Università degli Studi di Salerno

Dottorato di Ricerca in Informatica e Ingegneria dell'Informazione
Ciclo 33 – a.a 2019/2020

TESI DI DOTTORATO / PH.D. THESIS

Biometric systems in homeland security context

Paola BARRA

SUPERVISOR: **PROF. MICHELE NAPPI**

PHD PROGRAM DIRECTOR: **PROF. PASQUALE CHIACCHIO**

Dipartimento di Ingegneria dell'Informazione ed Elettrica
e Matematica Applicata
Dipartimento di Informatica

Contents

A Giulia.

Abstract

The mission of homeland security is ensuring the safety of living communities and protecting citizens from unforeseen events. In this research field, intelligent and advanced systems are extremely useful to prevent from tragic epilogues. Homeland security systems aim at supporting humans in those continuous and tiring activities of monitoring and detecting dangerous situations occurring in a surveilled area. Fatigue and distraction can reduce the human attention over time and be the cause of risks for safety and security. This thesis highlights the recent advances in this field and proposes some contributions on the state-of-the-art to deal with difficulties of the homeland security issues. The work focuses on a specific perspective view of the problem consisting in the use of biometrics to detect and recognize individuals. The biometric traits explored in this work are both hard biometrics, i.e. the face, and soft biometrics, i.e. the gait. Face is traditionally and widely used as a strong biometric trait for recognition and authentication. A reliable and robust face biometric recognizer is based on the assumption that facial features are good in quality and number. This is achieved when the face is detected in collaborative conditions and the pose is ideal to extract the features. The pose of the face is not always frontal therefore a preprocessing phase of facial recognition involves the estimation of the pose of an acquired face. As a contribution to the state-of-the-art of head pose estimation, three different methods have been proposed that encode the face thanks to the use of facial landmarks and extract the pose. The features extracted from the face can be both static and dynamic. With static facial features we extract information from

a face if its identity is not known. Dynamic facial features relate to lip movement and lip recognition. The landmarks that define the skeleton have been extracted from a series of videos of people walking; this made it possible to study the gait and classify people on the basis of gender and on the basis of their "cooperativeness", that is the aptitude to support the camera or to try to escape it. The results obtained and discussed in this thesis are strongly linked to the concepts of security, surveillance and trust and therefore may serve as insights to further explore the strengths and the limitations of software solutions applied to homeland security.

Chapter 1

Introduction

Biometrics has become an essential component of the most effective solutions for automatic person identification. A biometric recognition system is a system that exploits physical characteristics (such as fingerprints, iris, face, ear shape, etc.) and/or behavioral characteristics (ie, voice print, signature, writing, etc.) of a subject for his/her identification and recognition[1]. The physiological characteristics of an individual are quite stable, subject to only small variations over time. Behavioral ones can be influenced by the psychological situation of the individual and require constant updating. Biometric systems operate under the premise that these distinctive human characteristics can be effectively acquired through special sensors and represented in numerical form so that they can be processed, stored and, subsequently, compared.

1.1 Biometric features

Any morphological characteristic of a subject can be considered a biometric key for its recognition when it manages to meet the requirements of:

- **Universality**: each individual must possess that particular biometric characteristic. In practice this may not happen;

the minimum percentage of the population for which a requirement enjoys the property of universality is 99%;

- **uniqueness**: the trait must be different enough from individual to individual so as to be sufficiently discriminating. Ideally, two individuals should not share the exact same biometric trait;

- **permanence**: refers to the way in which the stroke varies over time. The biometric feature must remain unchanged over time. The degree of permanence of a trait has a strong impact on system design and long-term management of biometric data;

- **capturability**: the biometric characteristic must be capable of being acquired and quantitatively measurable. Furthermore, the collection of biometric data should be non-intrusive, reliable, robust and cost-effective.

- **performance**: requirement linked to the goodness of the technology used, the various stages of recognition must not be expensive in terms of time and space;

- **acceptability**: the acquisition procedure must be tolerated by the majority of the population;

- **elidibility**: the system must not be easily evaded, to prevent it from being cheated or misled.

Not all traits satisfy all requirements equally, the choice of one biometry over another depends on the nature and purpose of the biometric system in which it is to be used.

Some physical traits, such as the geometry of the hand, are more appropriate for authentication, others such as fingerprints, iris or face recognition are more suitable for identifying a subject as they are better able to discriminate the identity of an individual in a very broad context [2].

A further classification of biometrics is made according to their discriminating power; as in figure 1.1 biometrics can be classified into two categories:

- **Hard Biometric**: able to guarantee a unique identification of the subject. They allow two individuals to be distinguished in a marked way (strong uniqueness) and keep their measurability almost or completely unchanged over time (permanence). Hard Biometrics are the fingerprint, the iris and the ear.
- **Soft Biometric**: they do not ensure unique identification. There can be several subjects with the same so-called "weak" trait. Biometries such as voice, finger and hand geometry, in general those of a behavioral nature (gait, face dynamics, handwriting dynamics, etc.) are considered less unique and less stable. Most often these traits are used in association with each other, or with strong traits [3].

In a video surveillance context, some biometrics are more suitable than others. This type of acquisitions take place in an uncontrolled context and often without the knowledge of the subject, the most suitable biometrics are those without contact and which can also be acquired at a distance: for example the face, the gait and the way in which we speak. It is precisely these three biometrics that have been deepened within this thesis work.

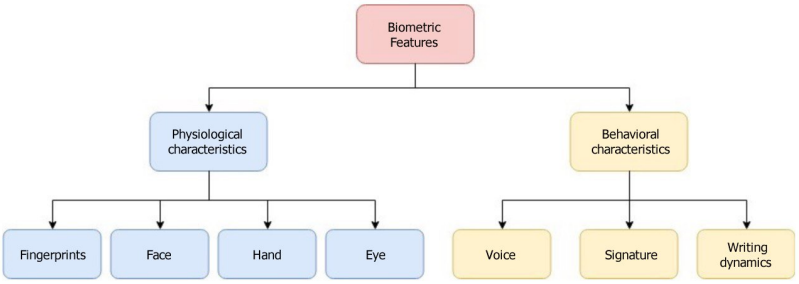


Figure 1.1: Some of the main physiological and behavioral biometric traits.

1.1.1 Homeland security context

Homeland security currently feeds undoubtedly one of the most explored lines of research and development; ports, airports, customs and border areas are extremely sensitive areas as they are crossed by millions of people every day. For this reason, intelligent and advanced systems are necessary for the protection of citizens and to prevent as much as possible unexpected events that could give rise to tragic epilogues. Traditional methods for recognizing violence in crowded and video-monitored environments today involve the intervention of a human operator who, through monitors, is required to personally and continuously check and recognize these situations. But human intervention in these contexts is not always effective for several reasons. The first reason concerns the need to have a human agent available at (almost) full time; the second reason regards the human error, as a distraction or misunderstanding of the situation detected on video. Furthermore, in a video surveillance context a human being can hardly pay the same attention to multiple monitors, so the risk of error is high. Consequently, attempts are made to integrate video surveillance systems more and more with artificial intelligence systems capable of intercepting anomalous situations.

