

The task of the classification is to compute the accuracy level of the whole process by assigning the feature vector provided to a class. The output of the classifier is typically a discrete selection among the set of available classes. All the preparation phases of the pattern recognition process are designed and tuned for improving the performance of the classifier and the degree of difficulty of the classification depends by the similarity relations between the patterns of different classes.

2.1.5. ASSESSMENT

The assessment evaluations aim to improve overall classification accuracy trying to minimize the classification error rate based on basis of the classification outputs. This phase is addressed to

measure the performances of the entire process and by means of the calculation of metrics and verifying the accuracy of the classifier.

2.2. BASIC CONCEPTS OF PATTERN RECOGNITION

Learning models and training process are two important elements of pattern recognition that are essential in order to define the approach to set the recognition model and the training mode.

2.2.1. LEARNING MODELS

A pattern recognition system is a mathematical model able to map the incoming patterns into the most suitable classes. Usually is not possible to establish a reliable mapping without the availability of data samples. Learning is any method able to incorporate information from a training set samples in a predefined pattern classification. There are three basic types of learning methods depending mainly by the nature of the pattern recognition task.

- **Supervised Learning:** In supervised learning the class labels or costs of training set samples are known by the classifier before that the training phase begins. The training phase computes the model which minimizes the total cost for the training set patterns. This kind of learning involves human effort, is typically the most used learning method, and has many applications in pattern recognition area.
- **Unsupervised Learning:** In unsupervised learning the training set images are not labelled and it is used to draw inferences from datasets consisting of input data without labelled responses. Unsupervised learning does not require human effort for labelling and has many applications in engineering, such as image segmentation and multi-spectral remote sensing.
- **Reinforcement Learning:** In reinforcement learning an agent learns from the consequences of its actions, rather than from being explicitly taught and it selects its actions on basis of its past experiences (exploitation) and also by new choices (exploration). The reinforcement signal that the classifier receives is a numerical reward which encodes the success of an action.

2.2.2. TRAINING

In order to evaluate the class that better fits the pattern samples to their corresponding classifiers, the recognition system must be trained by using an available training set samples. The performance of a pattern recognition system refers to its capability to associate an image with an object or even to recognize new samples not catalogued during the learning stage. However a pattern recognition system, which is usually trained in order to maximize the performance according a predefined training set samples, is not able not recognize new test samples without a new training phase and In pattern recognition systems there are two main causes that cause problems performances:

- the pattern recognition system can require accurate plan of the sessions of training;
- the number of features is too large against to the number of training samples and this is also called the curse of dimensionality.

2.3. FEATURE EXTRACTION METHODS

The selection of algorithm to perform the feature extraction is important in order to achieve an accurate classification. Feature extraction consists in identifying the most discriminative variables for data classification that are associated with the most relevant elements in the data distribution. Feature extraction is generally used for reducing the dimensionality of facial images making some linear or nonlinear transformation of the data with consequent feature selection so that each extracted feature is as representative as possible. Features can be observed under different points of view, what is not obviously present in one domain can become obvious in another domain and the images can be represented by using the original spatial representation or by means of frequency domain coefficients.

In pattern recognition the target of statistical classification is to use an object's features and to identify which class it belongs to. This target can be reached using a different approaches and a classifier can be based on the value of a linear or non-linear combination of the features. A linear model for classification separates input vectors into classes using linear decision boundaries.

The family of multivariate analysis methods for the feature extraction is usually used in order to reduce the data dimensionality by projecting the points towards the most relevant directions. For example in Principal Component Analysis (PCA) [34], that is one of the most used linear feature extraction methods in image processing, when the features and the target variables are

distributed nonlinearly, the related linear methods are not able to describe correctly the data distribution. In order to face this problem, different non-linear versions of PCA has been developed, which can address non-linear problems either by local approaches [35], neural networks [36] or kernel-based algorithms [37]. In particular the kernel methods project the input data to a high dimensional Hilbert space and define a linear method. The model is nonlinear with respect the input space and computes the nonlinear relations between data via a kernel (similarity) function implicitly. Kernel methods have in general good performance in the case of high dimensional problems and low number of training examples.

2.4. TECHNIQUES AND ALGORITHMS FOR FACE DETECTION

The face recognition is a particular case of the general problem of pattern recognition where the target is to achieve the best features with minimum classification error and with the lowest running time using sets of faces. In unconstrained cases the extraction of features [38] can become very complicated because the subjects in the images are captured in live situations without any rule and the available features can be so small that the processing algorithms can be not able to retain all particulars needed to have a clear representation of them. There are a lot of methods to make the processing of images in order to extract the features and it is not easy to compile a taxonomy because there is not a globally accepted grouping criteria. The extraction and classification criteria differentiate themselves between different scenarios that depend on the specific application.

Among the possible classification methods, a basic classification can be the following:

- holistic;
- feature-based (structural);
- hybrid.

2.4.1. HOLISTIC APPROACH

In holistic approach, also called global, the faces are taken as units and the extracted features pertain to the whole face. The best models that are nearest to holistic methods are Eigenfaces [39], Principal Component Analysis and Linear Discriminant Analysis (LDA).

These approaches are computationally expensive and suffer of problems such as sensitivity to face orientation, size, variable lighting conditions, background clutter and noise. Typically the holistic approaches contain transformations to subspaces in order to reduce the dimensionality which is important when dealing with large databases. One of the main problems in performance of these direct matching methods is that they attempt to perform classification in a space of very high dimensionality. In order to mitigate the curse of dimensionality, several other schemes have been proposed that employ statistical dimensionality reduction methods to obtain and retain the most meaningful feature dimensions before performing recognition. One of them was proposed by Sirovich and Kirby [40] that propose to use PCA [41][42] to represent the face images. They demonstrated that any particular face can be represented using the eigenvalues coordinate space and that any face can be approximately reconstructed by using a collection of eigenpictures and the corresponding projections, which are the coefficients, along each eigenpicture itself. On the basis of the work of Sirovich and Kirby, Turk and Pentland [39, 40] recognized that projections along eigenpictures could be used as classification features to identify faces. The significant features are, with this approach, the "eigenfaces," because they are the eigenvectors (principal components) of the set of faces and they do not necessarily correspond to features such as eyes, ears, and noses. The projection operation characterizes an individual face by a weighted sum of the eigenface features, and so to recognize a particular face.

Further enhancements have been proposed such by other authors as Sharif that introduced the appearance-based methods for recognition purpose in order to handle the problem of illumination as factor to enhance the performance of face recognition [43].

Among the subspace reduction, methods as Linear Discriminate Analysis LDA have been applied for face recognition under different light and pose conditions. Although LDA has good performance in detection of the features it has problems under Small Sample Size (SSS) problem [44]. Additional enhancements have been introduced using evolution of the classical PCA algorithm with the Kernel version (KPCA) [45] and, similarly, in LDA with LDA (KLDA) [46] where the kernel functions take as input the vectors in the original space and return the dot product of the vectors in the feature space. Figure 13 shows the eigenspaces produces using PCA and LDA.

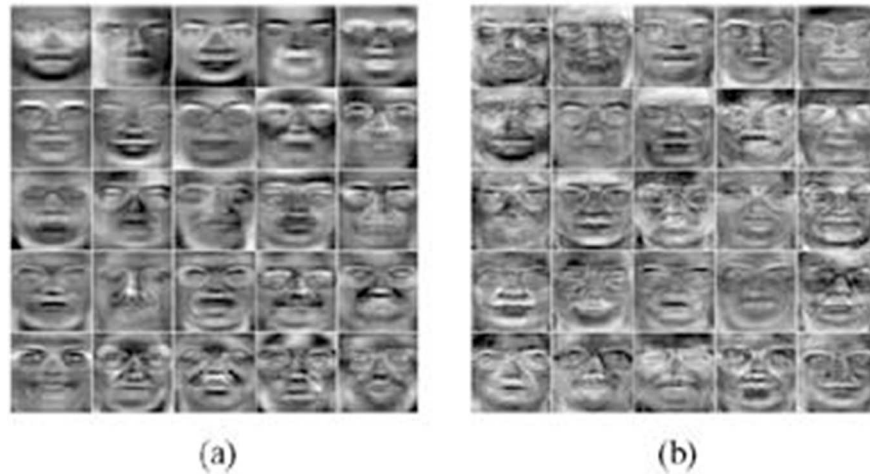


Figure 13 - Holistic basis images from a) Principal Component Analysis b) Linear Discriminant Analysis

Another interesting property of PCA is that it allows the conversion of 2D face images into 1D feature vector where each row of the 2D image matches a row in the feature vector. The spatial information of the face image can be lost during the conversion and a Small Sample Size problem can appear. In order to avoid this problem Two Dimensional PCA (2DPCA) can be used and an advantage over 1DPCA is that the feature vector is now two-dimensional so the problem of dimensionality is reduced [47]. With 2DPCA the co-variance matrix can be constructed via 2D image vector but more coefficients are required in the representation in the 2DPCA subspace and, due to linearity of 2DPCA, higher order vectors [48].

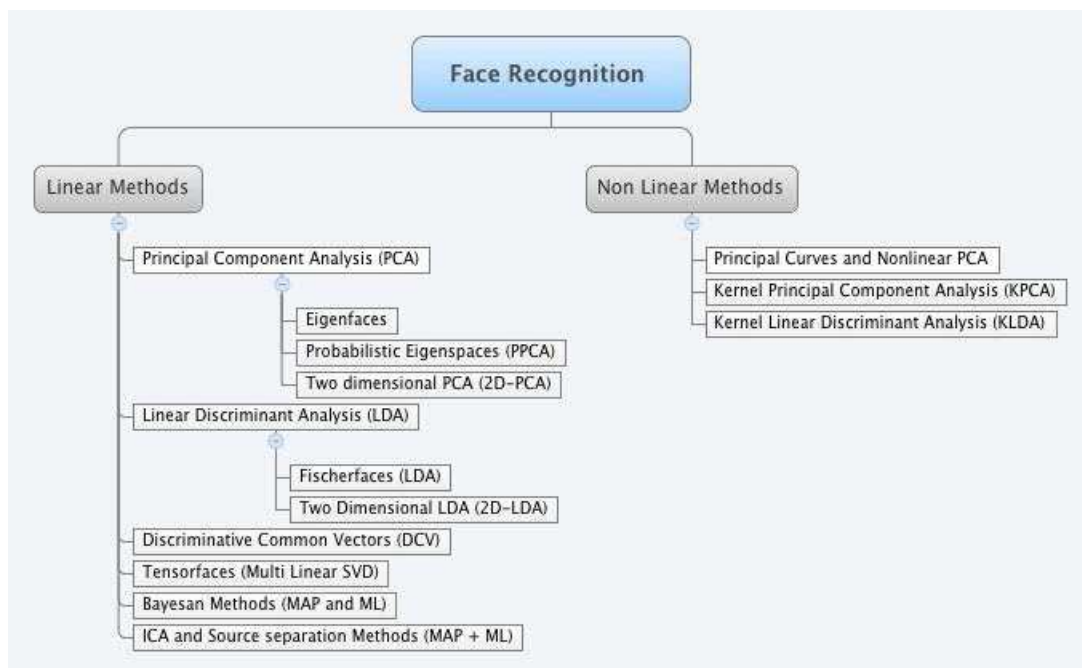


Figure 14 - Taxonomy of the face recognition methods based on PCA

The main advantage of the holistic approaches is that they perform the recognition using the whole face and assume that all the pixels in the image are equally important [49]. These techniques are not only computationally expensive but require a high degree of correlation between the test and training images, and do not perform effectively under large variations in pose, scale and illumination, etc.[50].

2.4.2. FEATURE BASED METHODS

Feature-based approaches are methods based on extraction of local features from the face images. These methods exhibit robustness to variations caused by expression or pose changes and are less sensitive to variations in illumination. As drawback the face image could not contain enough information for identification purposes and also this requires a precise and reliable feature localization. In feature-based methods the local features such as eyes, nose and mouth must be extracted and their locations and local statistics (geometric and/or appearance) are sent to a classifier in order to identify individuals.

Three basic different extraction methods can be distinguished:

- generic methods based on edges, lines, and curves;
- feature-template-based methods;
- structural matching methods that take into consideration geometrical constraints on the features.

The local features, each representing a single subsection of the image, can be matched with the features stored in the database and, although the images of the same object can be taken in different environmental conditions, they can provide the measurement of the similarity between different subjects. A difference between images can occur due to noise level, change in illumination, scaling, rotation and change in viewing angle. In order to match the images the basic concept is that the local features should be invariant to these differences and the performances of local feature based methods that has to perform the correct detection of local features which are highly distinctive and invariant to different capturing conditions.



Figure 15 - Match results for one of the test images

The application of the feature based methods is similar to the properties of the human recognition where people can identify face from very far distance although the details are ambiguous. That means the symmetry main features characteristic, such as nose or mouth, are enough to be recognized. There are differences in shape, size and structure of those organs, so the faces are differ in hundreds of particulars so one method of recognition can be the extraction of the shape of the biometric element, the eyes for example, and then distinguish the faces by distance and scale. Another method can be the application of a deformable model in order to describe the shape of the organs on face [51].

Examples include the use Gabor features [52], SIFT [53], SURF [54], HOG [55] and histograms of Local Binary Patterns (LBPs) [56]. Using these methods firstly the direction and edges of important component must be detected to build the feature vectors from the edges and direction. For example, the Gabor filter can capture salient visual properties such as orientation selectivity, spatial localization and spatial frequency characteristics. These filters have the form:

$$\psi_{u,v}(z) = \frac{\|k_{u,v}\|^2}{\sigma^2} e^{-\frac{\|k_{u,v}\|^2 \|z\|^2}{2\sigma^2}} \left[e^{ik_{u,v}z} - e^{-\frac{\sigma}{2}} \right]$$

Where u and v define the orientation and scale of the Gabor kernels $z = (x, y)$. The wave vector

is $k_{u,v} = k_v e^{i\phi_u}$ where $k_v = \frac{k_{\max}}{f^v}$ and $\phi_u = \frac{u\pi}{8}$ $\phi_u = \frac{u\pi}{8}$ with k_{\max} the being the maximum

frequency and f being the spacing factor between kernels in the frequency domain.

SIFT algorithm was proposed in 2004 by D. G. Lowe and is a powerful technique for general object detection. SIFT makes the processing of images in four stages. In the first phase it searches over all image locations and, then, for each candidate location it fixes key points on basis of on measures of their stability. In the third step, each key point is assigned with one or more orientations based on local image gradient directions. At last, the image gradient magnitudes and orientations are sampled around the key point location.

The SURF algorithm is similar to SIFT and promises a more compact feature descriptors and faster matching. SURF is invariant to translation, rotation and scaling and partially invariant to illumination and small transformations. Its process foresees routines to make the detection and description of the interest points using the determinant of the approximate Hessian matrix. The integral images are used in Hessian matrix in order to reduce the computation time. The sum of Haar-wavelet responses is used to describe the feature of an interest point. When the dominant orientation is estimated and included in the interest point information, SURF descriptors are constructed by means of the extraction of square areas around interest points.

The HOG descriptor is a method that can be used to describe the local object appearance and shape by means of the distribution of intensity gradients. The HOG descriptor is implemented by the division of the image in small spatial regions (cells) and then by the processing of the histogram of gradient direction or edge orientations. The local histogram of cells can be normalized in order to avoid problems in illumination change. The HOG descriptors operate on localized cells and not suffer of the problems caused by geometric deformation.

The LBP algorithm was introduced mainly for texture description and only afterwards it found application also in face recognition. It is able to evaluate the micro-structures of by analysing the texture of the skin. The original LBP operator identify each pixel of an image with an 8 bit number by using the grey scale of the central pixel in comparison with the neighbours in a Region of 3×3 pixels and the histogram of the generated values is used as a texture descriptor. Improvements of the LBP algorithms were introduced by means of two parameters: radius R and the number of sampling points P .

2.4.1. HYBRID METHODS

The hybrid methods have been developed for face recognition in order to use a combination of both holistic and feature extraction methods described in the previous sections. Usually the 3D images are used in these methods because the face image is addressed to detect the curves of the eye sockets or the shapes of the mouth. In hybrid methods even the profile of the face is useful because these systems use the axis of measurement in order to have enough information to construct a full face. The 3D system usually proceeds following this process: Detection, Position, Measurement, Representation and Matching.

- Detection: capturing of a face by a scanning a photograph or on video;

- Position: determining the location, size and angle of the head;
- Measurement: assigning measurements to each curve of the face to make a template with specific focus on a biometric feature (e.g. the curve of the eye or the angle of the nose);
- Matching: comparing the received data with faces in the existing database.