1.1.2 Human recognition in video surveillance context

The goal of this research is to deduce from a frame or video information about an individual about his identity, context and behavior, in a nutshell: who he/she is, where he/she is and what he/she is doing.

- **Who is he/she?** An individual can be identified by a system through facial recognition techniques. This is possible if the individual is registered in the system and if the image available to us is suitable for recognition. From the video it is possible to extract a face in the best pose for recognition purposes. To do this techniques of head pose estimation

[4, 5] are used. In the head pose estimation problem, the reference points on a human face are detected and the pose is determined in terms of angle degrees in pitch, yaw and roll. In a video surveillance system of a shopping center, for example, a photo of the face and its pose in pitch, yaw and roll degrees would be stored for each individual, so that they can be recognized again if they reappear on the system.

- **Where is he/she?** Context analysis is a technique used to classify the context in which a scene is set. This is not a biometric technique because it is not based on purely human characteristics, but provides information that can allow us to more faithfully reconstruct what is happening and determine the choices of the system; combined with the identity and action performed by the subjects being filmed, for example, the presence of suitcase abandoned in an airport can determine the choice of alerting the police to prevent a risk of attacks.

- **What is he/she doing?** It is a very important and challenging problem to monitor and understand user behavior through the videos taken from various cameras; the study of this problem is called action recognition. The techniques mainly employed use Computer Vision. The starting point for understanding the action performed by an individual in a video is to estimate a sequence of static poses for which it is necessary to detect and locate the main parts/joints of the body (e.g. shoulders, ankles, knee, wrist etc.). Thanks to the existing state-of-the-art pose detection techniques it is possible to extract the coordinates relating to the skeleton of individuals within a frame and therefore the distances between the various parts of the body. This information, inserted in a sequence and given as input to a well trained recurrent neural network (RNN), can return the binary classification related to gender (male or female) [6, 7], to the action performed by the subject (standing, sitting, running, walking, arguing, etc.) [8] or the type of interaction between

different individuals.

These three pieces of information are necessary to allow the system to "understand" if what is happening is normal in relation to the context or requires the need for an emergency human intervention (if necessary, then alert the police, firefighters, ambulances or other).

1.2 Contributions of the Thesis

Our contributions to the state-of-the-art in the field of soft biometrics are related to the use of biometrics in homeland security by analyzing in detail the information that can be extracted in video surveillance videos. The first obstacle in recognizing a face from a video surveillance camera is the face pose estimation in terms of degrees of pitch, yaw and roll that represent the rotation of the face with respect to a frontal position.

Regarding the head pose estimation problem, the aim is to reduce the error, that is the difference between the classified pose and the real one of the input photo. The following problems have been solved for this issue:

- classification of the facial pose by a quad-tree coding. This solution led to the development of an algorithm that through a quad-tree coding of the face classifies the facial pose up to the reduction of the error to 4.07° in yaw, 7.51° in pitch and 5.50° in roll;
- identification of a reference system called "Spider-web", for the coding of the facial pose. This encoding enabled the development of a pose classification algorithm that further reduces the error to 6.21° in yaw, 3.95° pitch and 4.16° in roll;
- last contribution in the head pose estimation is the choice to use regression instead of classification to estimate the pose. Regression, compared to classification, better approximate

the pose estimation. The pose coding algorithm used is that of the "Spider-web" which has improved performance through logistic regression, reducing errors to 3.12° in pitch, 2.31° in yaw and 1.88° in roll.

To allow us to experiment with pose recognition algorithms and body movements, we have built a dataset of videos. This dataset, called Gotcha-I, contains videos of 62 subjects walking in a controlled context with different lighting conditions. From this dataset it was possible to carry out the following experiments:

- an algorithm for recognizing the biological gender of an individual (man / woman) based on the pose of the body. This allowed us to recognize genre from a single frame with 78% accuracy.

- a gender recognition algorithm based on the gait of the body. This led us to 82% accuracy in indoor video with the light off and with the camera flash.

- the development of a recurrent neural network (RNN) to recognize from the gait if the individual is non-cooperative or cooperative (that is, if he/she escapes the camera or not). This algorithm achieves 97.58% accuracy.

Finally, the last part of the thesis is dedicated to two case studies in a real-world environment. We created two applications for biometric recognition.

- The first one consists of a software application that acquires facial features from people faces; on request there is also the possibility of grouping faces that share the same characteristics. This software application can be used to tag the facial features of a large number of faces within a database.

- We then created a facial recognition application to recognize the identity of an individual from the dynamics of the face. This application was created for the control of personnel in a company; to enter the company building the person in

question is required to pronunce a given sentence in front of the camera, the system will allow the entrance if it recognizes the subject's lip dynamics corresponding to that sentence.

1.3 Outline of the Thesis

This work is the result of an industrial doctorate course. This path includes one year and a half of research at the University of Salerno, six months of research abroad and one year of research in the company Softlab. My thesis is based on this path and concerns three different aspects of the same topic:

- In this first chapter we have introduced the ideas behind biometric recognition in video surveillance.
- Chapter 2: focuses on estimating head position as a pre-processing step of facial recognition. This chapter introduces three methods that use face geometry to solve this task.
- Chapter 3: contains the study carried out in Spain in collaboration with the Universidad Las Palmas de Gran Canaria and focuses on gait analysis studies, in particular on how to detect the gender and cooperation of a person by analyzing the way in which he/she walks. The experiments of these works were carried out on the Gotcha-I dataset collected specifically to deepen these studies.
- Chapter 4: contains the research carried out with the Softlab company, and focuses on biometric recognition through facial features, this is expressed both through the results conducted on a specific study on static features and through a tool created specifically to recognize dynamic features such as the labial.
- In chapter 5 we draw concluding remarks and future research issues.

Chapter 2

Head Pose Estimation

Biometric recognition focuses on recognizing individuals by physical or behavioral characteristics. The face is one of the most widespread biometrics and one of the most accepted for both authentication and identification of the person. In particular, the face is one of the features that tend to be affected by lighting, pose and expression (PIE) and sometimes even by low image quality. Occlusions from scarves or sunglasses and non-frontal head pose are sources of problems [9]. These conditions can complicate face detection and recognition especially for video acquisitions at a distance or into the wild. The first studies were conducted under controlled conditions, therefore with faces captured with uniform lighting, a front pose, a neutral expression and a face free of occlusions. However, in order to have feedback with situations in the real world, research must tackle increasingly challenging problems. Faces captured in the real world are affected by critical factors, such as uneven lighting, natural and / or artificial occlusions or self-occlusion, and the subject's pose or expression [10, 11]. These factors play a particularly relevant role when dealing with unassisted acquisition (i.e. no expert operator guides the operation) and acquisitions in the wild. A relevant example is found in video surveillance [12, 13], which involves partially or even totally unaware subjects. These factors particularly affect the processing and extraction of features such as face, ears, periocular region and

iris.

Among the factors mentioned, the pose is probably one of the most difficult to deal with, whether it concerns the entire body of a subject [14, 15], or the head alone [16]. Head rotation can hide / distort the discriminating features of the face; partial self-occlusion of the face due to the pose can complicate further processing [17]. Furthermore, the possible rotations of the head around the x, y and z axes (pitch, yaw and roll, respectively), shown in Figure 2.1 [18, 19], cause inherent deformations in the geometric relationships between the element faces that are quite difficult to adjust [20]. The accuracy of recognition depends on how far the pose moves away from the neutral pose, in terms of elementary rotations with respect to each axis. In figure 2.2(a) examples of poses and the corresponding degrees of pitch, yaw and roll are shown. When a face is captured at a distance by video surveillance devices it is more common to find the face rotated in a combination of the three axes of rotation, rather than in the neutral pose. In these scenarios, on one hand there is a high probability that in a randomly selected image the captured face is not in a pose suitable for recognition, but on the other hand, there is also a high chance that in at least one frame of the captured sequence the face will be close to the neutral pose. With neutral pose we mean the pose close to 0° pitch, 0° yaw and 0° roll (see figure 2.2(b)).

We define *optimal frame* of a video an extracted frame in which the subject has the pose closest to the desired one. In video surveillance we may need to extract a specific pose of the subject to allow more accurate recognition.

The ability to select that optimal frame, possibly in real time, could improve the recognition performance, knowing the degree of head rotation acquired. In the following the existing Head Pose Estimation (HPE in following) methods that evaluate the head rotations by detecting two or three axes are discussed.

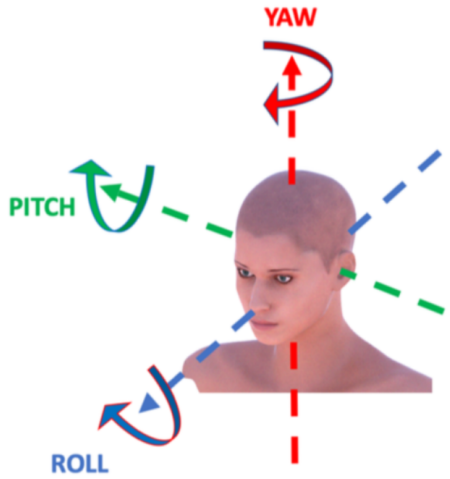


Figure 2.1: Axes of rotation of the head : Pitch, Yaw and Roll

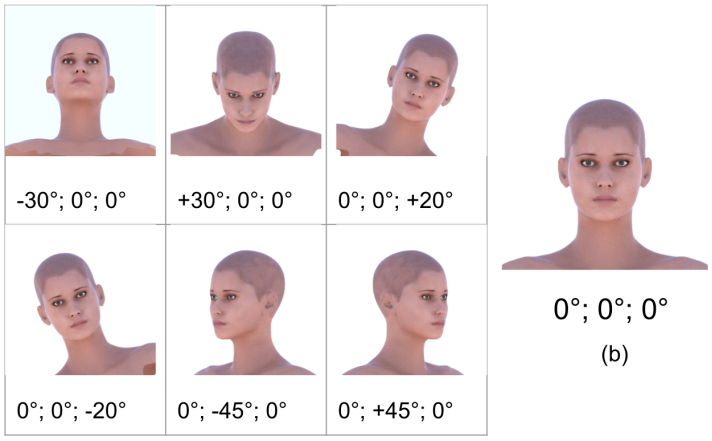


Figure 2.2: Examples of poses and the corresponding degrees of pitch, yaw and roll.

2.1 HPE state-of-the-art

Many HPE algorithms have been proposed over the years; here the techniques that have produced the best results at the state-

of-the-art are reported. When studying HPE, the most dividing choices are between the choice of 2D (intensity) or 3D (depth) images and the choice of whether to use deep learning techniques or use machine learning techniques (without neural networks).

Some methods are made for 3D images, known as RGB-D images; the D stands for "depth" image as, in addition to the color information, there is also the depth information. The use of 3D data implies the use of special sensors and cameras capable of capturing the subject and acquiring its depth; furthermore for this type of acquisition the operating distance between the camera and the subject is limited; for these reasons the use of the above methods in video surveillance contexts is very limited and the methods that use 3D images often use also 2D images. The following subsection describe the existing papers and the different choices adopted.

2.1.1 2D image methods

In the category of approaches working on 2D images there are many methods involving machine learning techniques in particular with the use of deep neural networks (DNN) and convolutional neural networks (CNN). The method presented in [21] estimates the horizontal and vertical alignment of the head (pitch and yaw) through a neural network. FSA-Net [22] is another method that estimates head pose based on the use of a neural network, which is based on regression and aggregation of characteristics. They propose learning a fine-grained structure mapping to spatially group features prior to aggregation. The method in [23] estimates head position by applying a deep neural network in a Coarse-to-Fine strategy. Two subnets are used jointly to classify the input image into four categories and then estimates the pose via a Fine Regression. The method in [24] uses the combination of two trained CNNs to identify both the head pose and the body pose; similarly, the head pose estimation approach in [25] uses information from the video scene to evaluate the orientation of the head using the direction of movement of the subject. QuatNet is a multi-

regression loss function applied in [26] to train a CNN to estimate head poses from RGB images with no depth information.

The proposal in [27] deals with a whole body estimation approach, consisting of three modules: the first module uses the HOG (histogram of oriented gradients) method to extract the characteristics related to the person's appearance; the second module updates a classifier with the person's tracking and direction information. Based on the direction in which he/she walks and the information of the first module, the third module estimates the orientation of the body, merging the information gathered from the previous modules. The authors in [28] analyze the region of the nose and, based on its orientation, they evaluate the pose of the face. The experiments show that this information has a high discriminatory power to determine the orientation of the head compared to the techniques that are based on the analysis of the entire facial region. In [29] and [30] through transfer learning approach two well-known neural networks are used: Multi-Loss ResNet50 and Hyperface. ResNet50 is used to predict the Yaw, Pitch and Roll angles of faces, directly from the image; Hyperface uses a CNN to detect the face, locate the reference points and estimate the pose. In [31] the authors address the face alignment problem with KEPLER: an iterative method for Keypoint Estimation and Pose prediction of unconstrained faces by Learning Efficient H-CNN Regressors (KEPLER) for addressing.

In [32] it is proposed the use of a combination of linear regressions that learns to map high-dimensional feature vectors extracted from the face bounding boxes on the head pose angles and the bounding box displacements, so that they are predicted in robust way also in presence of unobservable phenomena. In the method presented in [33] the pose estimation is formulated as a mixture of linear regression problems. The method maps the HOG-based descriptors extracted from the face bounding boxes to the corresponding head poses. The paper in [34] addresses the problem of estimating head position over a wide range of angles from low resolution images using chrominance-based functions. A linear auto-associative memory is obtained by training with the

Widrow-Hoff correction rule.