CHAPTER 3

3. THE AUTOMATIC TAGGING OF FACES IN VIDEO SEQUENCES

One of the main applications of face recognition is surveillance for security purposes or for the multimedia video automated tagging. As example, the recognition in the surveillance applications requires the real-time capturing of faces from an image sequence acquired by a video camera. Automatic video annotation finds application in also in many fields where is necessary to get timely and useful information from a multimedia content and converting it in knowledge. The achievement of a reliable automatic process is still an important open problem in multimedia and computer vision.

The current technology of video search tends to be ineffective and expensive in terms of human effort. In professional field there are a lot of activities that require the automatic video annotation with the identities of the persons tagged for archiving and indexing purposes. For the end users of most video sites, such as YouTube and or Google Video, the simplification of searching in videos on basis of the information associated with the whole videos can modify the service experience. Just think of the case where an elderly or sick person has difficulty in remembering the faces of people you see on TV, these services can improve the quality of life. Even in Web search services such technologies can change the user experience where the engines can be enabled to provide results on the basis of researches performed on images rather than on the sequences of text. Additionally the automatic tagging can also modify the results retrieved by the search engines that could be able to provide the information that often remains hidden allowing a deeper mining. Facial recognition in videos is a complex of techniques used to verify the presence of biometrics within a frame or a sequence of frames and the subsequent association by means of the identity of an individual through the physical features of the face. As explained in the previous chapter, from a technical point of view, the facial recognition is a particular specialization of the problem of the recognition of objects in which the items to be recognized have special properties that join them (e.g. position of eyes, mouth and nose).

Also the local image characteristics (morphology of the eyes, nose), in addition to the global

parameters, are fundamental in perception of face. The quality of the response of the automatic systems depends directly from the limitations that are imposed on the type of input images and search features within the frame. The constraints on the images can include the type of format, the spatial resolution of the image, the brightness and tint that are the elements on which to operate in order to segment the image itself into elementary components. Usually these techniques are applied on still images and not on multimedia video where many parameters such as the resolution of the single frame, for example, can be not suitable to obtain acceptable performances. The recognition of faces in videos in fact is a more challenging problem because it occurs often under unconstrained environment and, only after a long set of processes of pre-elaboration, an image can be ready for the classification with an automatic system.

The automatic face recognition in video is therefore an essential technology for the video tagging of large amounts of multimedia content. The video tagging can be defined as associating of useful information and description of the content of video (frames, scenes or shots) available in an easily usable mode and enabling the user to perform contextual searches.

In order deal with this problem, different video tagging schemes can be designed in order to help in extracting relevant content from a video file and hence support the activities of searching videos with specific content. The design of such schemes can face the problem starting from different scenarios and be applied alternative approaches.

For example a simple mechanism could allow to an end-user to apply tags (association of textual, semantic information with a relevant individual video frame or a scene) to people in a video and then use these tags in order to make available index services. Another mechanism could enable the automatic tagging of other videos starting from the knowledge contained in the element already tagged. Another one can be the association of techniques of speech recognition in order to discover names or frequent words that can be associated with the content. The approaches can be various and different. In the context of this work it was decided to use the Web as repository of knowledge.

In order to approach this problem, the study spans over the following areas:

- Extraction of relevant content (faces of people) from the video data including the content extraction from individual frames, shots and scenes of the video data.
- Schemes for defining and representing the matching between images collected using the Web and the video content.

A video-based face recognition system usually consists of three basic modules:

- detection of the face;
- tracking;
- recognition.

An automated system should allow the scan of the multimedia content choosing a few good frames and then apply one of the recognition techniques over those selected frames in order to identify the individual. As example, the system would allow an end user or a professional to tag relevant face images of the videos making the association of the name as annotation. The system should suggest relevant scenes, frames and shots in the video and then insert a tag.

Afterwards a search mechanism would be easily able to successfully respond to the contextual queries such as “The President Obama” would result in:

- videos which have the related faces with similar functionalities of current search engines;
- videos which contain scenes where is present the face of a character, as President Obama, with relates scenes appropriately marked in the resulting videos such that end users would be able to play exactly the scenes where, for example, “Obama” is present.

The resulting video would be tagged with the names of the people and the contextual information would be associated with each marked person in order to develop a knowledge base able to assist in identification persons.

3.1. FACE RECOGNITION IN VIDEOS

The acquisition of the face can be performed on a sequence of frames on a fixed image, and according to this distinction, the problem of recognition can be divided in static and dynamic. The static recognition is normally characterized by a high image quality in terms of pose, resolution, and lighting. The dynamic recognition is instead characterized by heterogeneity of backgrounds, poses and lighting which is partially offset by the availability of a large number of frames, it occurs in video sequences and is the main topic in this work.

Beyond the basic image-to-image matching of faces, according the definition of NIST [57], a classification in face recognition in videos can be provided as:

- Video-to-Still (V2S);

- Still-to-Video (S2V);
- Video-to-video (V2V).

FIVE (Face in Video Evaluation) [58] is a NIST program dedicated to assess the capability of face recognition algorithms to correctly identify or ignore persons appearing in video sequences. The Still-to-still case (S2S) is not interesting for NIST because the still images do not take advantage of the temporal information present in the videos.

This program aims to determine which algorithms are most effective and whether any are viable for the following operational use-cases:

- high volume screening of persons in the crowded spaces (e.g. an airport);
- forensic examination on videos from a crime scenes (e.g. in medium/small stores);
- persons participating to a business meetings (e.g. for video-conferencing);
- persons appearing in television videos.

The activities of NIST are addressed to assess several application scenarios and include one-to-many identification tests for video sequences according the following enrolment table:

	Video-to-Video	Video-to-Still	Still-to-Video
Enrolment dataset	N enrolled video sequences	N enrolled stills	N enrolled video sequences

Table 2 – Enrolment table provided by NIST

All identification functions operate on such multi-image or multi-frame templates and the number of images per person depends on the application area such as:

- in civil identity credentials such as passports or driving licenses where the images are acquired approximately uniformly over time (for example ten years for an Italian passport). While the distribution of dates for such images of a person might be assumed uniform, a number of factors might undermine this assumption;
- in criminal applications the number of images would depend on the number of arrests. The distribution of dates for arrest records for a person (i.e. the recidivism distribution) is modelled using the exponential distribution. NIST currently estimates that the number of images not exceeds 100.

The Table 3 shows the main video features for the datasets foreseen from NIST.

	Dataset T and B		Other datasets - Undisclosed
Collection, environment	Indoor public space with individuals walking mostly toward cameras as could occur on a transit terminal		Television footage, indoor and outdoor
Number of individuals in field of view	Multiple, usually below 20 many not fully visible but usually more than 1.		Few, most often 1, occasionally others in background
View angle	Various pitch due to different heights of camera installation, some yaw also due to subject behaviour		Pitch variation present, but yaw angles vary more due to subject behaviour
Video frame size	1920 x 1080	Various	Various
Eye to eye distance (typical)	10-100 pixels	10-150 pixels	10-120
Camera properties	Consumer-grade video	Professional-grade video	Professional-grade video cameras
Camera motion	Fixed geometry, fixed optics		Usually camera is still or slowly panning or zooming
Frames per second	24	Up to 30	Up to 30
Similar composition to	Compare to the i-Lids ¹ (Imagery Library for Intelligent Detection Systems) data but with higher spatial resolution on the face.		Similar to YouTubeFaces in that typically one subject is present and in the foreground
Accompanying stills	Yes, for video-to-still and still-to-video searches, high-resolution stills approximating ISO/IEC 19794-5 are available. In addition, off-angle images exist with many combinations of pitch and yaw. In addition, less formal "social-media" like stills are available also. Various galleries will be formed from these images. Images for which interocular distance exceeds 240 pixels will be downsized.		Stills usually resemble frames from the video. ISO/IEC 19794-5 images are not usually available.

Table 3 - Video features for the datasets foreseen from NIST

¹ <https://www.gov.uk/imagery-library-for-intelligent-detection-systems>

The detection and recognition processes are articulated usually in four phases:

- segmentation of scenes in order to select the images or sequence of images in a video on which make the application of the process of detection and recognition according specific requirements. That because, in the case of an image in a video, the face can placed in a position unuseful for the detection in the presence of important occlusions or poses excessively defilate. In the case of a video sequence is necessary that in the set of frames, at least one frame must have a minimum characteristics of pose, lighting and resolution such as to be suitable to pick up the image.
- localization of the area of interest that consists in the exact localization of the face or faces and in segmentation of significant parts that compose it. In recognition of face is useful to evaluate the consistency of the object detected by the characterization of some characteristic physical features (eyes, nose, etc.). The main issues in image processing depend on:
 - definition of the mathematical model of the face and the application of algorithms for the detection;
 - distance of the subject from the sensor of capture in order to have a sufficient number of pixels for the detection;
 - lighting conditions for which the uniform projection allows to highlight the main features;
 - robustness with respect to the alignment of the face and spin;
 - complexity of the background image that can allow processing it.
- extraction of biometric elements: the extraction of specific biometric element within an area portion of an image is a process necessary to distinguish and select a face from another. The higher the quality of the image the higher the probability of minimizing the number of false positives and false negatives during the detection;
- face recognition through association with a database of faces sample: After the detection phase, the face must be cut, cropped and resized in order to associate it with vector of values that can be compared to a database of faces known sample.

3.2. THE PARAMETERS OF VIDEO SEQUENCES

The videos transmitted both from broadcast digital TV and Web providers, or even captured from a videocamera, are a sequence of frames composed as a grid of pixels. The horizontal component is defined by pixels (or samples) and is known as a video line, the vertical component is defined by the number of lines.

The Advanced Television Standards Committee (ATSC) in North America, Integrated Services Digital Broadcasting (ISDB-T) in Japan and in South America and DVB-T in Europe, Africa and Oceania have defined standards for digital television that include how sound and video are encoded and transmitted. They also provide guidelines for different levels of quality. All of the digital standards are better in quality than previous analog signals and HDTV standards now are on the top tier of all the digital signals. A video format defines the main parameters in which video is recorded and stored. It normally specifies:

- codec/compression;
- video frame resolution;
- frame rate and the scanning system (interlaced or progressive);
- pixel aspect ratio.

Many of the parameters used to describe video formats originate in analogue TV standards and the lowest quality digital format is about the same as the highest quality that an analog TV can display. The main problems in TV broadcast video are related to:

- Aspect ratio - Standard television has a 4:3 aspect ratio instead HDTV has a 16:9 aspect ratio.
- Video frame resolution - The lowest standard resolution (SDTV) will be about the same as analog TV and will go up to 704 x 480 pixels inherited by American NTSC. The highest HDTV resolution now available is 1920 x 1080 pixels and the near future will be the SHD at 3840 x 2048 pixels (4K).
- Frame rate and scanning system - A set's frame rate describes how many times it creates a complete picture on the screen every second. DTV frame rates usually end in "i" or "p" to denote whether they are interlaced or progressive. DTV frame rates range from 24p (24 frames per second, progressive) to 60p (60 frames per second, progressive).

The aspect ratio is independent of absolute image size or resolution and can be expressed as absolute dimensions (4x3), a ratio (4:3), a fraction ($\frac{4}{3}$), or as the decimal equivalent of a ratio (1.33:1, or simply 1.33). Aspect ratios are generally expressed according to the following conventions:

- Video aspect ratios are often written as ratios, such as 4:3 for SD video or 16:9 for HD video.
- Film aspect ratios are often written as decimal equivalents, such as 1.33, 1.85, and 2.40. The larger the decimal number, the wider the image. An aspect ratio of 2.40 is wider than 1.85, and 1.85 is wider than 1.33.
- Digital video resolutions are usually written as absolute pixel dimensions, such as 720x480, 1280x720, 1920x1080, and so on.

Regarding the frame resolution, there are a lot of formats both coming from computer graphics and multimedia video. The ratio of the width to the height of an image is known as the aspect ratio, or more precisely the DAR (Display Aspect Ratio).

The Full HD video uses 1080 lines and 1920 pixels per line (1920x1080) instead a HD video with 720 lines uses 1280 pixels per line (1280x720). Both of these formats have an aspect ratio of 16:9. Common video frame sizes are shown in the Table 4.

Broadcasters are free to decide which of these formats, when and whether use them in broadcasting in standard and high definition. Also the electronics manufacturers get to decide which aspect ratios and resolutions implement on their TVs devices. Finally also the consumers get to decide which resolutions are most important to them and buy their new equipment based on that.

Digital television standards			
Standard	Resolution	DAR	Pixels
	(dots × lines)	(H:V)	
SDTV 480i, EDTV 480p	640 × 480	4:3 or 16:9 or 3:2	307,2
	704 × 480		337,92
	720 × 480		345,6
	852 × 480		408,96
SDTV 576i, EDTV 576p	480 × 576	4:3 or 16:9	276,48
	544 × 576		313,344
	704 × 576		405,504
	720 × 576		414,72
	768 × 576		442,368
720p (HDTV)	1280 × 720	16:09	921,6
	1366 × 768 (FWXGA)		1,049,088
1080i, 1080p (HDTV, Blu-ray)	1920 × 1080	16:09	2,073,600
2160p (UHDTV)	3840 × 2160	16:09	8,294,400
4320p (UHDTV)	7680 × 4320	16:09	33,177,600

Table 4 - Digital television standards

3.2.1. PIXEL ASPECT RATIO BASICS

A pixel refers to a physical picture element that emanates light on a video display. When digital content is stored into a file or on a disc or transmitted is deployed with a particular frame size and aspect ratio called SAR (Storage Aspect Ratio) which is the ratio of pixel dimensions. If the DAR (Display Aspect Ratio) matches the SAR then the process of the displaying the stored video is simply a problem of proportionally scaling it to the correct size. An example of this might be a 16:9 display showing video stored with a frame size of 1280x720 pixels where both have the same aspect ratio.

In other cases the video can be stored or transmitted with an aspect ratio SAR that does not match the display so the process of displaying the video introduces a distortion in the SAR in order to make it to match the desired viewing DAR. An example is a 16:9 display showing video stored with a frame size of 720x480 pixels. The SAR is $720:480 = 3:2$, that is an aspect ratio which does not match the 16:9 display so the stored video must be stretched horizontally or squeezed vertically to match the display correctly (Figure 16).

This situation is referred to as anamorphic video and to correct for it is necessary the introduction of a third parameter in aspect ratio that is the Pixel Aspect Ratio (PAR). The basic relationship between the three aspect ratios is $DAR=PAR*SAR$.

In digital video, the pixels used on a display are considered to be square (i.e. with equal width and height) and the pixels of the stored video are considered to be either square or non-square. If pixels are square, then the PAR is 1:1 and $DAR = SAR$ as in the first case above. If the pixels are non-square, then the PAR is not 1:1 and acts as a correction factor for the SAR. Since $DAR=PAR*SAR$, in the second case above with a DAR of 16:9 and SAR of 3:2, the PAR is 32:27.

The pixel distortion is a problem for the face recognition systems because the face contained in a frame can have significantly difference in biometric parameters.

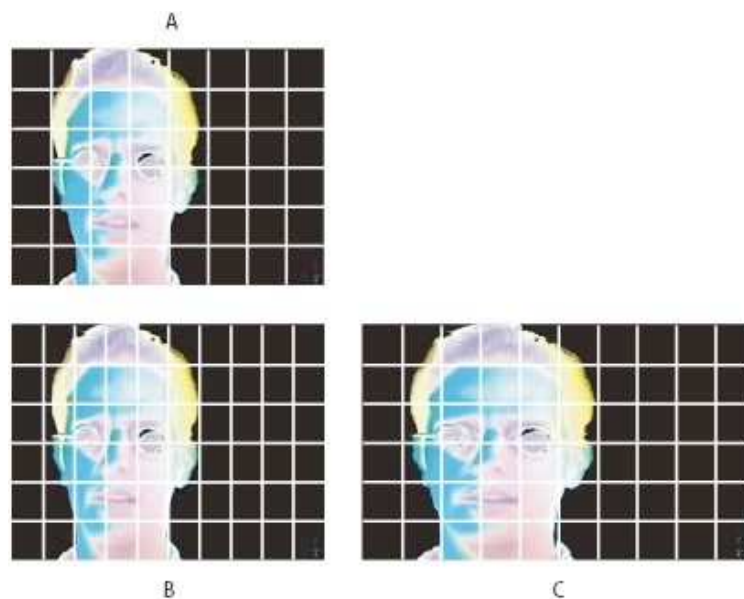


Figure 16 - Square pixels and 4:3 frame aspect ratio (A), nonsquare pixels and 4:3 frame aspect ratio (B), and nonsquare pixels displayed on a square-pixel monitor (C) (Source: www.adobe.com)

3.2.2. INTERLACED AND PROGRESSIVE VIDEO

For some video formats, images consist of two interlaced fields that together make up a frame. This approach was introduced when TV was first invented, due to a technical limitation that

prevented a full frame from being "progressively" drawn on the monitor (from top to bottom) without a noticeable visual shuttering (Figure 17). By breaking up the image into two fields (halves) and displaying one after the other this artifact was eliminated.

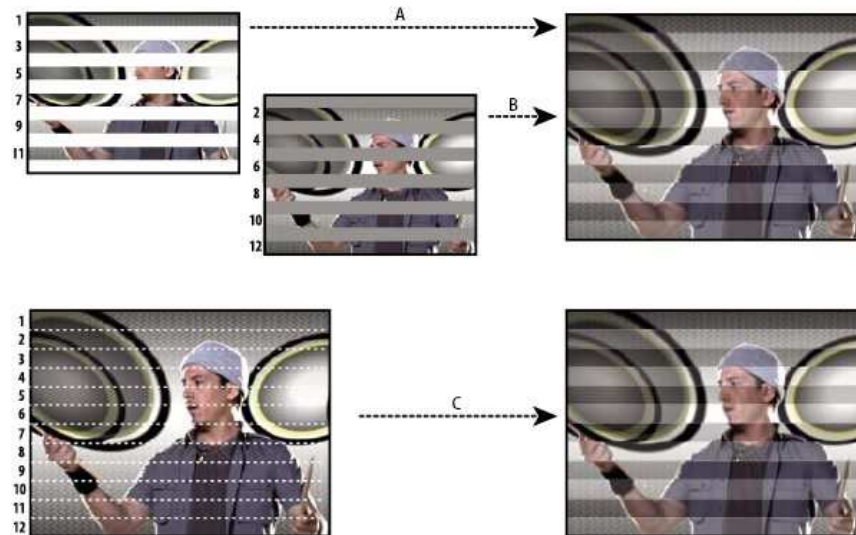


Figure 17 - For interlaced video, entire upper field (odd-numbered lines) is drawn to screen first, from top to bottom, in one pass (A). Next, entire lower field (even-numbered lines) is drawn to screen, from top to bottom, in one pass (B). For noninterlaced video (Source: www.adobe.com)

Although many newer video standards for high-definition television and all digital cinema and Web formats use progressive video (images are drawn in one pass from top to bottom), many HD standards use interlaced video for temporal resolution issues. For a given data rate, the final user can have either a perfect whole image every n units of time or a less perfect image every $n/2$ units of time. Many TV broadcasters currently still use the interlaced video in transmission.

Progressive scan video cameras usually have the ability to switch back from progressive scan to interlaced video. Typical frame rates are 60p, 30i, 30p and 24p.

When the face recognition processes works with progressive videos there is no need to de-interlace the video clips before processing them, instead in interlaced case the video frames must be de-interlaced by combining the two fields to create a single whole frame. This latter operation often produces artifacts as shown in Figure 18.

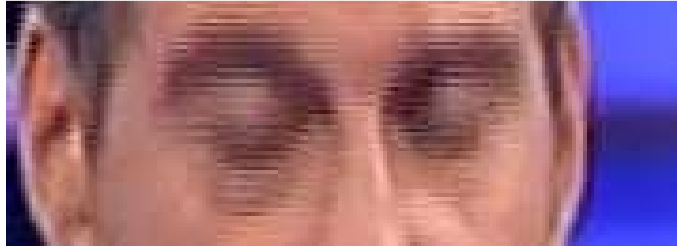


Figure 18 – Distortion produced by interlacing over a frame acquired from broadcast TV

3.3. VIDEO SEGMENTATION AND CROPPING

For the testing of the face recognition algorithms, an editing phase on video clips is required. The images contained in the videos must be selected and adapted because the subjects in them are taken in uncontrolled environment and, for this purpose, the frames must be selected and processed before the recognition processing steps. The images suffers of different problems such as occlusion, false detection or presence of multiple faces.

3.3.1. PROCESSING OF VIDEO SEQUENCES

The video contents are a sequence of frames containing a lot of elements that, even including faces, contain many not useful objects. The video streams must be segmented and filtered in order to obtain useful frames for the recognition. Additionally the video contents are encoded and adapted in order to be transmitted. Web providers, as You Tube, and TV Broadcasters use different standards of transmission according each specific transmission requirements.

The transitions within a video are commonly known as scene changes and the act of segmenting a video sequence into shots is called scene change detection. Scene change detection involves the identification and retention of frames at which the content of the scene is different from the previous frames. This involves the setting of thresholds which can be used to determine the frames that are significantly different with respect to a specific metric. Shots can present themselves in two ways:

- scene cuts, where a frame of one shot is succeeded immediately by a frame from another shot;
- gradual changes, such as dissolve, panning, and zooming, where the changes are performed gradually.

The detection of gradual changes is much more difficult in the second case, because the criteria used to determine the measurement of a change in the visual information between two images can be smooth and hard to be detected in quantitative format. The main problem in these cases is how to quantitatively characterize the difference in the visual content of two frames and to choose a threshold to determine whether this difference is significant enough to identify a scene cut.

For purposes of this thesis were considered the following dissimilarity functions in order to perform the scene detection:

- edge change ratio;
- intensity histogram.

3.3.1.1. EDGE CHANGE RATIO

Edge detection is the process of localizing pixel intensity transitions. The edge detection is used for object recognition, target tracking, segmentation, and etc. Several edge detection methods (Sobel, Prewitt, Roberts, Canny) exist and have been proposed for detecting transitions in videos. The Edge Change Ratio (ECR) methods attempt to compare the content of two consecutive frames. It transforms both frames in edge pictures that is the extraction of the meaningful outlines of objects within the pictures (Figure 19). According to hard cuts, fades, dissolves and wipes exhibit a characteristic pattern in the ECR time series. Hard cuts can be recognized as isolated peaks or during the fade in/fade out transitions. Temporal visual discontinuity usually comes along with structural discontinuity that is the edges of the objects in the last frame before the hard cut cannot be found in the first frame after the hard cut, and the edges of objects in the first frame after the hard cut in turn usually cannot be found in the last frame before the hard cut. The basic idea of edge change ratio method is summarized as:

- detect edges in two contiguous frames f_n and f_{n+1} ;
- count the number of edge pixels in frame f_n and f_{n+1} ;
- define the entering and exiting edge pixels.

If the edge change ratio is larger than the predefined threshold it is considered as a cut between the frames. This is repeated for all the frames in the video and the hard cuts can be detected.

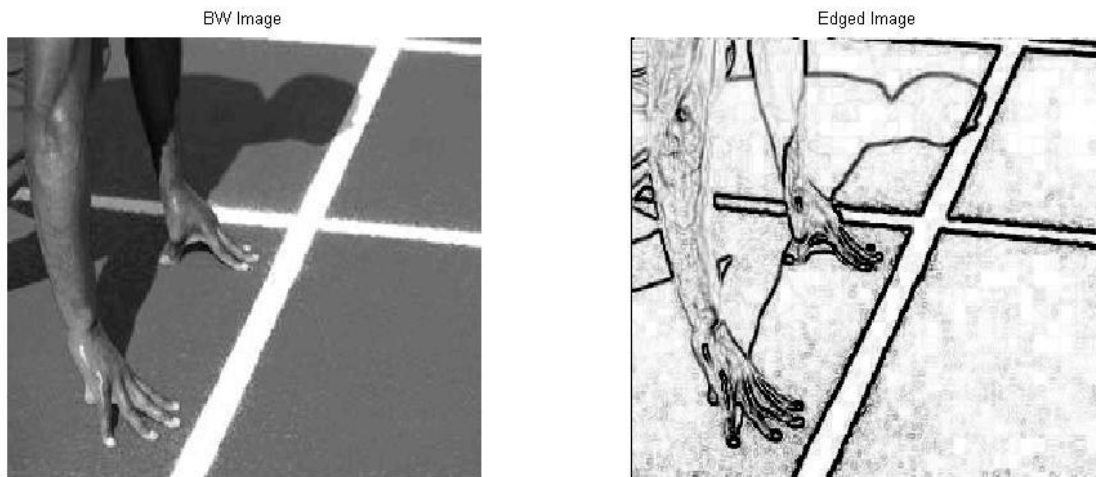


Figure 19 – Example of Edge Detection

3.3.1.2. HISTOGRAM ANALYSIS

The histogram analysis is useful in order to extract a particular information content by examining all pixels in the image and assigning a value to each depending on the local light intensity. An example can be the grey-level/colour histogram where a discrete function associates to each value of light intensity the number of pixels of the image with the same magnitude indicating, at the same time, the number of pixels of the image for each level of grey or colour intensity. From this type of histograms can be extracted information as:

- the contrast if most of the pixels takes the same values grouped within a small range;
- the light intensity are predominant low intensity or high intensity and then the image is dark or bright.

The techniques of analysis of the histograms have been applied because they are useful to detect the cases in which there is a rapid scene change with a relative quick change of light intensity.

3.4. SCENE DETECTION

The development of the script for the Scene Detection was made to achieve a segmentation of the scene mode alternative to the use of black. The black is not used for an element of discontinuity between long-term programs such as, for example, the news. For the development of the algorithm have been taken in account 2 mutually exclusive types of thresholds:

- number of blocks changed in the representation of the edges contained in the image;
- difference between histograms over a fixed threshold.

The use of edge detection to sense the change of scene in a video stream enables the segmentation of the video by means of the detection of the edges in order to make the drawing less sensitive to small changes. In principle the algorithm finds the edges in two consecutive frames and, on the basis of these edges, compare sections, or blocks, of frames among them. If the number of sections changed between one frame and the other exceeds a specified threshold a flag is triggered in order to mark time in which the scene has changed. The number of pixels used for each block was equal to 32 (Figure 20). This figure shows the events of scene transition occurred over a news program of over than 40 minutes long whereon which where detected 967 scene changes.

The algorithm of Edge Detection is used in parallel with the comparison between the histograms of two consecutive frames in order to avoid false negative in the event of rapid changes of illumination in which the Edge Detection fails to function properly.

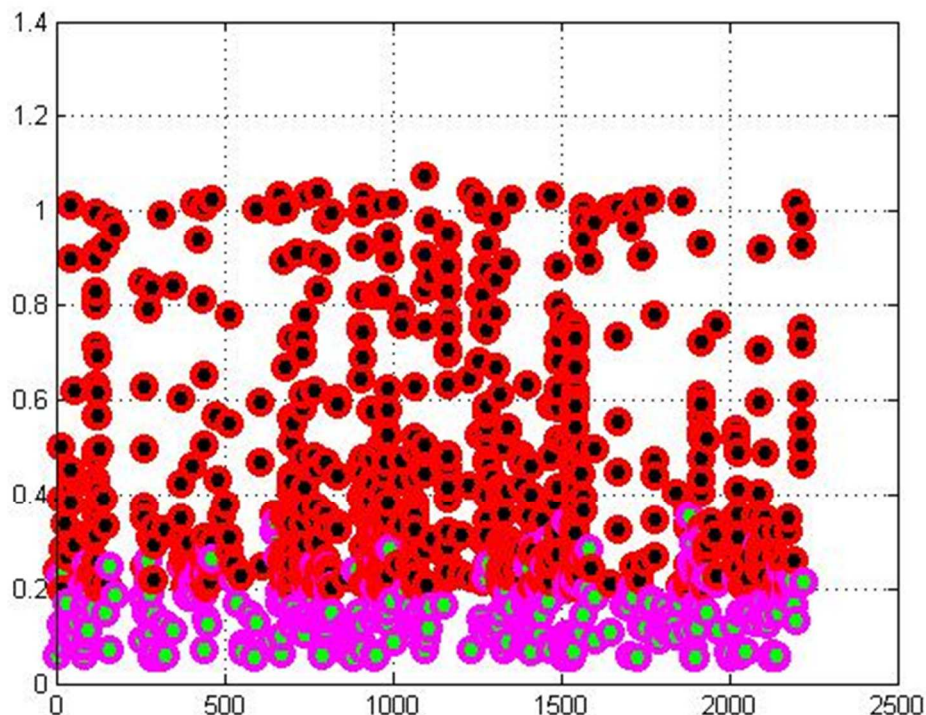


Figure 20 - Diagram of the events identified by the algorithm for scene change detection (Time: 2250s – 967 transitions)

3.5. FACE DETECTION IN VIDEOS

The face detection algorithm used in within this work is based on the procedure developed by P. Viola and M. Jones and subsequently improved by R. Lienhart and others [60][61].

The various approaches for face detection can be classified into four categories:

- Knowledge-based methods based on the definition of rules to determine which areas of the image represent a face under which encode the concept of face;
- Feature invariant approaches based on the extraction of features that characterize the human face that is invariant to the changing conditions of brightness or pose;
- Template matching methods which are defined in the "template pattern" of features that describe the prototype of a face and that are used to analyse an image by searching the similarity with a model.
- Appearance based methods in which the patterns that describe the prototype of a face are derived from a set of images of learning achieved according to the purposes that the observer try to get (e.g. Profiles). AdaBoost-based face detection also belong to this class.

Any of the methods can involve colour segmentation, pattern matching, statistical analysis and complex transforms, where the common goal is classification with least amount of error. Paul Viola and Michael Jones presented a fast and robust method for face detection that relies on the use of simple Haar-like features. Based on the concept of an "Integral Image" it generates a large set of features and uses the boosting algorithm AdaBoost to reduce the over-complete set. The introduction of a degenerative tree of the boosted classifiers provides for robust and fast interferences.

The model introduced by Viola and Jones can be summarized in three basic steps:

- introduction of a structure for the representation of images called Integral Image to perform calculations of particular geometric configurations very rapidly;
- application of a method for the construction of a classifier capable of selecting a small number of configurations using Adaboost;
- combination of classifiers later more complex in a cascade structure, which dramatically increases the speed of the tracking of faces, focusing on regions, in which the probability of finding a face is greater.

To increase the speed of a face recognition algorithm uses the information in the single image converted to grayscale and does not work directly on the intensity of the image. The algorithm, which was developed initially as a research technique of the faces, also been applied for the

localization of other biometrics such as eyes and mouth. The image input to the algorithm can be in colour or grey scale and, in the case of colour images must be provided to a conversion in grayscale images because the colour space would be sufficient to take into consideration the component relative to the brightness. The image then undergoes a transformation that re-encodes the information of individual pixels to create the Integral Image, also called summed Area Table (Figure 21). The integral image at a given location x, y contains the sum of the pixels above and to the left of x, y , including:

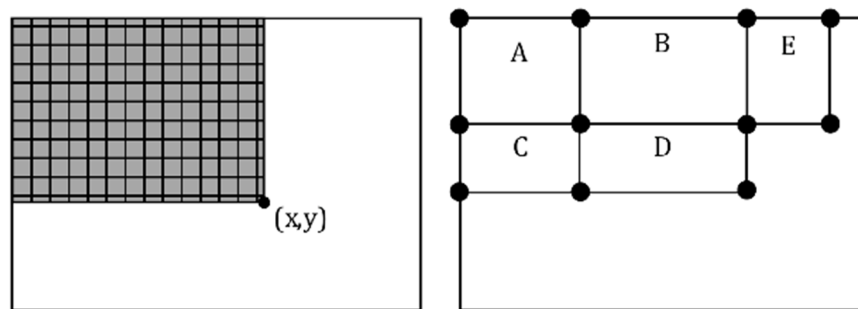


Figure 21 - Representation by Integral Image

$$ii(x, y) = \sum_{x' \leq x; y' \leq y} i(x', y')$$

Where $ii(x, y)$ is the integral image and (x, y) is the original image.

Using the following pair of recursive formulas:

$$s(x, y) = s(x, y - 1) + i(x, y)$$

$$ii(x, y) = ii(x - 1, y) + s(x, y)$$

Where $s(x, y)$ is the cumulative sum of row, $s(x, -1) = 0$ and $ii(-1, y) = 0$, and the integral image is calculated in a step from the original image.

The fundamental element of the algorithm for detection is the mechanism of feature extraction that takes the name of Haar-like features and which derives from the Haar filter (extraction of wavelet), whose calculation becomes fast in the images represented by the Integral Image.

The reasons for which are used the Haar wavelet are:

- the encoding of the differences in average between regions with different orientation allows an analysis with more steps;
- using this approach the facial features are invariant with respect to the representation.

Since within each sub-image window the total number of configurations of type Haar is very large it is necessary to exclude most of the configurations available to speed up the process.