2.1.2 3D image methods

The majority of the existing solutions operate on 2D images, but 3D imaging has also been exploited. For example, [35] addresses the problem of head pose estimation from depth data. They synthesize a large amount of annotated training data using a statistical model of the human face. Experiments show that the approach is capable of handling real-world data presenting large pose changes, partial occlusions and facial expressions, even if it is trained only on synthetic facial data. In [36] 3DDFA (3D Dense Face Alignment) is proposed, which adapts a dense Morphable 3D model (3DMM) of a face to an image via cascading CNNs. In [37] a very large 2D dataset is synthetically expanded by converting the annotations of the 2D landmarks into 3D and unifying all the existing datasets, leading to the creation of LS3D-W. The method presented in [38] introduces a robust method in the case of variable lighting and rotation. Head pose is estimated from 2D key points drawn in two consecutive frames in the head region and their 3D projection on a simple geometric model. In the automotive field, [39] presents a solution for monitoring the driver's head. By combining 2D and 3D information, head position is estimated and regions of interest identified. They use this methodologies to detect special driver-related events such as drowsiness or inattention.

2.2 Head Pose Dataset

The Head Pose is a biometric trait closely related to the face. It is studied in the preprocessing phase of the face detection before carrying out the recognition. This technique is used on surveillance videos to extract the frame with a certain pose in terms of degrees of pitch, yaw and roll. A person's identity or facial features are easily recognized by whoever collects the dataset and labels the

data. However, an user cannot easily classify head rotations without special devices capable of gathering depth information. In the HPE, images that have RGB-D depth information in addition to the three color channels are preferred.

Table 2.1 shows the main characteristics of the most popular datasets. For each dataset popularity ("Pop") represents the number of recent HPE documents that used that dataset over the past five years, to the best of our knowledge.

Table 2.1: Characteristics of the most used datasets for HPE; (nd stands for "Not Declared").

Dataset	**Year**	**Type**	**#Subj**	**#Frames**	**Pop**
BIWI	2013	Depth+RGB	20	+15K	17
ICT-3DHP	2012	Depth+RGB	10	1400	6
SASE	2016	Depth+RGB	50	+30K	3
Pointing'04	2004	RGB	15	2790	12
AFLW	2011	RGB	20	25K	9
AFLW2000	2018	RGB	nd	2000	10
300W_lp	2016	RGB	nd	+61K	4
Gotcha-I	2020	Video	62	+137K	1
UPNA	2016	Video	10	36K	1
UBIPOSE	2016	Video	nd	+10K	1

In the following studies, RGB images belonging to three of the main datasets used to study HPE were used: Biwi Kinect Head Pose Dataset, the Annotated Facial Landmarks in the Wild, Pointing '04 and Gotcha-I.

2.2.1 Biwi Kinect Head Pose Database

The Biwi Kinect Head Pose dataset [40] comprises 24 sequences of 20 different subjects (14 men and 6 women, 4 people with glasses) recorded while sitting about one meter away from the sensor. All subjects rotated their heads trying to span all possible ranges of yaw and pitch angles, but also some roll is present in the data. To

label the sequences with the position of the head and its orientation, the data has been processed off-line with a template-based head tracker [41], as illustrated in Fig. 2.3. A generic template was used to match each person's identity as follows: first, a sequence of scans of the users' neutral face recorded from different viewpoints were registered and fused into one 3D point cloud as described in [42]; then, the 3D morphable model of [43] was used, together with graph-based non-rigid ICP (Iterative Closest Point) [44], to adapt the generic face template to the point cloud. Each sequence was thus tracked with the subject's template using ICP [45], obtaining as output for each frame the 3D location of the head (and thus of the nose tip) and the rotation angles. Over 15k frames have been annotated using such automatic method to acquire the ground truth for this database; the mean translation and rotation errors were around 1 mm and 1° respectively. The resulting Biwi Kinect Head Pose Database contains head rotations in the range of around ±75° for yaw, ±60° for pitch, and ±50° for roll. Faces are 90 x 110 pixels in size, on average. In addition to the depth data used for the tagging algorithm, the corresponding RGB images are also available, as shown in Fig. 2.4.

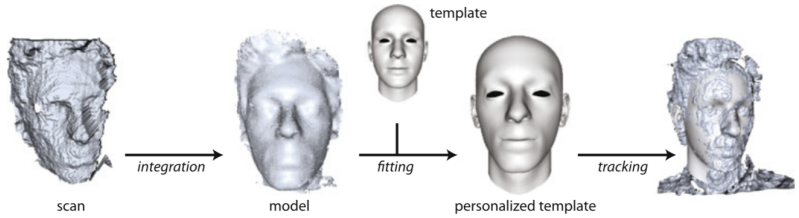


Figure 2.3: Automatic pose labeling. A user turns the head in front of the depth sensor, the scans are integrated into a point cloud model and a generic template is fit to it. The personalized template is used for accurate rigid tracking.

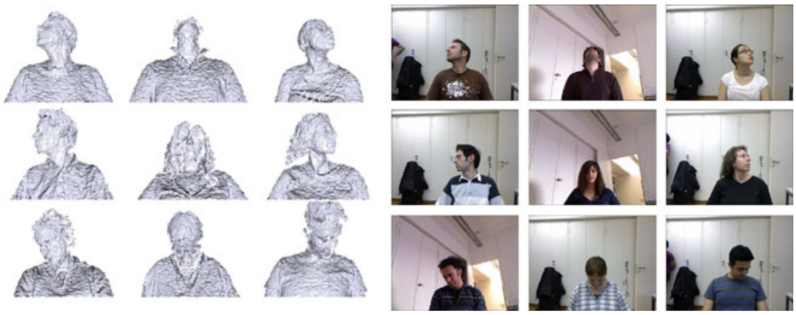


Figure 2.4: Example frames from the Biwi Kinect Head Pose Database. Both depth and RGB images are present in the dataset, annotated with head poses.

2.2.2 The Annotated Facial Landmarks in the Wild (AFLW)

Annotated Facial Landmarks in the Wild dataset (AFLW) provides a large-scale collection of annotated face images gathered from the social network Flickr, exhibiting a large variety in appearance (e.g., pose, expression, ethnicity, age, gender) as well as general imaging and environmental conditions. In total about 25k faces are annotated with up to 21 landmarks per image (figure 2.5).

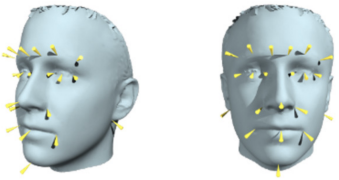


Figure 2.5: The points on the faces represent the 21 landmarks acquired for each image.

Of these faces, 59% are tagged as female, 41% are tagged as male; some images contain multiple faces. No rescaling or cropping has been performed. Most of the images are color and some of them are gray-scale imges. The facial landmarks are annotated

Figure 2.6: Some sample images taken from the AFLW2000 dataset, in different poses of the head.

upon visibility. So, if a facial landmark, e.g., the left ear lobe, is not visible no annotation is present. The database is not limited to frontal or near frontal faces. AFWL can be downloaded at [46].

2.2.3 Pointing'04

Pointing '04 Head Pose Image Database [21] is from the PRIMA Lab (INRIA) group. It includes 2,790 384 x 288 pixel resolution images from 15 subjects. This dataset is heavily researched for HPE regarding pitch and yaw, but does not include roll angle. During the acquisition, 93 post-its were placed in the room and the subject was asked to look at one post-it at a time by moving his head and not his eyes. The annotations shown are not precise as many subjects also moved their eyes to look at the post-its. Despite this, Pointing'04 is highly regarded and used in HPE research. Figure 2.7 shows a small subset of images from the Pointing '04 dataset.

2.2.4 GOTCHA-I

GOTCHA-I [**?**] contains videos of 62 subjects in 11 different environments, for a total of 682 videos. Each frame was extracted from a video of the subject's head starting from the right ear to

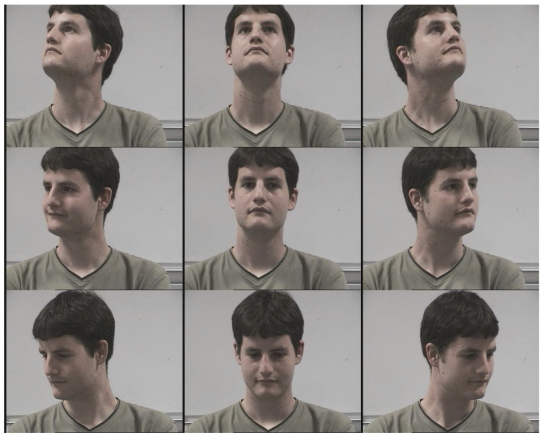


Figure 2.7: Some example images taken from the Pointing '04 dataset, in different poses of the head.

the left ear, framing the face (180° video), as in figure 2.8. From each 180° video a 3D model was built for each subject, and then obtained 2,223 head pose labeled images for each subject in the range of −40° and +40° in yaw and −30° and +30° in pitch and −20° and +20° in roll, with a step of 5°. The entire dataset has in total 137,826 labeled images.

Figure 2.8: Frames extracted from a 180° video of the Gotcha-I dataset.

2.3 Our contribution to Head Pose Estimation

In our studies we used an algorithm that extracts 68 facial landmarks [47] which is among the most used ones. This method makes a prediction of the position of the points on the face and even in presence of occlusions or low quality images, the output will always be composed of the coordinates of all the 68 points (none excluded). Furthermore, this method has excellent performance in the case of 2D RGB images. In figure 2.9 we see how the 68 face landmarks are positioned on a face.

Figure 2.9: 68 facial landmark and their positions in a face.

The structure of the landmarks was coded based on the poses represented by Lara, the synthetic dataset we created for this study. Lara is the name we gave to the 3D model of the face in the figure 2.10; using Blender, we extracted 2,223 poses of Lara at 5° degrees each. In this way we created 2,223 classes, each represented by a facial pose in a range of ±45° for Yaw, ±30° for Pitch and ±20° for Roll which represent the method's discrete search space. These ranges have been selected to reduce the search-space for practical considerations such as the statistical prevalence of the yaw rotation values compared to pitch and roll

values as well as the working limits of most face detection and facial landmark localization algorithms. Similar reasoning applies to the angular granularity of 5° adopted, which is reasonably small, however still visually significant. A smaller angular step would be barely noticeable, though it would have a significant impact on the efficiency of the method. It is worth to note that the proposed approach has no inherent limitations in terms of angular range and could work on large poses as well. Actually, apart from the considerations made above, the greater limit is in the landmark localization algorithm we used, that provides optimal results within limited angular ranges and suggested the current number of poses distributed on the three axes.

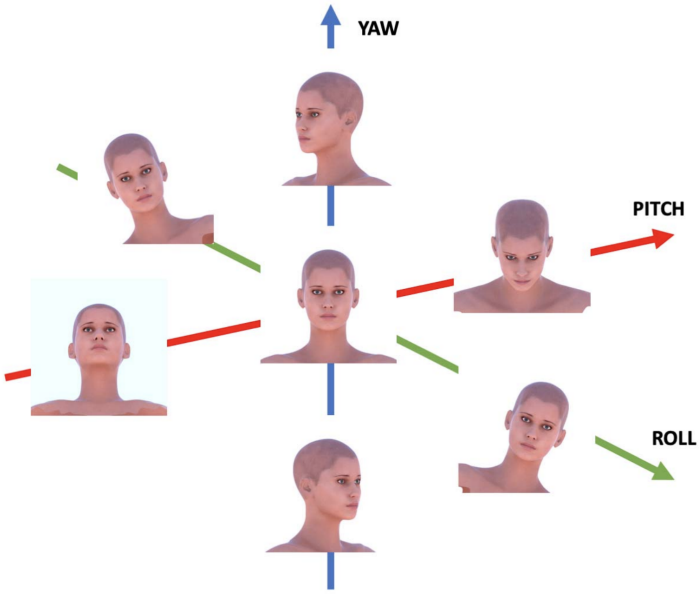


Figure 2.10: Variations on pitch, yaw, and roll in Lara Dataset.

2.3.1 QT PYR: The Quad-Tree approach

We have carried out several studies and experiments using the QuadTree to estimate HP [4]. The method consists of three main

steps:

- 1. Facial landmark extraction.
- 2. QuanTree decomposition.
- 3. Classification of Head Pose.

1. Facial landmark extraction. This task is accomplished using the method presented in [47].

2. QuadTree decomposition. The image containing the landmarks is recursively decomposed on the base, i.e. of the presence of landmarks. The image with 68-face landmarks is the root of the tree; since it contains at least one landmark, it is split into 4 sub-images. This methodology is applied recursively to each sub-image, so splitting each one in turn whenever it contains at least one landmark. This subdivision stops if there are no landmarks in a sub-image or if the sub-image is 4x4 pixels large, as shown in fig. 2.11. The QuadTree is then encoded in a binary array: a "1" denotes that the image has been split into 4 sub-images, a "0" denotes that the image has not been splitted. Fig.2.12 shows how the images are splitted recursively.

This encoding has the property that the resulting binary vector has a fixed length regardless of the encoded pose. The length of the vector is of 1.356 binary items.