The basic element of Viola-Jones algorithm is the use of an algorithm to distinguish between faces and non-faces. In this case, the algorithm used, is a variation of Adaboost built as a linear combination of weighted classifiers in order to achieve maximum discrimination between the two classes. The result is that each stage of the process is a combination of multiple classifiers in cascade with a subsequent segmentation of the image into regions in which the probability of finding a face is greater. One of the terms of the extent of this approach is the false negative rate, defined as the regions in which the object is present but is not detected by the system. Another parameter for evaluating the performance of the algorithm is the false positive rate defined as those areas that the system considers of interest but do not contain the desired object.

With the techniques for the Face Detection exposed is possible to realize systems with performance such as to obtain less than 1% false negative and 40% of false positives using classifiers that can be realized with filters that reduce by about half the number regions of space where the image detector final should be assessed. The sub-windows that are not discarded by the classifier are processed iteratively by a sequence of cascaded classifiers making a decision maker tree. The waterfall complete for face detection can have up to 32 classifiers (Figure 22). The cascade structure is extremely fast in the average time of finding it, though. On a data set containing 507 faces and 75 million sub-windows, the faces are detected using an average of 270 instructions per microprocessor sub-window.

At every step of the cascade of classifiers increases the number of features that are evaluated, to pass from one stage to the next is necessary that we have reached a preset threshold so that the decision maker features evaluated in an internship can be classified as belonging to a face. If the response is negative the classifier weighed stops the waterfall and discards the image area because it does not contain a face. If the cascade of classifiers ends getting a positive response in all the stages in the image area is classified as a face.

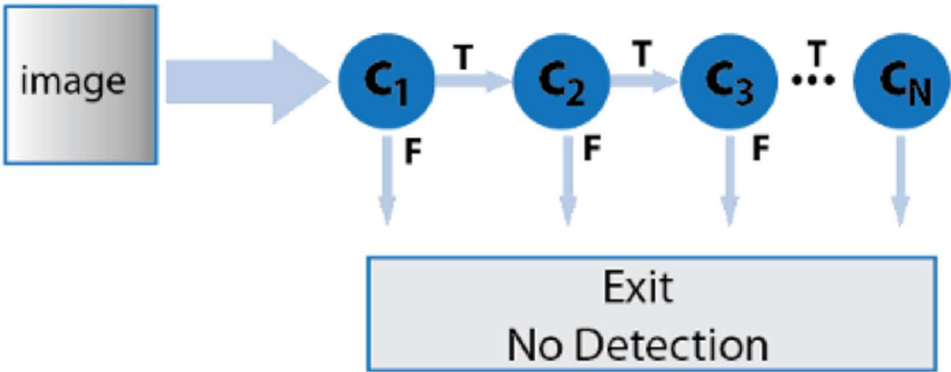


Figure 22 – Classifier Cascade

The cascade of classifiers is performed in all possible windows contained in the image, moving the window within the image starting from the top right, and moving from the top down and from right to left. The minimum size of the search window is given by the size of the images of faces and the background that are used for learning and usually the dimensions are 20x20 or 24x24 pixels.

3.5.1. QUALITY OF FACES

In order to assess an image in terms of quality was chosen to apply an algorithm for the detection of the relevant features within the face. The idea is to be able to rank the frames of a video sequence in order to detect faces inside them and make on them a classification in terms of quality. In this scenario the face, or the set of faces, with the largest number of significant elements have been be considered as the best candidates to be submitted to the identification process.

The search the frame with the best features using the algorithm KLT. The algorithm of Kanade-Lucas-Tomasi (KLT) [62][63][64] is the method for the extraction of the relevant characteristics of a face that has been used. KLT makes use of spatial information intensity detect the position of the relevant space. The image can be characterized by dunes functions $F(x)$ and $G(x)$ which represent the values in each position x , where x is a vector, for which the algorithm wants to find a vector h which measures the difference between $F(x+h)$ and $G(x)$ in the region of interest R . The application of KLT algorithm function returns the set of feature points detected in input image converted to grayscale. This function uses the algorithm developed by minimum eigenvalue of Shi and Tomasi. This algorithm is based on the Harris corner detector in which the selection criterion is calculated by assigning a score to each pixel. If the score is above a certain value of the pixel is marked as a corner. The rating is calculated using two eigenvalues that are used by a function to assign a value.

3.5.2. AUTOMATIC CROPPING AND ZOOMING OF IMAGES

The algorithm for automatic cropping of the images was made to allow automatic resizing of faces detected no loss of proportionality space. The faces are cut proportionally to fit to the database of faces for face recognition using images of 180x200 pixels.

The size of the rectangle of the face are set in relation to the ratio of the images of the database

and, in relation to these, the reference side is selected on the basis of which to perform the clipping with the proportionality ratio closer to 1. Therefore, the factor scaling is chosen according to the side with the ratio closer to 1 with respect to the horizontal axis or the vertical axis. This operation was performed in order to minimize the loss of pixels when the image is cropped. After the resizing is completed each image has been filled on the opposite axis in order to have again an image of the desired size in terms of pixels. If the image size is larger than the size of the requested images from the database, the clipping is performed, but if it is lower, our algorithm makes first a gradual filling on both sides of the line of pixels by copying recursively outer and performing, at the end, makes a further crop.

The result obtained is the extraction of all images of the same proportionality respect the original. Figure 23 and Figure 24 show a frame of a movie in which two faces are detected with different dimensions that are scaled in both figures 180x200 pixels without loss of proportionality.



Figure 23 - Scene with the detection of two faces



Figure 24 - Result of the automatic scaling of faces in which two images of different sizes initials have been reported both to the size of 180x200 pixels without loss of proportions

3.6. RESULTS OF VIDEO PRE-PROCESSING ACTIVITIES

The videos were examined in automatic mode frame by frame in order to discard the frames in which biometric elements were not detected.

The purpose of the algorithm of face detection is the discovery of one or more faces in a scene. The Viola-Jones algorithm was applied several times on the same image in order to detect over other biometrics such as nose, eyes, mouth or other body parts.

The machine used for the simulations has the following characteristics:

- Hardware: i7 CPU Q720@1.60Ghz- 4GB Ram – HD 500 GB
- Software: Windows 7 Home Premium 64bit
- Matlab 2013a

3.6.1. FACE DETECTION E TRACKING

There is a fundamental element face by a function that uses the algorithm Viola-Jones. The algorithm of VJ uses a cascade of classifiers to process the areas of an image to detect the presence of a particular object. At each stage in the cascade binary classifiers are applied more and more complex to eliminate plane regions that do not contain the object to detect. If the object is not found during any step of the cascade of classifiers image is deleted. The process is based on a set of functions derived from the transformed Haar and does not work properly in the case where in the image are present many rectangles for which they were made control functions of further biometric parameters. The threshold has been set to the default and is equal to 4. Once you locate a face this movement is tracked using the information on the hue. The diagram in Figure 25 shows the main steps of the code.

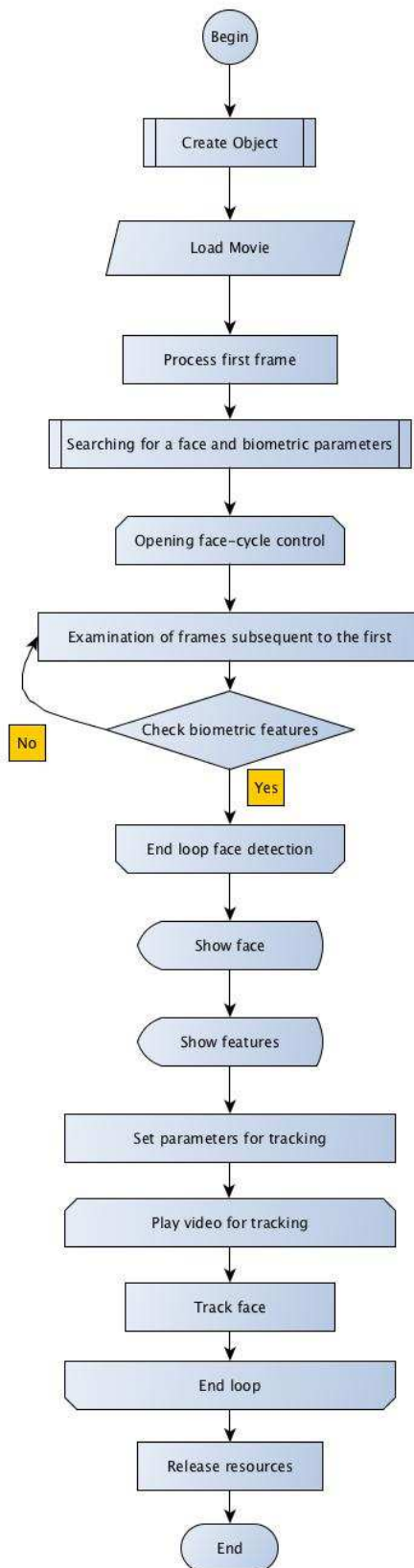


Figure 25 - Flow chart of the Detection and Tracking algorithm

3.6.2. ANALYSIS OF PERFORMANCES

There have been some tests on the following list of movies extracted from television events:

- anchorwoman_face_on_background.avi
- news_faces.avi
- Politician_woman.mpg
- Politician_man.avi
- President_Obama.mpeg

In the video “President_Obama.mpeg” have been detected all three faces in the sequence since from the first frame. On other videos can be seen that although the faces are correctly detected, the detection of other biometrics inside does not always happen and, therefore, the tracking does not work properly on all faces. This is the case of “Politician_man.avi” in which the detection of the face started after several frames because in the first several false detection occurred.

Detection and tracking was always correct in “Politician_woman.mpg” in which there is only one face.

3.6.2.1. SEARCH OF THE FACE WITH THE HIGHEST NUMBER OF KEY FEATURES

In the algorithm of KLT the eigenvalues were used to determine whether or not each pixel represents an important element on the face. A code in Matlab has been developed in order to find the best frame in terms of number of features within the face region. The target was the evaluation of the processing constraints in case of pre-processing of the movie in order to find a candidate frame with the best quality according a predefined parameter. As parameter was selected the number of features calculated by means KLT algorithm.

The Figure 26 shows the main steps of the code.

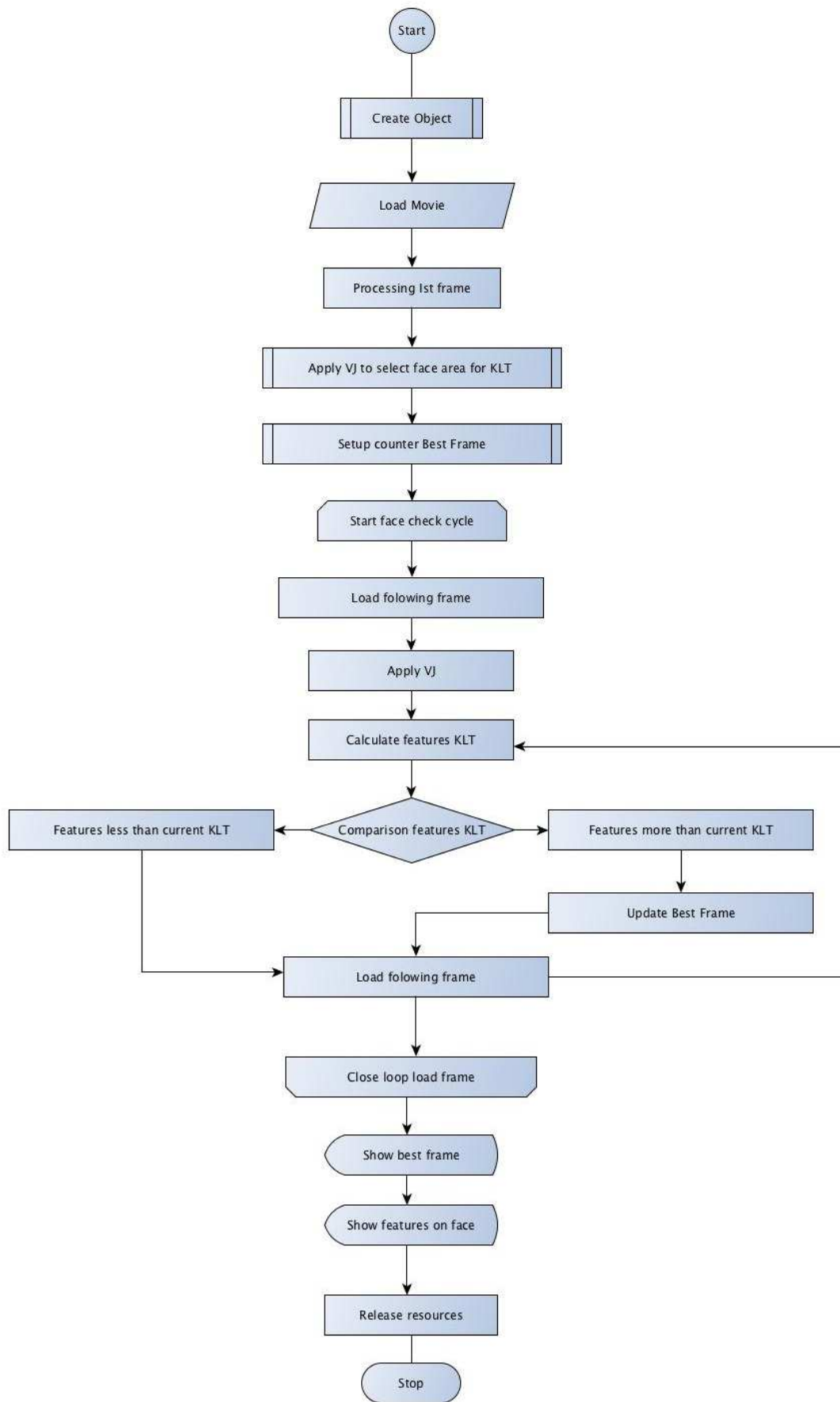


Figure 26 - Algorithm for finding the best frame in a video sequence

The following table shows the results of the test on the video sequences.

Movie	Time	Resolution	Duration (s)	Size of the face in pixel	Max number of features detected with KLT
anchorwoman_face_on_background.avi	26"	1024x576	389,56	94x94	125
news_faces.avi	14"	1024x576	85,88	59x59	65
Politician_woman.mpg	8"	704x576	112,13	163x163	449
Politician_man.avi	31"	704x576	377,12	212x212	756
President_Obama.mpeg	6"	704x576	74,31	68x68	75

Table 5 – Performances in application of KLT algorithm

Figure 27 shows an example of detection of a face within a sequence of a news program with application of the algorithm KLT.



Figure 27 – Case of application of KLT algorithm: On the right the frame with a face with a number of features with overlap of features

In case of clip “Politician_woman.mpg” the duration is of 8s and the time taken to search for the frame with the highest number of features was 152.483473s.

3.6.2.2. FACE DETECTION RESULTS

The tests were carried out on a series of recordings of the News of two national broadcasters. We examined, at the beginning, a total of 212 clips whose statistics are reported in the following table with the relative percentage error and correct identification.

TOTAL EVENT EXAMINED	
Number of events	212
Number events surveys correct number	169
Number events surveys NOT correct number	43

Table 6 – Total number of events

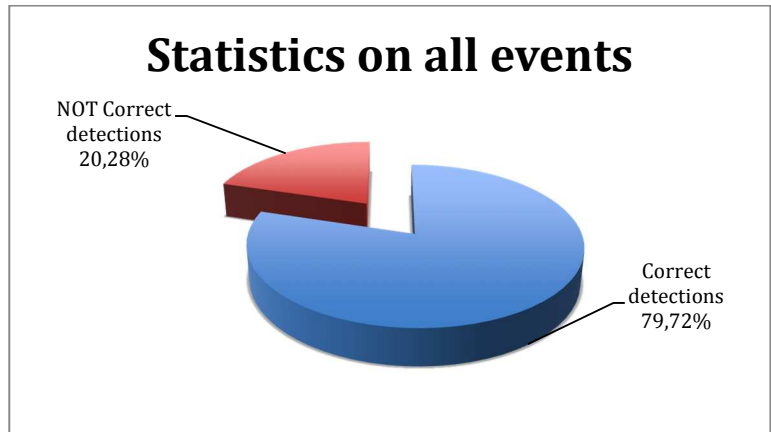


Figure 28 - Statistics on total number of events

The following table reports the classification of segmentation according to the type of events.

Events TG#1	72
Number events surveys correct number	61
Number events surveys NOT correct number	11

Table 7 - Number of events TG #1

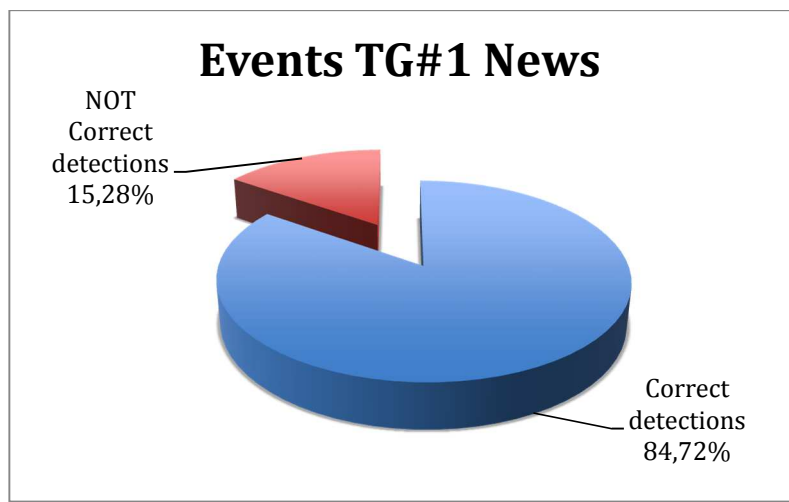


Figure 29 - Statistics on total number of events TG#1

Events TG#2	140
Number events surveys correct number	108
Number events surveys NOT correct number	32

Table 8 - Number of events TG #2

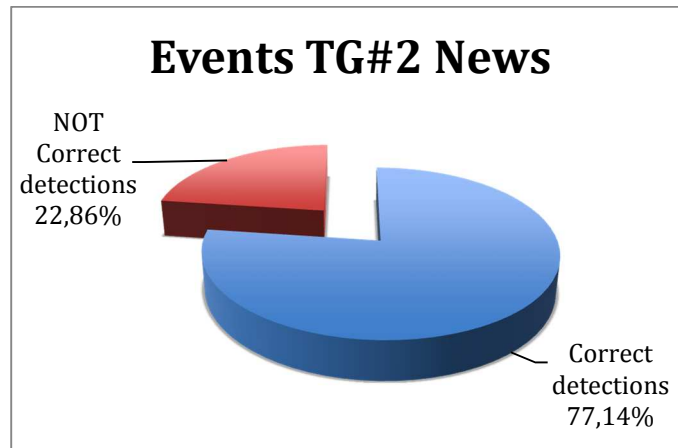


Figure 30 - Statistics on total number of events TG#1

In the latest part of this work an analogous work was carried out over additional 24 video clips (12 in HD and 12 in SD) acquired by broadcast TV at the same time using 2 video recorders with settings at different resolutions. Figure 31 shows two samples of frames acquired in the latest phase of work both in HD and in SD.



Figure 31 – Samples of frames acquired in the latest phase of work. On the left a sample in HD and on the right a sample in SD.

3.7. CRITICAL CASES IN FACE DETECTION

The reasons for which the probability of detection are equal to 100% are mainly:

- biometric traits can not be identified;
- dyed leather confused with colour background;
- background with regular geometric elements;
- presence of partial or total occlusions;
- presence of false faces.

- ***SUBCASE#1: Biometric traits unidentifiable***

The Figure 32 shows the case in which the secondary biometric traits were not detected by the algorithm that is not capable of producing positive results when an area of colour is uniform and the differences are not appreciable. The map shows the colour as the face to appear as a single dark spot.



Figure 32 - Map of matching face with biometric elements not detectable

- ***SUBCASE#2: Skin colour confused with colour background***

Another relevant case is the presence of backgrounds with coloured homogeneous than face. The case of the Figure 33 shows the occurrence in which the two faces is detected in only one case while the other one is not detectable by the background providing a false negative. The map of the greys shows that the second face of the right has continuity with respect to the background.



Figure 33 - Map of the hue in which a face in the foreground has a uniform colour from the background

- **SUBCASE#3: Background with regular geometric elements**

A detection operation is particularly complex in the cases where the background has a regular structure and it has an impact on the performances. The Figure 34 shows the case in which the detection algorithm was able to resolve the ambiguity only after several frames. The first frame, not shown in the figure, had numerous areas of false detection placed in the upper right of the image in correspondence with the area in which there are numerous colour discontinuity and stroke with a regular pattern. The area is wide recognition in many frames. The following images were obtained with the algorithm for the detection of the best frames within a sequence of 31s whose processing has had a total duration of 574.728038s. In many frame box inside which is identified the presence of a face has presented a wide area, greater than that shown in Figure 34.



Figure 34 - Case of a figure with a background with regular geometric elements

- **SUBCASE#4: Presence of partial or total occlusions**

There are several cases in which the presence of partial or semi-total occlusions for the detector produces a result of false positive or false negative. One of these cases is shown in the Figure 35 where, despite the almost total occlusion, the detector is able to capture a face although is almost entirely occluded. Although an evident false detection occurs, it can be considered a case of correct detecting with a situation of false positives in which the face is lacking of some biometric elements.

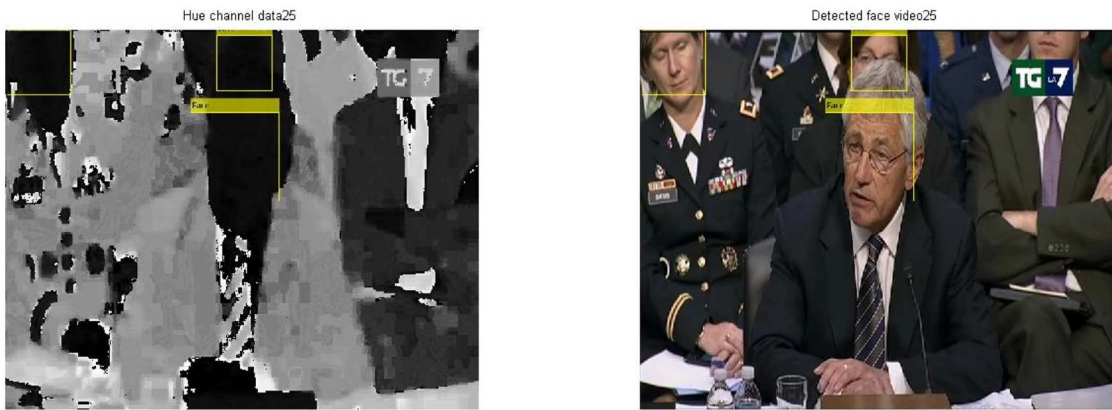


Figure 35 - Case of detection of a face almost totally occluded

Figure 36 shows a case similar to the previous one, however, in which the occluded face is not detected. Also in this case, on the face at the lower right, some biometric traits are recognizable but the detector returns a correct result and not a false positive.



Figure 36 - Case of a correct detection of a face is not occluded

In other cases, compared with a partial occlusion, the detector returns a false negative result in which a partially occluded face, while presenting all the major biometrics, is not detected by the algorithm (Figure 37).

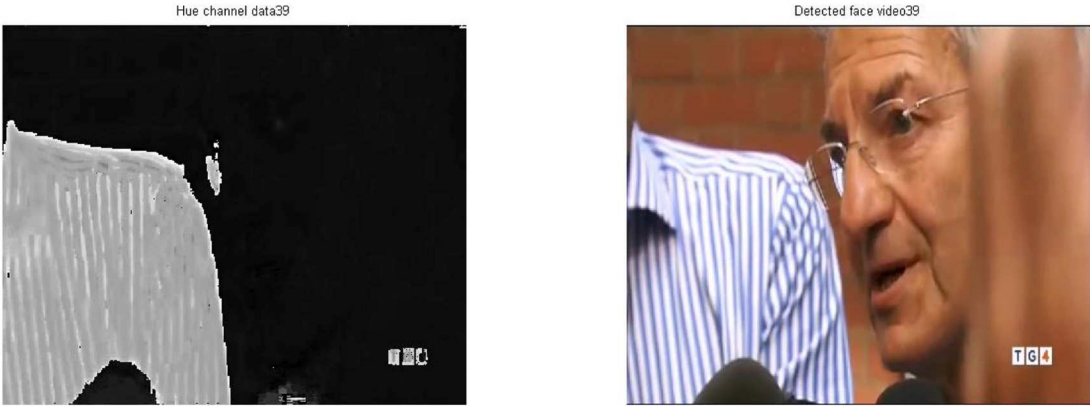


Figure 37 - The case of a false negative in the face of partial occlusion

- **SUBCASE#5: Presence of false faces**

There are cases in which they were detected false faces. Figure 38 shows the case of a false positive in which the algorithm detects a face in a piece of furniture that is in the background at on the left.

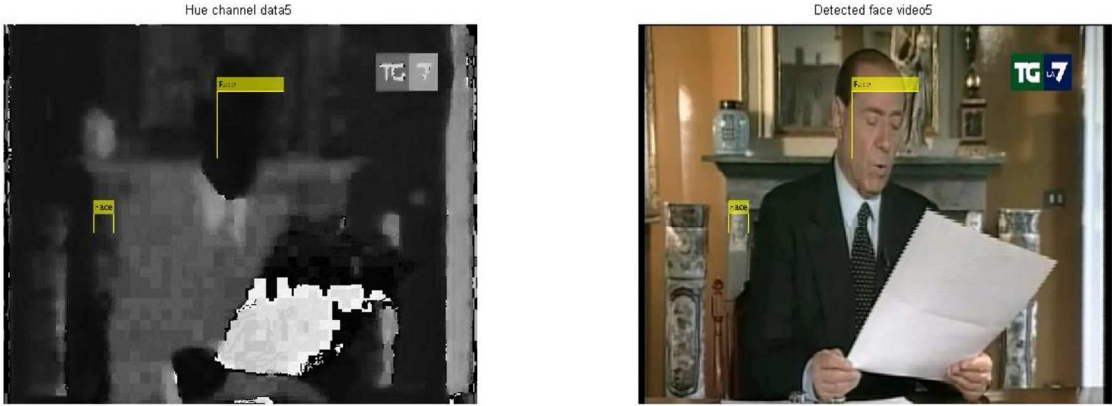


Figure 38 - The case of a false positive detected a piece of furniture

CHAPTER 4

4. EXPERIMENTATION AND RESULTS EXAMINATION

4.1. THE ADOPTED EXPERIMENTAL MODEL

The work presented in this thesis is aimed to develop a system able to perform the automatic recognition of characters in a video stream by exploiting images collected from the Web as a database as reference. This process requires as first step the creation of a database containing faces collected from the Web, followed by the analysis of the video stream in order to extract the faces of the characters and then finding the best match between the face under test and the ones in the database.

In a video sequence acquired from a broadcast television, the faces are often in pose and lighting conditions not optimal for classical face recognition systems, therefore the use of a single technique can lead to unsatisfactory results. The approach adopted in this work is based on a weighted combination of results obtained using different techniques. In particular we are considering techniques able to handle the problem of face rotation.

The proposed approach is based on two parts:

- video analysis for face detection;
- face recognition exploiting both
 - eigenfaces;
 - SIFT.

Figure 39 shows a scheme of the adopted approach.

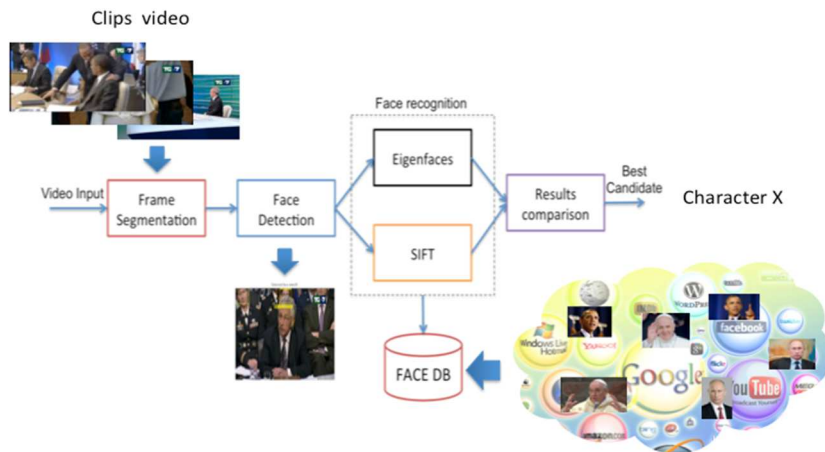


Figure 39 - Block diagram of the proposed detection and tracking process

4.1. ALGORITHMS FOR THE MANIPULATIONS OF IMAGES AND THE EXTRACTION OF THE FEATURES OF THE FACE

4.1.1. EIGENFACES

Eigenfaces algorithm allows the performing of face recognition in an efficient, simple, and accurate way. In this work, we propose the use of eigenfaces algorithm in different cases. In the first one we applied it for exploiting the time component in order to improve the recognition rate. The idea is to discover the presence of the same character in consecutive frames to reduce the probability of missed detection and increasing the probability of correct identification. In the second case we applied eigenfaces jointly with another algorithm more suitable to detect the features of the faces mainly in case of rotation.

The first successful demonstration of machine recognition of faces was made by Turk and Pentland in 1991 using eigenfaces [39]. Their approach covers face recognition as a two-dimensional recognition problem. The flowchart in Figure 39 shows on the top of the diagram the different stages in the developed eigenface based recognition system.

- The first stage is to process the images of the gallery by zooming, cropping and naming the training set.
- The second stage is to create the eigenfaces space by adding all the images of the gallery.

- The final stage is the submission of all images of the probe set in order to get the euclidian distance.

Eigenfaces are made by extracting characteristic features from the faces. The input images are normalized to line up the eyes and mouths. They are then resized so that they have the same size. Eigenfaces can be extracted from the image data by using PCA. When the eigenfaces have been created, each image is represented as a vector of weights so the system is ready to accept entering queries. The weight of the incoming unknown image is found and then compared to the weights of those already in the system.

If the input image's weight is over a given threshold it is considered to be unidentified. The identification of the input image is done by finding the image in the database whose weights are the closest to the weights of the input image. The image in the database with the closest weight will be returned as a hit to the user of the system.

Let z_i be the i^{th} frame of a video sequence. Denoting with S_j the subset of sample images of the j^{th} subject, we can write:

$$z_i = s_{k_i}^{(j)} + n_i$$

where $s_{k_i}^{(j)} \in S_j$ and n_i is a sample from a Stationary zero mean Gaussian White noise with variance σ_N^2 modelling both model mismatch and imaging system noise. Then, the conditional probability density of z_i , given $\{s_{k_i}^{(j)}; i = 1, \dots, M\}$, is

$$p\left(\frac{Z_1^N}{s_{k_1}^{(j)}, s_{k_2}^{(j)}, \dots, s_{k_M}^{(j)}}\right) = \prod_{i=1}^N p\left(\frac{z_i}{s_{k_i}^{(j)}}\right) = \frac{1}{(2\pi\sigma_N^2)^{\frac{M}{2}}} \exp\left\{-\frac{\sum_{i=1}^M [z_i - s_{k_i}^{(j)}]^2}{2\sigma_N^2}\right\}$$

Since the actual $s_{k_i}^{(j)} \in S_j$ are unknown we can approximate the likelihood of the observed sequence Z_1^M with respect to the i^{th} hypothesis as follows:

$$\ln \Lambda(Z_1^N; H_j) = -\frac{M}{2} \ln 2\pi\sigma_N^2 - \frac{1}{2\sigma_N^2} \sum_{i=1}^M \min\left\{\left[z_i - s_{k_i}^{(j)}\right]^2\right\}$$

Thus, the maximum likelihood estimator selects, among the candidates, the one for which hold the following relation:

$$\hat{j} = \text{Arg} \left\{ \text{Max}_j \left[\ln \Lambda \left[Z^N; H_j \right] \right] \right\} = \text{Arg} \left\{ \text{Max}_j \left[-\frac{M}{2} \ln 2\pi\sigma_N^2 - \frac{1}{2\sigma_N^2} \sum_{i=1}^M \min \left\{ \left[z_i - s_{k_i}^{(j)} \right]^2 \right\} \right] \right\}$$

This condition is also equivalent to:

$$\hat{j} = \text{Arg} \left\{ \min_j \left[\sum_{i=1}^M \min_{s_{k_i}^{(j)} \in S_j} \left\{ \left[z - s_{k_i}^{(j)} \right]^2 \right\} \right] \right\}$$

Finally, the candidate that less differs from the query for the whole clip is the one corresponding to the query. Figure 40 shows an example of faces associated to the Euclidean distance from the eigenspace.