3. Classification of Head Pose using the Lara model. This encoding is done for each Lara head pose, thus obtaining 2223 binary vectors. The QuadTree encoding described above is performed to test an image. The binary array resulting from the QuadTree encoding is compared to those in the dataset using the Hamming distance. Let us recall the definition of the Hamming distance. Given two strings s and t of length n, it is defined as

$$d(s,t) = \sum_{i=1}^{n} \delta(s_i, t_i) \tag{2.1}$$

where $\delta(s_i, t_i)$ is the following function

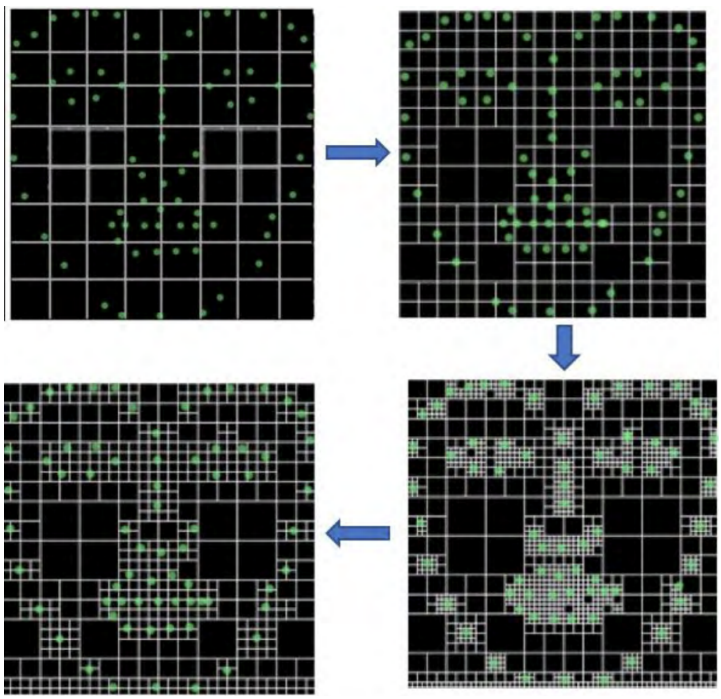


Figure 2.11: Example of four subsequent quad-tree subdivision steps from the coarser to the finer level.

$$\delta(s_i, t_i) = \begin{cases} 1, & \text{if } s_i \neq t_i \\ 0 & \text{if } s_i = t_i \end{cases} \tag{2.2}$$

This metric is particularly fast for our purpose. In our case the length of the strings is the length of the binary array; we will show that it does not require to be high to reach a good precision for HPE. For face extraction this is resized to 128x128 pixels. Considering the lower limit of 4x4 pixels for each QuadTree encoding, we obtain, for each pose, a binary array of size 1,365. Once the 2,223 arrays have been sorted, the comparison is performed as in a binary search. So instead of performing 2,223 comparisons for each image to be tested, the comparisons will be at most 8 ($log_2 1,356$); this reduces the computational time for each image to 0.044 seconds.

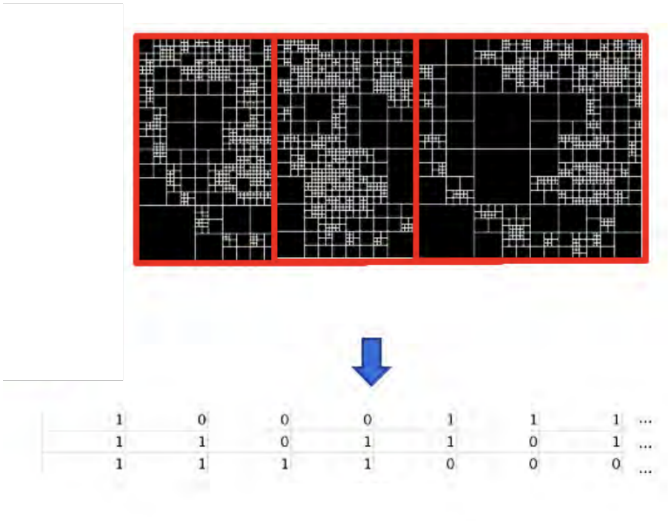


Figure 2.12: Each head pose is encoded as a binary array.

This method is named QT_PYR and its overall framework is shown in Figure 2.13.

2.3.2 QuadTree PY+R and PYR

The results of this method are presented for the datasets BIWI and AFLW in table 2.2. Given a face, it is classified in one of the 2,223 poses wrt the pitch, yaw and roll axes of rotation. Starting from QT_PYR, another methodology was experimented. This methodology, called QT_PY + R method, pre-processes the image by normalizing the face of the subject based on the roll rotation. The roll normalization is carried out starting from the measurement of the angular coefficient of the straight line passing through the two points represented by the external corners of the eyes. In Table 2.2 we can see the results of both the methods (QT_PYR and QT_PY+R) on the BIWI and AFLW datasets. Err_yaw, Err_pitch and Err_roll represent the differences in degrees between the predicted pose and the actual pose along the yaw, pitch and roll axes respectively. The MAE value is the Mean Absolute Error and rep-

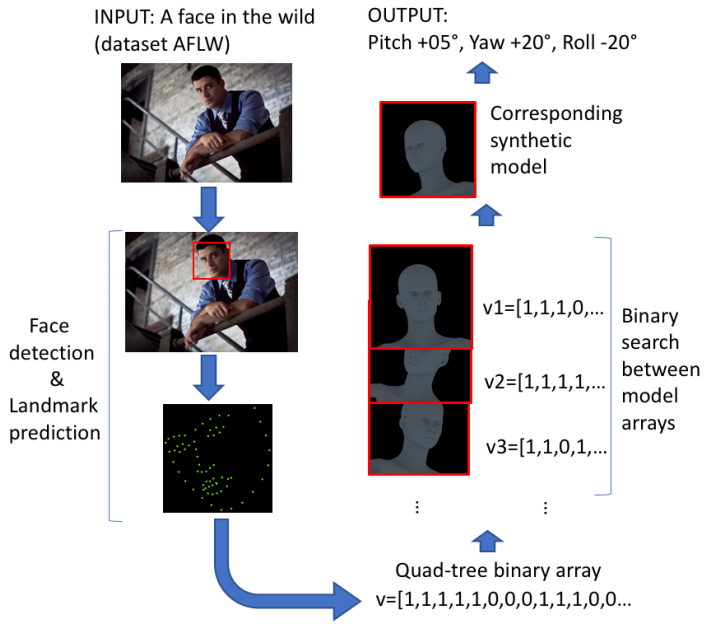


Figure 2.13: The QT_PYR workflow.

resents the distance between the predicted and the ground truth poses, as defined by the following equation 2.3:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \tag{2.3}$$

where y_j is the actual pose and $\hat{y}_j$ is the predicted pose, in our case the predicted angular value. We calculated the MAE for Pitch, Yaw and Roll separately and also an overall MAE of the error along the three axes.

In the results in the table 2.2 we can also observe that evaluating the roll separately does not improve the error in the estimation of the pose.