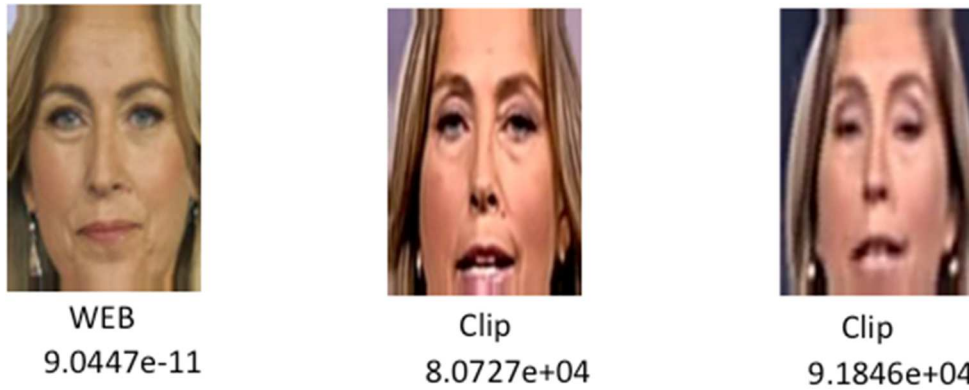


Figure 40 – Euclidean distance of a face collected from Web and two images captured from videos from the Eigenspace of the same subject

Where:

- The weight on k^{th} face is: $\omega_k = u_k^T (\Gamma - \Psi)$
- The Euclidean Distance is: $\varepsilon^2 = \left\| \Omega - \Omega_k \right\|^2$

4.1.2. SIFT

Scale Invariant Feature Transform (SIFT) is a local image features descriptor invariant to orientation and scale usually used as sparse feature representation [65]. SIFT features are extracted from images to help in reliable matching between different views of the same object. The extracted features are invariant to scale and orientation and they provide a robust feature matching with respect to several range affine distortion, 3D view point distortion, illumination changes, and additive noise.

A SIFT feature is a spatial area within the image, called keypoint, associated with a descriptor. The keypoints are extracted by the SIFT detector and their descriptors are calculated by the SIFT descriptor.

The SIFT detector uses as keypoints image structures similar to “blobs” and by searching for blobs at multiple scales and positions, the SIFT detector is invariant, or more precisely covariant, respect to translation, rotations and re scaling of the image.

A SIFT descriptor is a spatial histogram of the image gradients characterizing the form of a keypoint. The gradient at each pixel is can be viewed as a three-dimensional elementary feature vector, formed by the pixel location and the gradient orientation. The samples are weighed by the gradient and accumulated in a histogram which is the SIFT descriptor of the region. An additional Gaussian weighting function is applied in order to decrease the weight of the gradients far from the keypoint center (Figure 41).

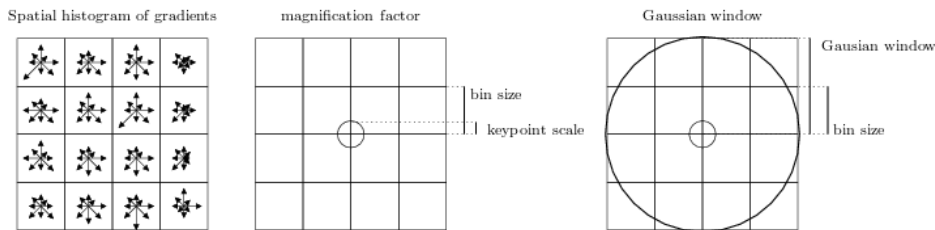


Figure 41 – The SIFT descriptor is a spatial histogram of the image gradient (<http://vision.princeton.edu/>)

In order to search for image blobs at multiple scale, the SIFT detector construct a scale space, using the Gaussian Scale Space as the collection of smoothed images:

$$I_{\sigma} = g_{\sigma} * I \quad \text{with } \sigma \geq 0$$

Where I denote an idealized infinite resolution. Nevertheless the image I is available only conceptually and in the real applications the input is I_{σ_n} that is a pre-smoothed image at a nominal level $\sigma_n = 0,5$ to account for the finite resolution of the pixels. Considering that the Gaussian kernel is:

$$g_{\sigma}(x) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2} \frac{X^T X}{\sigma^2}\right)$$

In practice the scale space is computed by means:

$$I_{\sigma} = g_{\sqrt{\sigma^2 - \sigma_n^2}} * I_{\sigma_n} \quad \text{with } \sigma \geq \sigma_n$$

Scales are sampled at logarithmic steps given by

$$\sigma = \sigma_0 2^{\frac{o+s}{S}}$$

with $s = 0, 1, \dots, S-1$ and $o = o_{\min}, \dots, o_{\min} + O - 1$.

Where $\sigma_0 = 1,6$ is the base scale, o_{\min} is the first octave index, O the number of octaves and S the number of scales per octave where an octave corresponds to doubling the value of σ .

Blobs are detected as local boundaries of the Difference of Gaussians (DoG) scale space, obtained by subtracting successive scales of the Gaussian scale space:

$$DoG_{\sigma(o,s+1)} = I_{\sigma(o,s+1)} - I_{\sigma(o,s)}$$

For each candidate keypoint the interpolation of nearby data is used in order to determine its position. The interpolation is done using the quadratic Taylor expansion of the Difference-of-Gaussian scale-space function, $DoG_{\sigma(o,s+1)}$ with the candidate keypoint as the origin. This Taylor expansion is given by:

$$DoG(x) = D + \frac{\partial D^T}{\partial x} x + \frac{1}{2} x^T \frac{\partial^2 D}{\partial x^2} x$$

Where D and its derivatives are evaluated at the candidate keypoint.

The first step is aimed to identify the locations of potential interest points in the image by detecting the maximum and minimum of a set of DoG filters applied at different scales all over the image. Then, these locations are refined by discarding points of low contrast. Finally, the orientation is then assigned to each key point based on local image features. An example of feature matching is shown in Figure 42.

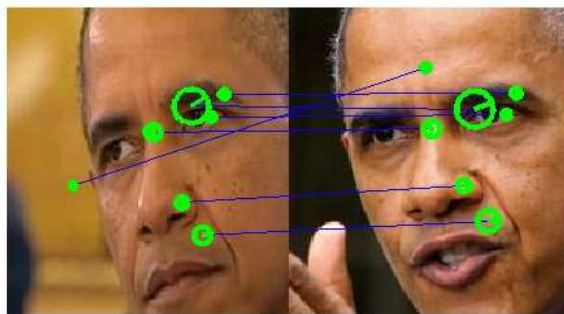


Figure 42 – Comparison of images by means SIFT of the same subject

A SIFT keypoint is represented as a circular image area with the orientation and it is described by means of four parameters:

- the keypoint center coordinates x and y ;
- the scale through the radius of the region;

- the orientation through an angle in radians.

4.2. THE DATABASE OF FACES CREATED USING WEB IMAGES

Experimental tests have been carried out for evaluating the effectiveness of the proposed system. The faces from Internet have been manually annotated and they have been used as reference for identifying images extracted from the clips. In this work the database has been made in two phases. In the first phase, 100 faces of 10 characters have been used for training the system and in a second phase other 100 has been added. Each character's training set is composed by a collection of 10 images of 180x200 pixels originating from pictures downloaded from Web. These images have been resized, cropped and filled in order to deploy a database of pictures with all elements with the same dimensions usable for the experiments.

To analyse the performances of the face recognition system, experimental tests were carried out on datasets created by downloading from the Web annotated images. The final database contains 20 individuals (mostly male) with 10 images each. The total number of images used in this experiment was 100 during the first test and 200 at the end. For each dataset, we created 10 subsets via randomly selecting the training images per individual. Figure 43 shows a set of the database of faces downloaded from Web and in the experimentation phase. The faces were scaled and cropped to be adapted to the size of 180x200 pixels (Figure 44). Additionally too similar poses have been discarded in many cases and for each character the set of 10 images has been completed in sequential phases. At the beginning 4 or 5 picture of the character have been added. Each following image included in the dataset has been tested by running the eigenvectors algorithm trained with all the faces included until that moment.



Figure 43 – A subset of annotated faces DB downloaded from Web

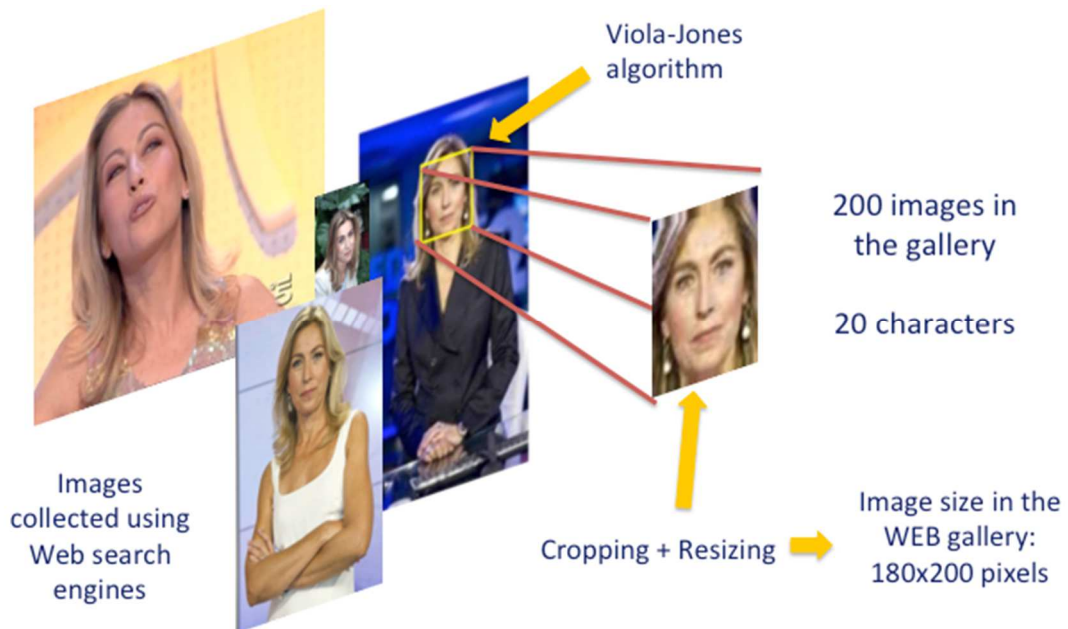


Figure 44 - Face detection, zooming and cropping

4.3. SELECTION OF FACES FROM VIDEOS

An important step of the proposed approach was the video analysis. The aim is the characters identification and their annotation (recognition) by comparing the extracted faces with the annotated ones collected in database containing the Web faces. To this aim, several video sequences have been recorded by using a commercial TV Tuner. In particular, we considered two Italian TV broadcasting companies. In a first face 212 clips have been recorded and a database of

10 characters has been created, in a second phase another 20 clips, both in SD and HD, have been recorded and another set of 10 characters has been extracted.

Also in this case, to identify the regions containing faces the Viola Jones algorithm has been applied. Viola-Jones detector is a strong, binary classifier based on a cascade of weak detectors. During the learning phase, the cascade of detectors is trained up to obtain the required detection missed detection rate by using the Adaboost algorithm.

The generic frame is then partitioned in several rectangular patches that are the inputs to the classifiers cascade. If a rectangular image patch passes through all of cascade stages, then it is classified as “positive” that is, containing faces. The process has been iterated at different scales. In Figure 45 and Figure 46 example of faces detected by the Viola- Jones algorithm in video frames are reported. As can be noticed in Figure 46, two faces of the same news are characterized by different pose and expression.

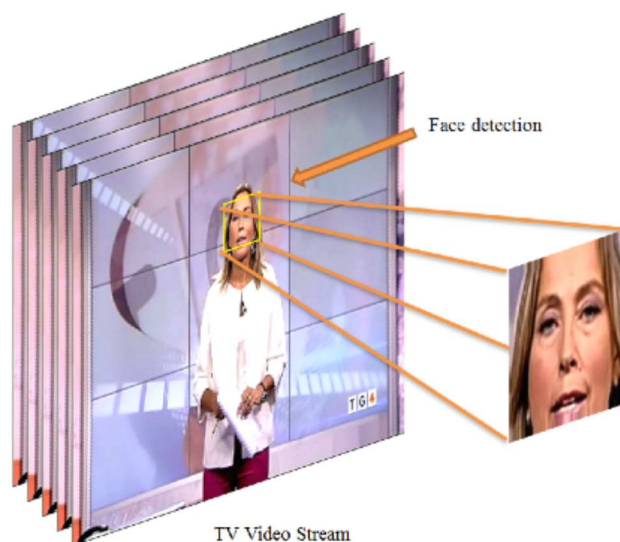


Figure 45 - Example of region containing face.



Figure 46 - Regions containing faces extracted from the same news stream at different times. As can be noticed in different frames, the same character image differs from pose, expression, and lighting conditions.

In both phases more than 400 faces were extracted and from the video clips and over a selected set has been performed the recognition. Figure 47 shows a subset of the selected faces.



Figure 47- Face DB extracted from the recorded TV streams

The results, concerning the face detection phase, have demonstrated a capability of correct detection probability of 80%. The following cases are the main motivations for which the probability of detection is not equal to 100% as explained in the previous chapter:

- the biometric features can not be identified;
- the skin colour is too similar to the background colour;
- the presence of background texture characterized by regular geometric elements;
- the presence of partial or total occlusions;
- the presence of objects that look like faces.

Another possible cause is in the quality of the received signal. In fact, transmission of compressed video over broadcast TV channels can produce errors, packet losses that can significantly degrade the image quality. Additionally the pixel aspect ratio can be modified significantly during the format conversion or the frame adaptation provided by the broadcaster.

However the performances of the Viola-Jones detector may be further improved by exploiting recent modifications of the original algorithm in order to lead the detection probability to 98% [66].

4.4. FACE RECOGNITION PROCESS

In the following, the details of recognition process are described. The faces, after being detected, have been scaled to the same size without loss of proportionality in order perform the recognition process. After faces have been detected, state-of-the-art tools are used for extracting the features that have been used for face recognition: Eigenfaces [39] and Scale SIFT [53].

The recognition has been performed in two phases:

- the sequences of the faces extracted from the streams have been processed using both Eigenfaces and SIFT methods obtaining a separate ranking with each algorithm;
- the results of each method have been combined using a weighted index.

Each frame of the streams has been compared with all images in the database of the faces downloaded from the Web containing images of 180x200.

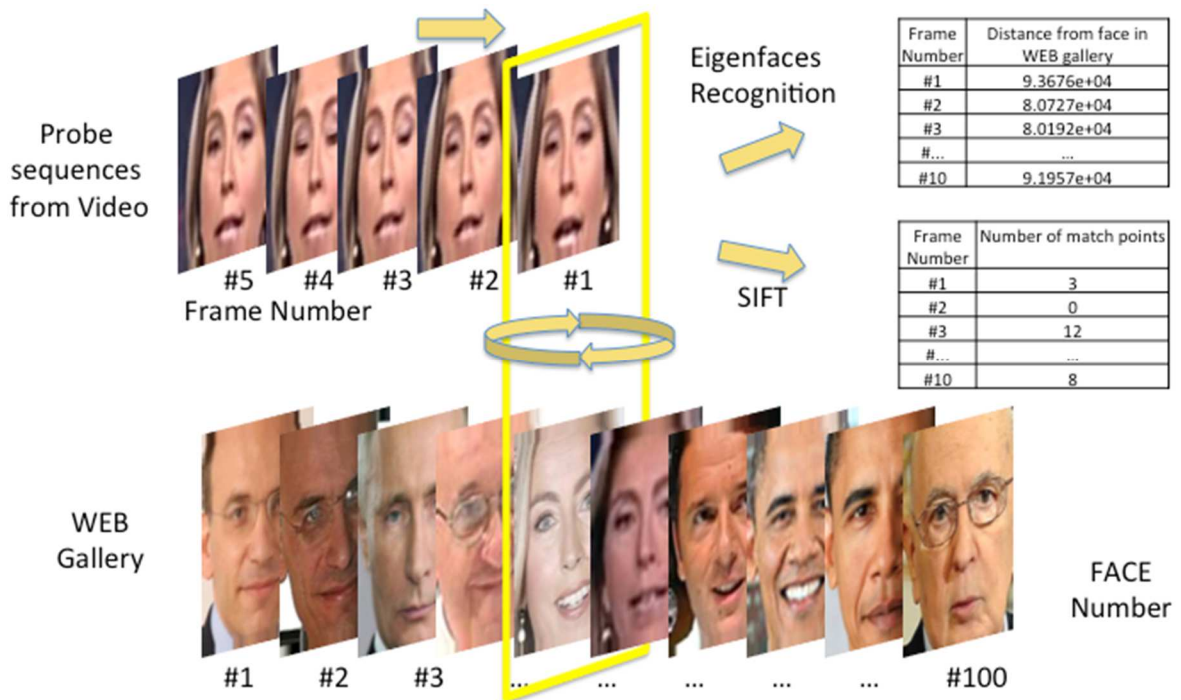


Figure 48 – Recognition process implemented in experimentation

In order to improve the effectiveness of the recognition technics, the similarities provided by the two methods have been linearly combined. In more details, the Eigenfaces method evaluates face similarities in terms of distance computed in the Eigenfaces' space. The face with the smallest distance is then considered as the best candidate for the Eigenface method. On the other hand, the SIFT methods evaluates similarity counting the number of matching points. The face with the highest number of matching points is considered to be the most similar to the searched one.

In the proposed technique, given a pair (y, z) of images, characterized by an eigenface distance $d_{EF}(y, z)$, and a SIFT matching points score $p_{SIFT}(y, z)$, the overall similarity has been computed as

$$s(y, z) = w_1 p_{SIFT}(y, z) - w_2 d_{EF}(y, z)$$

Where w_1 and w_2 are the weighting factors.

By minimizing the combined similarity, we were able to correctly identify the 80% of the video faces in one of the experiments when we were exploiting both Eigenfaces and SIFT for improving the recognition scores with a dataset of 100 faces of 10 characters as described in the following paragraph.

4.5. PERFORMANCE COMPARISON

The work has been carried out in three steps:

- a first set of faces of database has been deployed and the recognition process has been carried out using the Eigenfaces method;
- using the original database a combined approach using Eigenfaces and SIFT has been performed;
- the database has been improved and a comparison between SD and HD videos has been performed.

At the beginning we focused on experimental tests on a datasets created by downloading from the Web containing annotated images and containing 10 individuals, (mostly male), with 10 images each. The total number of images used in this experiment was 100. For each dataset we created 10 subsets via randomly selecting the training images per individual. The database is organized in groups of faces belonging to the same character, each group contains 10 faces of the same character and in the tables, and each group is identified with G_i . For each character identified in video sequences has been extracted a sequence of 10 faces identified with F_i . For each extracted face the ranking of the i^{th} face over all G_i group has been calculated.

Considering N faces of the video content and K elements in the database extracted from Web we performed $K*N$ matches.

In this first stage, the faces were analysed with the algorithm of eigenfaces obtaining a success rate equal to 96.55% using only three selected characters over the 10 available. Figure 49 shows the cumulative distance between the face under test and the ones in the databases.

In all these cases the best candidate is the one that shows a lower Euclidean distance that in Figure 49 is represented by the dotted curve [67].

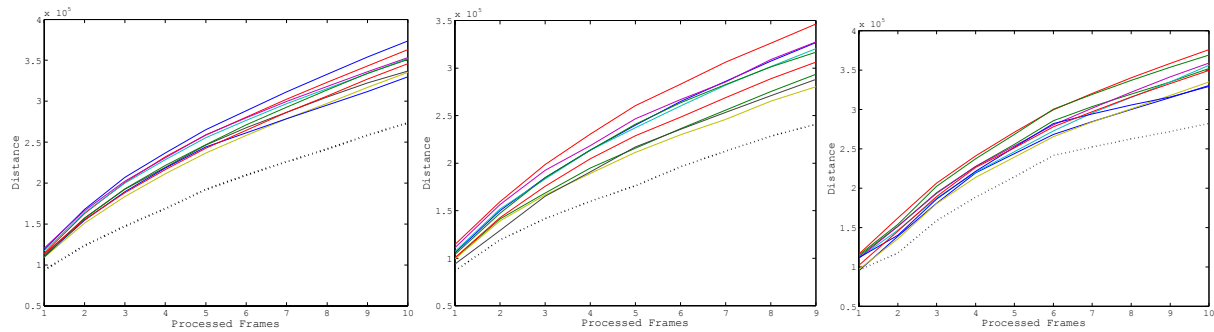


Figure 49 – Best candidate calculating the Euclidean distance

With eigenfaces method the ranking has been carried out using the minimum of the Euclidean distance (Table 9). Using SIFT instead the number of the correspondences of the local features has been computed by taking as reference the maximum of occurrences for each frame (Table 10). Afterwards the results have been normalized and combined with separate scale factor for each methods in order to select the face candidate as the nearest the face of one character included in the database. The combination of ranking between Eigenfaces and SIFT has been useful in order to improve the performances of a single method. As the results of the combination model an overall percentage of 90% of recognition is achieved [68].

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
G1	1	10	9	10	4	7	3	8	10	10
G2	8	1	4	4	5	4	6	4	5	9
G3	2	9	1	8	6	6	10	10	2	6
G4	7	7	8	1	10	8	7	7	8	8
G5	10	8	7	9	8	10	8	9	9	4
G6	3	3	2	6	1	1	4	2	6	5
G7	6	4	6	2	2	2	1	3	3	1
G8	4	2	3	3	3	3	2	1	4	2
G9	5	6	10	7	9	5	9	6	1	7
G10	9	5	5	5	7	9	5	5	7	3

Table 9 – Eigenfaces results

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
G1	1	3	10	3	1	4	5	7	2	5
G2	9	1	6	10	7	9	6	3	10	8
G3	4	5	1	4	2	1	7	4	4	2
G4	10	8	4	1	5	8	2	8	8	9
G5	5	2	2	5	3	2	3	1	5	6
G6	6	9	7	2	9	6	8	10	6	10
G7	2	7	8	8	6	7	1	6	3	4
G8	7	10	9	7	4	3	9	2	9	7
G9	8	6	3	6	10	8	5	5	1	3
G10	3	4	5	8	8	10	10	9	7	1

Table 10 – SIFT results

Experimental tests have been carried out also in HD for evaluating the effectiveness of the proposed system. To this aim, several video sequences have been recorded by using different commercial TV tuners. In particular, we considered several Italian TV broadcasting companies. Clips both using Standard Definition Channels @720x576i and HD channels @1920x1080i have been recorded and a database of 200 subjects has been created, improving the previous case, with faces downloaded from Internet. The faces from the Internet have been manually annotated

and they have been used as ground truth for identifying random faces extracted from the clips. In this work, 200 faces of 20 characters have been used for training the system. In the gallery database the training set of each character is composed by a collection of 10 images of 180x200 pixels that have been downloaded from the Internet resized, cropped, and filled in order to have a database of pictures with all elements with the same sizes.

The set of test images was obtained from the clips of the same TV channel recorded simultaneously both HD and in SD. Also in this case the images have been resized, cropped, and filled in order to have a database of pictures with all elements with the same sizes with the gallery database [69].

The images of the clips have been compared both using the Eigenfaces and the SIFT methods with the gallery database in order to detect cases where the different acquisition source, in HD and SD, can produce different results in the recognition process. The frames acquired using the HD standard have size 4 times larger than the SD ones and the original size of the picture produce differences in the output after the operations of resizing and cropping. Usually a face captured in SD is about 50-60 pixels in the shorter dimension and 70-80 in the longer. The images in video clips suffer of several distortions due mainly to:

- subject pose;
- lighting and colour settings;
- artefacts of motion compensation algorithms;
- interlaced video sources;
- anamorphic distortion caused by image squeeze during the video editing.

The use of images recorded during TV broadcasts affects the performance of the algorithms used to perform the association between the gallery of characters in the database created with Web images and faces detected in video sequences. Figure 50 shows the case of the same algorithm, SIFT, applied to a pair of still images of the same character. In one case the photos are downloaded from the Web and the other the comparison is between one of the photos downloaded from Web an image acquired in a frame of a video sequence. In the second case, the algorithm detects a number of matches reduced in comparison to the first case due to mainly to the distortion of the image.

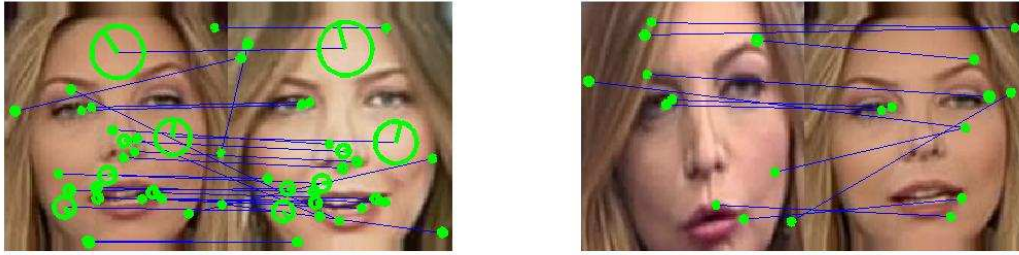


Figure 50 - Comparison between still to still and video to still images using SIFT

In the third case, the database has been improved up to 20 character each group contains 10 faces of the same character and in the tables, and each group is identified with G_i . For each character identified in video sequences has been extracted a sequence of 10 faces identified with F_i . For each extracted face the ranking of the i^{th} face over all G_i group has been calculated over the whole database of 200 faces. Also the probe gallery has been improved with faces captured both from videos @720x576i and @1920x1080i.

With Eigenfaces method the ranking has been carried out using the minimum of the Euclidean distance and for the SIFT the number of the correspondences of the local features has been computed by taking as reference the maximum of occurrences for each frame on a sample of 10 characters on the whole set of 20 because the images in HD were available only on the set of faces added in the last phase of the work. Afterwards the results have been normalized and combined with separate scale factor for each methods in order to decide if the face candidate can be selected as the best for the correct recognition Table 11.

Character	SD	HD
G1	NO	NO
G2	NO	OK
G3	NO	NO
G4	OK	OK
G5	OK	OK
G6	NO	NO
G7	NO	NO
G8	NO	NO
G9	NO	OK
G10	OK	OK
Score	3/10	5/10

Table 11 – Results of comparison SD/HD

The score shows that the SD video contents contains a lot of video distortion that downgrade the performance of the recognition. Using the HD video recordings the performances can be improved overcoming some problems that occur during the image rescaling and resizing. Also the anamorphic distortion caused by image squeeze during the video editing produce some additional artefacts in the probe images that before the cropping must be de-interlaced and recoded in mp2 with the correct aspect ratio.

Figure 51 and Figure 52 plot the receiver operating characteristic (ROC) curves, which are generated by sweeping the threshold through the entire range of possible values both for SD and HD video content. The difference between AUC (Area Under Curve) in case of SD and in HD case shows a better performance in case of use HD video clips. The results of each test have been combined according the similarity function described in the previous paragraphs as shown in the following table:

	Rate of success (%)
Results HD	
Failure	46
Success	55
Results SD	
Failure	62
Success	38

Table 12 – Global score combining Eigenfaces and SIFT

In machine learning the ROC the Area Under Curve (AUC) statistics are useful for model comparison and, in this case, can be interpreted as the highest probability of recognition of the

characters in the HD video clips. The value $AUC=0,5$ marks the threshold for a random test and in both cases it is higher. Nevertheless the value for SD is not sufficient because it is classified usually as poor in AUC scale, instead is fair for values between 0,7 and 0,8.

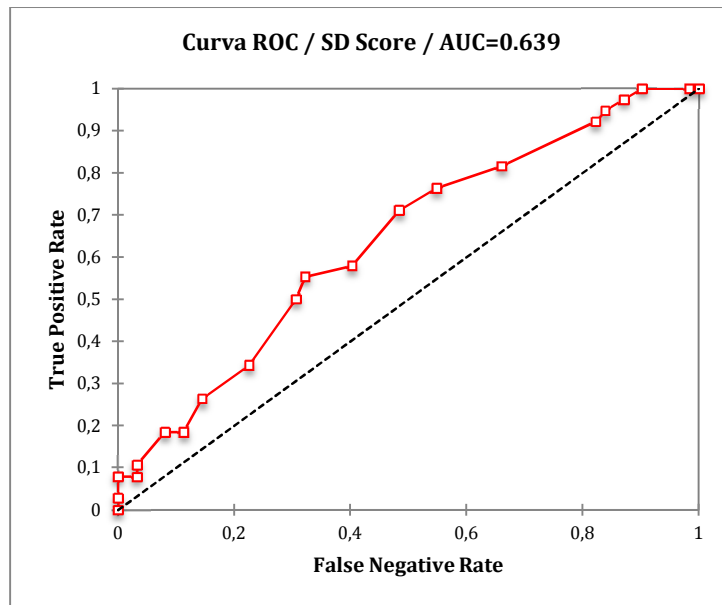


Figure 51 - ROC curve using SD test set

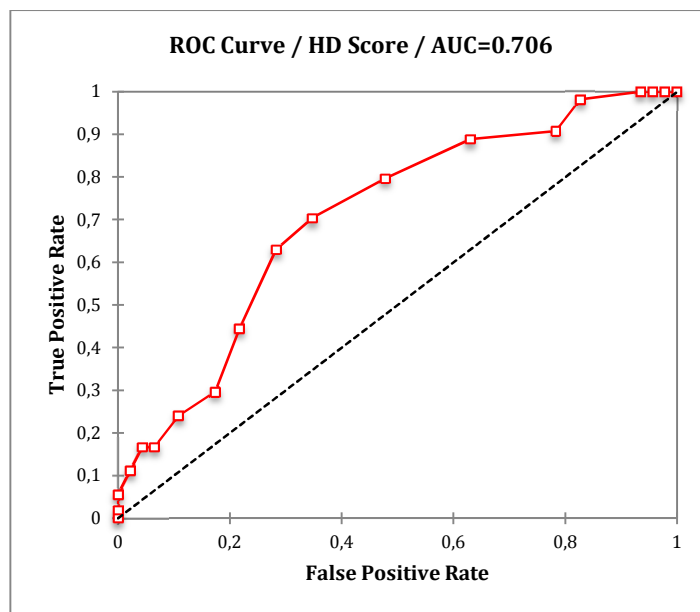


Figure 52 - ROC curve using HD test set

These results have to be analyzed based on the considered simulation scenario. The conditions of face selection for the gallery from the Web and from the set of the video images in SD and HD

were unsupervised. Surely the evolution to systems for the acquisition with high resolution sensors and diffusion of higher performance TV broadcast system (eg. 4K) will allow the improvement of the performance of the technologies for the face recognition.

CHAPTER 5

5. CONCLUSIONS AND FUTURE WORKS

The work of this thesis is aimed at study the possibility of creating a system of automatic face recognition system able to use images collected on the Web as a tool to create a knowledge base continuously updated.

The capability to develop an automatic system able to assist in recognition by using the Web search capabilities is a challenging target. Humans make use of face as basic capability in order to identifying people. This makes of the automatic face recognition a very important tool from the point of view of a wide range of commercial and law enforcement applications. Although a lot of significant works has been done on this subject, the current systems are still very far from the human perceptual system.

The principal studies and technologies on face recognition research are limited to recognizing faces from still images and, also in the most successful cases, the main results have been reached using supervised systems. Face recognition has attracted a lot of attention in the last decades and achieved high recognition rate under controlled environment. More and more researchers now focus on face recognition in the wild, which is difficult because of the variance of pose, illumination, occlusions and so on.

As described in Chapter 1, the constrained recognition processes use images captured in controlled environment where the subjects are collaborative and where they know that are being photographed for face recognition galleries. More challenging and similar to the human perception is the unconstrained case where both the subjects to identify and the gallery of the possible candidates are collected in an uncontrolled environment without the explicit collaboration or knowledge of the subjects. Metaphor of this scenario is this work which is aimed at emulate the human behaviour and where on one side there is a set of videos recorded by broadcast TV programs and, on the other side, there is a collection of faces extracted by images retrieved using the Web search engines.

In this scenario, the capability to use images to retrieve information from other information contents is an important feature that can be viewed as the evolution of the concept of IoT in order to reinforce concept of VIoT. VIoT is a concept derived from the IoT that is focused on the combination of the visual and image processing technologies with IoT itself. It focuses on the combination of the visual related technologies to traditional IoT, which consisted of cameras, information transmission networks and intelligent image or video analysis algorithms. With the wide diffusion of Web cameras, the application of VIoT can retrieve the object location via image of the information of the scene, attach a visual label to the object and then return the label to the information network. Applying this paradigm, the instance of an object on the Web will be an abstract element and will represent the connection point that will enable producers and consumers of information to remain hooked to the object itself and allow the development of new services.

5.1. CONCLUSIONS

The work described in this thesis describes an approach for face recognition in an unconstrained environment using as reference a collection of faces collected on Web. The stimulus was within of some research projects where some companies wanted to develop new tools, to improve the current audiovisual investigation workflows, by using automated systems in order to analyse the TV contents.

Specifically, the requirement was the improvement of the performance in tagging the videos with the indication of the presence of a character in a sequence of TV frames in order to create summary reports with minimal human intervention.