Table 2.2: Results of QT (QT_PYR and QT_PY+R) methods on BIWI and AFLW.

Dataset	Config	Err_yaw	Err_pitch	Err_roll	MAE
BIWI	QT_PYR	4.07	7.51	5.50	5.69
BIWI	QT_PY+R	6.28	14.95	4.12	8.45
AFLW	QT_PYR	7.6	7.6	7.17	7.45
AFLW	QT_PY+R	9.33	17.84	3.44	10.20

2.3.3 The Web-Shaped Model (WSM)

The Web-Shaped Model (WSM) [48] differs from the previous one in the image encoding with facial landmarks. In the proposed method a spider-web model is applied to encode the 68 facial landmarks. The spider-web is constructed by placing the center of a spider-web on the landmark corresponding to the tip of the nose (landmark 33).

Figure 2.14: The main steps of WSM.

The radius of the spider-web is given by the distance between the point O and the most distant landmark from the point O itself, i.e. the maximum distance to collect all the landmarks. Then the size of the spider-web adapts to the size and pose of the face.

Let us now introduce some terminology used in the Web-Shaped Model:

- by *circles* we mean the concentric circumferences that compose the spider-web (the red circles in fig. 2.15-a);

- by *quarter* we mean a quarter of the spider-web determined by a pair of Cartesian axes passing through the origin O centered on the nose (the part in blue in figure 2.15-a);
- with the term *slice* we refer to a slice of the spider-web delimited by two radiuses (the black radiuses in fig. 2.15-a);
- by *sector* we mean the shape delimited by two concentric circles and two radiuses (the green section in fig. 2.15-a).

Once we have defined the number of circles and slices that our spider-web has, the pose of the face is defined according to the number of landmarks that fall in each sector. For example, if we have a spider-web with 4 circles and 3 slices for each quadrant, we will have a total of 48 sectors = 3 (slices) * 4 (quadrants) * 4 (circles). The resulting array of this encoding will have length 48 and the values within the vector will correspond to the number of landmarks that fall in each sector, as shown in fig.2.16. The reading of the array from the spider web will proceed from the outside to the inside, as shown in figure 2.15-b.

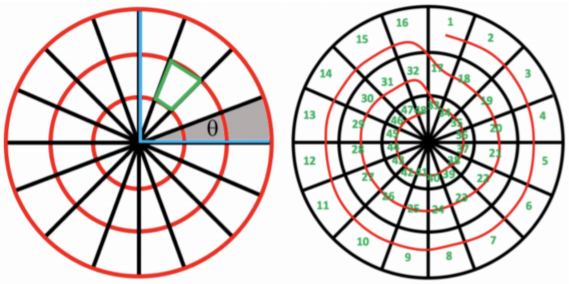


Figure 2.15: a) The structure of the spider-web; b) the numbering of the sectors.

Thanks to this technique, each pose is encoded in its corresponding array. The method was tested on 2,223 poses of the synthetic Lara model. Lara's poses were extracted with 5° deviations for the following ranges:

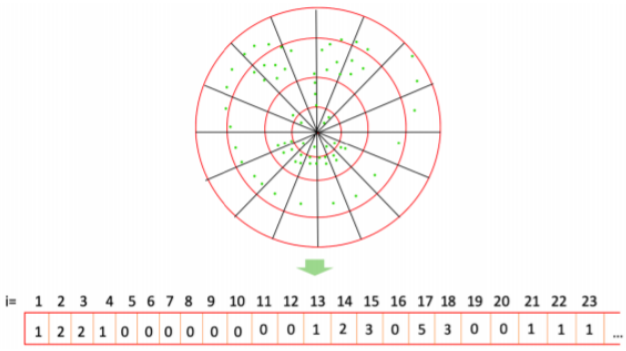


Figure 2.16: The image shows how the spider-web built on the facial landmarks turns into an array.

- pitch: [-30°, +30°]
- yaw: [-45°, +45°]
- roll: [-20°, +20°]

The model takes in input the images to be classified; then it encodes them using the spider-web method; finally the resulting encoding is compared with the Lara's images to extract the pose. The comparison is made using the Hamming distance. It is important that the images of Lara and the images to be classified are encoded with the same spider-web configuration. At the beginning of the algorithm the configuration and the number of slices, circles and sectors of the spider-web are fixed. In fact, a too high or too low number of circles and slices give worse results: a number too high of sectors could result in an excessive discrimination; on the other hand, a too low number could collapse more poses and be not sufficiently discriminating. Moreover, to understand which configuration gives the best results, several experiments with different configurations have been made. The best results have been obtained with the configuration with 4 circles and 4 slices.

The experiments were conducted on the BIWI, Ponting'04 and AFLW datasets introduced in Section 2.2. The results are shown

in the table 2.3

Table 2.3: Results of the WSM model. The Pointing'04 dataset does not contain the Roll information.

Dataset	**Err_yaw**	**Err_pitch**	**Err_roll**	**MAE**
BIWI	6.21	3.95	4.16	4.77
Pointing'04	10.63	6.34	\	8.485
AFLW	3.11	4.82	2.25	3.39

To evaluate the performance even in more competitive conditions, the method was also tested on the videos of the Gotcha-I dataset. Given a video of a walking person, the frame in which the face has the position closest to the neutral one was extracted (fig. 2.17).

Figure 2.17: Frames from video. The last frame reproduces the one chosen by the procedure using the proposed pose estimation. The same frame (the 15-th one) is highlighted in the sequence by a yellow rectangle.

2.3.4 WSM with Regression

In the previous method [48] we made a comparison between the pose feature vector extracted and those stored in the dataset to perform the pose classification, as in figure 2.18. The output obtained is the pose, whose reference vector has the lowest distance (Euclidean) from the extracted vector of the input image.

With classification the output class can only fall into one of the 2,223 classes. For example, if the pose is (0°, 3°, 0°), the classification can give us the class (0°, 0°, 0°) or (0°, 5°, 0°) as output, because the classes are a discrete number of values. Example in figure 2.18-B). In the classification of the output it is a continuous number of values in the range; so in the previous example the regression method is able to predict the exact degree.

Starting from this approach, we use regression instead of classification, to outperform results. So, for each experiment, three different regression models were built, for pitch, yaw and roll:

- a regression model for pitch prediction that returns a number in the continuous range [- 30, +30];
- a regression model for yaw prediction that returns a number in the continuous range [- 45, +45];
- a regression model for roll prediction that returns a number in the continuous range [- 20, +20].

The method is represented in Figure 2.18-C). In doing so, the minimum error can be less than 5° unlike the previous method that uses classification.