The basic proposed approach is based on hybrid model using, at the same time, holistic and features based methods. In a video acquisition from the broadcast television, the faces are often in pose and lighting conditions are not optimal for classical face recognition systems, therefore the use of a single technique can lead to poor results. The approach adopted in this work is based on a weighted combination of results obtained using different techniques to face the problem of face rotation adopting the Eigenfaces method in order to have the benefits of holistic approaches and the SIFT in order to have the advantages of the feature based methods.

The work has been divided mainly in two phases:

- Web images and video analysis for face detection;

- face recognition exploiting Eigenfaces and SIFT.

The process of recognition has followed the following phases:

- Face detection: in the first stage of the process of face recognition requires the detection of the faces within the sequences in the original video and it was developed using the algorithm Viola-Jones.
- Treatment of face images: any image containing a face in the original video was resized, cropped and saved in an archive for the recognition phase.
- Comparison of faces: the comparison between the faces has been carried out using different recognition algorithms to find the person who, in the archive of reference, is the one that is most likely to be similar to the face detected in the movie.

The study has been conducted in order to develop a system able to perform the automatic recognition of characters in a video stream exploiting images collected from the Web as a database as reference gallery. This process has required as first step the creation of a database containing faces collected from the Web, followed by the analysis of the video stream in order to extract the faces of the characters and then finding the best match between the face under test and the ones in the database.

For this reason the same content were recorded, at the same time on the same channel, using 2 PVR in parallel keeping in conversion original size in terms of resolution and pixel size. Subsequently, the contents were deinterlaced and from them were extracted sequences of faces in a similar position.

At the end of this operation the faces sample were processed with the algorithms of face recognition and has been made of the comparison.

As described in the previous paragraphs, the use a hybrid technique improves recognition performance than using one and the use the content at HD resolution achieves, on the same content, a significant improvement in performance.

The development of new technologies in this field has to face the problems of the contents available in an unconstrained real scenario. The progress in the quality of the contents and in the communication media will allow the improvement of the performances. While the images available on the Web will increase in terms of resolution, also the HD technologies will evolve towards 4K solutions and also the Next Generation Networks will enable new services requiring ultrabroadband performances. The counterpoint to this evolution will be the increase of the processing capability required to perform the detection and recognition processes. The

innovations in face recognition will have therefore as basic requirements high computing performances, the deployment of ultrabroadband networks and the availability of high resolution images and video sensors.

5.2. FUTURE WORKS

Innovations on face recognition technologies will have significant impacts in order to improve the levels of security in sensitive locations, in development of new services and in change of the user's experience. In addition, the application of VIoT paradigm is aimed to apply the concept of a connected world of things, which consists of intelligent information processing. The advantage of VIoT is the ability to extract labels from the appearance of objects without explicit labels attached on the objects. Based on algorithms in biometrics, computer vision or pattern recognition, VIoT is the specialization of IoT in labelling of objects with name, identity, colour, shape or other attributes.

The lines of future development of this work can be several. The main task can be the improvement of the database. At this moment the database is populated with 200 pictures of 20 characters. The target could be the increase of the number of the pictures included in the gallery changing the size, which now is 180x200 pixels, but also light intensity, hue or contrast. The images contained currently in the database are related mainly to characters well known in the last three years on the news programs. Some of them should be replaced by some new or, in other cases, an update of the current photos could be made. The characters in the image change continuously and some characteristics could not more up-to-date.

The collection of videos was acquired, in some cases, more than two years ago. Now the hardware and software tools for recording have improved the performances so, also considering the improvement of transmission systems, we could have available better samples on which perform experiments. As we experienced in the last part of the work, the availability of images with higher resolution improves the performances of the recognition. The face images cropped and resized from the videos suffer less of problems of resolution in the matching phase. In our case all the images have been resized to 180x200 pixels but, while in the case of SD the starting size of the detected region was around 50-70 pixels per 80-90, in case of HD the initial size was about twice along each axis.

The commercial applications can be several and the industry is careful to the development of this field. Remembering the Minority Report scene where Tom Cruise was walking into a building and all the advertisements were immediately tailored, some companies as Tesco is experimenting sending tailored ads offering products and services [70].

While Google is experimenting also facial recognition apps for Google Glass [71], a company called Emotient is building tools for the headset in order to measure the human emotions and allow the shop assistants to offer better customer service [72].

Also Facebook with Deep Face [73] and Microsoft [74] with Kinect are working on this subject. The Facebook's technology promise an accuracy of 97.35% on the Labeled Faces in the Wild (LFW) dataset, reducing the error of the current state of the art by more than 27%, closely approaching human-level performance. Microsoft wants to take shopping into the high-tech, slightly dystopian future by deploying facial recognition cameras in retail locations to create a highly personalized in-store experience.

The path towards commercial applications is just at its beginnings and promises several innovative applications despite the privacy problems. For this reason, after receiving negative feedback from consumers and USA Congress, Google decided to slow down the plans to include facial recognition abilities in Glasses. Nevertheless some government agencies, as Homeland Security in USA, are investigating new facial recognition technologies in order to provide us with higher security levels. We are only at beginning of the arrival of new services.

6. REFERENCES

- [1] <http://www.nleomf.org/museum/news/newsletters/online-insider/november-2011/bertillon-system-criminal-identification.html>
- [2] M. Geuss, S. J. Purewal - Facebooks facial recognition flops. http://www.pcworld.com/article/230318/facebook_facial_recognition_flops.html, PCWorld, 2011.
- [3] Zhuoxuan Jiang, Yaping Lin and Stan Z. Li, 2013. Accelerating Face Recognition for Large Data Applications in Visual Internet of Things. *Information Technology Journal*, 12: 1143-1151.
- [4] <http://www.nist.gov/itl/iad/ig/face.cfm>
- [5] <https://requester.mturk.com/>
- [6] R. Singh and H. Ohm, "An overview of face recognition in an unconstrained environment," in *Procs. IEEE Second International Conference on Image Information Processing (ICIIP)*, 2013.
- [7] M. Fischer, H. Ekenel, and R. Stiefelhagen, "Interactive person reidentification in tv series," in *Procs. International Workshop on Content-Based Multimedia Indexing (CBMI)*, pp. 1–6, June 2010.
- [8] J. Sivic, M. Everingham, and A. Zisserman, "Who are you? – learning person specific classifiers from video," in *Procs. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.
- [9] M. Everingham, J. Sivic, and A. Zisserman, "Hello! My name is... buffy - automatic naming of characters in tv video," in *Procs. of the Workshop of British Machine Vision Association*, 2006.
- [10] M. del Pozo-Banos, C. Travieso, J. Alonso, and M. Ferrer, "Face identification based on tv videos," in *Procs. International Carnahan Conference on Security Technology*, pp. 119–125, Oct 2009.
- [11] R. Cinbis, J. Verbeek, and C. Schmid, "Unsupervised metric learning for face identification in tv video," in *Procs. of IEEE International Conference on ComputerVision (ICCV)*, pp. 1559–1566, Nov 2011.
- [12] M. Tapaswi, M. Bauml, and R. Stiefelhagen, "Knock! knock! who is it? probabilistic person identification in tv series," in *Procs. of IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [13] Satoh, S. and Kanade, T., "Name-it: Association of face and name in video," In *Proc. Intl. Conf. on Computer Vision and Pattern Recognition*, pages 368–373, (1997).
- [14] Yang, J., Chen, M. and A. G. Hauptmann, "Finding person x: Correlating names with visual appearances." In *Proc. Int. Conf. on Image and Video Retrieval*, pages 270–278, (2004).
- [15] Yang, J., Yan, R. and Hauptmann, A. G., "Multiple instance learning for labeling faces in broadcasting news video," In *Proc. ACM International Conference on Multimedia*, pages 31–40, (2005)
- [16] Zhai Y. and Shah, M., "Tracking news stories across different sources," In *Proc. ACM International Conference on Multimedia*, pages 2–10, (2005).
- [17] <http://www.vision.caltech.edu/html-files/archive.html>
- [18] <http://www.nist.gov/itl/iad/ig/feret.cfm>

- [19] Phillips, P. Jonathon, Moon, Hyeonjoon, Rizvi, Syed A., and Rauss, Patrick J. - The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 10 (October 2000), 1090–1104.
- [20] <https://www.sheffield.ac.uk/eee/research/iel/research/face>
- [21] <http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>
- [22] <http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/>
- [23] <http://www.ee.surrey.ac.uk/CVSSP/banca/>
- [24] Stan Z. Li, Anil K. Jain - *Handbook of Face Recognition* - Springer Science & Business Media, 15 mar 2005
- [25] P. J. Phillips, P. J. Flynn, T. Scruggs, K. Bowyer, J. Chang, H. Jin, K. Hoffman, J. Marques, J. Min, W. Worek - Overview of the Face Recognition Grand Challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2005)*.
- [26] G. B. Huang, M. Ramesh, T. Berg, E. Learned-Miller . Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Report 07- 49, Univ. of Mass., Amherst, Oct. 2007.
- [27] <https://www.bioid.com/About/BioID-Face-Database>
- [28] L. Wolf, T. Hassner, I. Maoz - Face recognition in unconstrained videos with matched background similarity. In *Proc. CVPR*, 2011.
- [29] Z. Jiang, Y. Lin, S. Z. Li, 2013 - Accelerating Face Recognition for Large Data Applications in Visual Internet of Things. *Information Technology Journal*, 12: 1143-1151.
- [30] Q. Jia, T. Fan Feng, Q. Lei - “Rfid technology and its applications in internet of things (IoT),” in *Consumer Electronics, Communications and Networks (CECNet)*, 2012 2nd International Conference on, 2012, pp. 1282–1285
- [31] Gartner Inc - "Forecast: The Internet of Things, Worldwide, 2013." – December 2013
- [32] W. Zhao, R. Chellappa - *Face Processing - Advanced Modeling and Methods*, Eds., Academic Press, 2006
- [33] S. Theodoridis, K. Koutroumbas - *Pattern Recognition* – Elsevier – 2009
- [34] I. T. Jolliffe, *Principal Component Analysis*, Springer, 1986.
- [35] S. T. Roweis, L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, no. 5500, pp. 2323–2326, December 2000.
- [36] M. A. Kramer - “Nonlinear principal component analysis using autoassociative neural networks,” *AIChE Journal*, vol. 37, no. 2, pp. 233–243, 1991.
- [37] J. Shawe-Taylor, N. Cristianini - *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.
- [38] M. Sharif, M. Javed, S. Mohsin - “Face Recognition Based on Facial Features,” *Research Journal of Applied Sciences, Engineering and Technology*, vol. 4, no. 17, pp. 2879-2886, 2012.
- [39] M. Turk, A. Pentland, “Eigenfaces for recognition,” *J. Cognitive Neuroscience* 3, pp. 71–86, Jan. 1991.

- [40] L. Sirovich, M. Kirby, "Low-dimensional Procedure for the Characterization of Human Faces," *Journal of the Optical Society of America A: Optics, Image Science, and Vision* , Vol.4, pp.519-524, 1987.
- [41] A. K. Jain, R. C. Dubes, *Algorithms for Clustering Data* . New Jersey: Prentice-Hall, 1988.
- [42] K. Fukunaga - *Introduction to Statistical Pattern Recognition* , second ed. Boston, MA: Academic Press, 1990.
- [43] M. Sharif, S. Mohsin, M. J. Jamal, M. Raza - *Illumination Normalization Preprocessing for face recognition - 2010 2nd Conference on Environmental Science and Information Application Technology*
- [44] M. Murtaza, M. Sharif, M Raza, J. Shah - "Face Recognition using Adaptive Margin Fisher's Criterion and Linear Discriminant Analysis," *International Arab Journal of Information Technology*, vol. 11, no. 2, pp. 1-11, 2014.
- [45] B. Schölkopf, A. Smola, K. Müller - "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation*, vol. 10, no. 5, pp. 1299-1319, 1998.
- [46] J. Lu, N. Plataniotis, N. Venetsanopoulos - "Face Recognition using Kernel Direct Discriminant Analysis Algorithms," *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 117-126, 2003.
- [47] V. Nhat, S. Lee - "Improvement on PCA and 2DPCA Algorithms for Face Recognition," in *Proceedings of Lecture Notes in Computer Science*, Berlin , pp. 568-577, 2005
- [48] J. H. Shah, M. Sharif, M. Raza, A. Azeem - *A Survey: Linear and Nonlinear PCA Based Face Recognition Techniques - Department of Computer Sciences, COMSATS Institute of Information Technology, Pakistan - The International Arab Journal of Information Technology*, Vol. 10, No. 6, November 2013
- [49] R. Cendrillon, B. C. Lowell, "Real-Time Face Recognition using Eigenfaces" in *Proceedings of the SPIE International Conference on Visual Communications and Image Processing* , Vol.4067, 2000, pp.269-276.
- [50] C. Beumier, M. Acheroy, "Automatic Face Recognition," in *Proceedings symposium IMAGING*. Eindhoven, The Netherlands, 2000, pp.77-89.
- [51] H. G. Wang, D. Q. Liang and Y. Tian, - "Extracting Face Features Using Corner Detection, Zernike Moments and Neural Network", *Journal of Xi'An JiaoTong University of China*, 1999, 33(12), pp 88-91.
- [52] S. Xie, S. Shan, X. Chen, W. Gao - V-LGBP: Volume based Local Gabor Binary Patterns for face representation and recognition. In: *ICPR*. (2008)
- [53] D.G. Lowe - Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60 (2004) 91-110
- [54] P. Dreuw, P. Steingrube, H. Hanselmann, H. Ney - SURF-face: Face recognition under viewpoint consistency constraints. In: *BMVC*. (2009)
- [55] A. Albiol, D. Monzo, A. Martin, J. Sastre - Face recognition using HOG-EBGM. *Pattern Recognition Letters* 29 (2008) 1537-1543
- [56] T. Ahonen, A. Hadid, M. Pietikainen - Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28 (2006) 2037-2041

- [57] P. Grother, M. Ngan - Face Recognition In Video Recognition Accuracy Quick Look – NIST- September 18, 2014 - Global Identity Summit, Tampa, Florida
- [58] P. Grother, M. Ngan - Face In Video Evaluation (FIVE)- Concept, Evaluation Plan, and API -NIST – October 3, 2014
- [59] <https://www.gov.uk/imagery-library-for-intelligent-detection-systems>
- [60] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in Procs. of IEEE Conference on Computer Vision and Pattern Recognition, 2001.
- [61] P. Viola, M. Jones, “Robust Real-time Object Detection”, Second International Workshop On Statistical And Computational Theories Of Vision – Modeling, Learning, Computing, And Sampling, Vancouver, Canada, July 13, 2001.
- [62] B. D. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. International Joint Conference on Artificial Intelligence, pages 674-679, 1981.
- [63] C. Tomasi and T. Kanade. Detection and Tracking of Point Features. Carnegie Mellon University Technical Report CMU-CS-91-132, April 1991.
- [64] J. Shi and C. Tomasi. Good Features to Track. IEEE Conference on Computer Vision and Pattern Recognition, pages 593-600, 1994.
- [65] L. Lenc, P. Král - Matching Methods for Automatic Face Recognition using SIFT - 8th Artificial Intelligence Applications and Innovations (AIAI 2012) Confence, Halkidiki, Greece, 27-30 September 2012, pp. 254-263.
- [66] S. Wang and A. Abdel-Dayem, “Improved Viola-Jones Face Detector,” in Procs. International Conference on Computing and Information Technology (ICCI), 2012
- [67] M. Leo, F. Battisti, M. Carli, A. Neri - Probabilistic person identification in TV news programs using image Web database - Paper 9019-14 – SPIE 2014 – San Francisco, California, USA- 2 - 6 February 2014
- [68] M. Leo, F. Battisti, M. Carli, A. Neri - Video news face retrieval based on Web image datasets – EMTC 2014 – Naples - November 13, 2014 – In proceedings
- [69] M. Leo, F. Battisti, M. Carli, A. Neri - Face retrieval in video sequences using Web images database - Poster 9399-33 – SPIE 2015 – San Francisco, California, USA- 10 February 2015
- [70] <http://www.theguardian.com/business/2013/nov/03/privacy-tesco-scan-customers-faces>
- [71] <http://www.forbes.com/sites/andygreenberg/2013/12/18/google-glass-face-recognition-app-coming-this-month-whether-google-likes-it-or-not/>
- [72] <http://emotient.com/>
- [73] <https://research.facebook.com/publications/480567225376225/deepface-closing-the-gap-to-human-level-performance-in-face-verification/>
- [74] <http://research.microsoft.com/en-us/news/features/kinectfacereco-103111.aspx>