2.3.5 Supervised Learning: Regression vs. Classification

Supervised Learning (SL) refers to a class of algorithms that learns a function f that maps an input space X to an output space Y based on a sequence of input-output pairs. There are two main groups of Supervised learning (SL) methods, namely *classification* and *regression*, depending on the nature of the output space. Classification methods predict discrete responses and aim to assign a label $y_j \in Y$ at each input element $x_i \in X$. Regression models predict continuous responses [49]. Relationships between two variables are modeled by linear regression trying to fit linear equation to observed data. Consequently, classification techniques provide

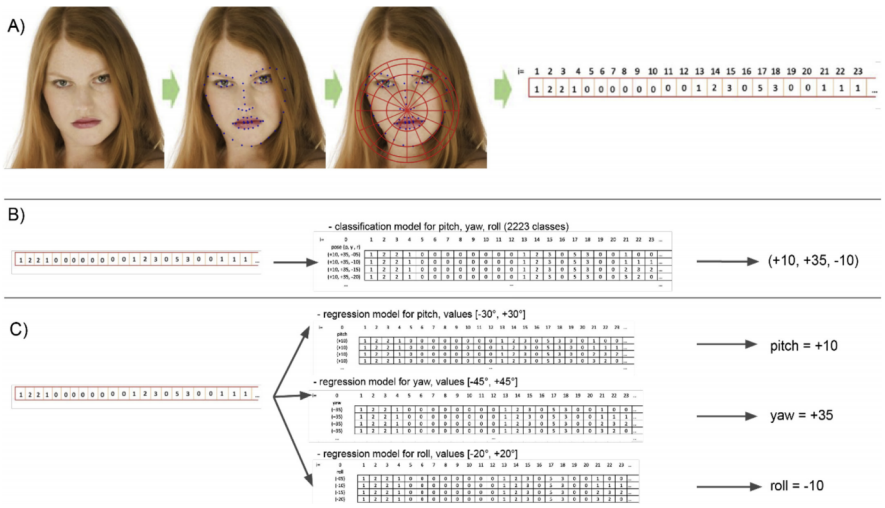


Figure 2.18: Representation of the method: A) Summary of the WSM approach; B) Classification; C) Regression.

the model or function that predicts new data in discrete categories; conversely, regression methods model functions at constant values, which means that it predicts data in continuous numeric data. Our approach stimulates the sensitivity of the regression methods to identify the head pose estimation. The goal is to predict the value of the dependent variable for the three angular values, respectively for pitch, yaw and roll axes associated to head's degrees of freedom, for which some information relating to the explanatory variables is available, in order to estimate the effect on the dependent variable.

2.3.5.1 Linear Regression

A linear relationship between an independent variable x, usually referred to as a predictor variable and a dependent variable y, i.e. a criterion variable, is expressed by the following equation:

$$y = mx + b \tag{2.4}$$

where m is the slope of the relationship and b is the y intercept. Linear Regression (LR) is employed to fit a predictive model to the set of training observations (x, y) [50]. The result is the prediction equation that gives the best estimate of y in terms of x. Then, the fitted model is used to make predictions of y for new instances of x.

2.3.5.2 Bayesian Ridge Regression

Bayesian Regression estimates a probabilistic model using regularization parameters in the procedure [51]. It assumes that the response y results from a probability distribution rather than estimated as a single value. Formally, to obtain a fully probabilistic model, the output y is assumed to be Gaussian distributed around X_w:

$$p(y|X, w, \alpha) = \mathcal{N}(y|0, X_w, \alpha) \tag{2.5}$$

where α is again treated as a random variable that is to be estimated from the data. A Bayesian view of Ridge Regression (BRR) is obtained in Eq. 2.6. The spherical Gaussian is adopted for the prior of the coefficient w:

$$p(w|\lambda) = \mathcal{N}(w|0, \lambda^{-1}, \mathbf{I}_p) \tag{2.6}$$

The priors over α and λ represent the gamma distributions. The parameters w, α and λ are estimated jointly during the fit of the model, the regularization parameters and being estimated by maximizing the log marginal likelihood [52].

2.3.5.3 Lasso Regression

Lasso (LsR) is a linear model that reduces the regression coefficients towards zero by penalizing the regression model with a penalty term called l_1 *norm*, which is the sum of the absolute coefficients [49]. Mathematically, the Lasso model is described by the following equation, in order to minimize the objective function:

$$min \frac{1}{2n_{samples}} ||Xw - y||_2^2 + \alpha ||w||_1 \tag{2.7}$$

The parameter α is a constant and $||w||_1$ represent the l_1 norm of the coefficient vector.

2.3.5.4 Logistic Regression

In the existing multiple regression models, Logistic Regression (LgR) represents a particular case of the generalized linear model. It is a regression model applied in cases where the dependent variable y is dichotomous [51], [49]. Therefore, LgR allows to analyze the relationship between a dichotomous variable and one or more explanatory variables (both continuous and categorical). In general, the Logistic model can be represented by the following equation:

$$y = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}} \tag{2.8}$$

where x is the input value, α and β the coefficients of the input value (constant real numbers) and y the predict value. Our implementation fit the model with *L2* regularization. More details about the LgR algorithm can be found at [53].

2.3.6 Results and Discussion

The four regression methods discussed in the previous section were used in the experiments.

The 70% of the datasets images was used to extract the model reference dataset, and the remaining ones are used for testing. To evaluate the proposed approach it is used a performance index commonly present as an evaluation criterion in HPE, namely the *Mean Absolute Error* (MAE). The MAE represents the distance between the predicted and the ground truth poses (MAE formula is presented in Section 2.3).

The results of our methods on the BIWI dataset have been compared with those known at the state of the art found in the

last rows of table 2.4; the results on the AFLW dataset are shown in table 2.5; instead table 2.6 shows the results of the experiments carried out on the Pointing'04 dataset. Also in table 2.7 we compare the computational time needed to carry out the experiments; it can be noted that WSM takes much less time than QuatNet but has worse results.

Table 2.4: Mean Absolute Error of Pitch, Yaw, and Roll Angles Across Different Methods over the Biwi Dataset

Method	Yaw	Pitch	Roll	MAE
Multi-Loss ResNet50 [29]	5.17	6.97	3.39	5.177
GPR [54]	7.72	9.64	6.01	7.79
PLS [55]	7.35	7.87	6.11	7.11
SVR [56]	6.98	7.77	5.14	6.63
hGLLiM [32]	6.06	7.65	5.62	6.44
FSA-Net [22]	4.27	4.96	2.76	3.99
Coarse-to-Fine [23]	4.76	5.48	4.29	4.84
QuatNet [26]	4.01	5.49	2.93	4.14
WSM [48]	6.21	3.95	4.16	4.77
QT PYR [5]	4.07	7.51	5.50	5.69
QT PY+R [5]	6.28	14.95	4.12	8.45
WSM-LR	3.63	3.44	2.15	3.07
WSM-BRR	3.61	3.35	2.11	3.02
WSM-LsR	3.63	3.36	2.16	3.05
WSM-LgR	**3.12**	**2.31**	**1.88**	**2.43**

2.4 Conclusions

In this chapter we have examined the algorithms behind head pose estimation and show our contribution to the state of the art. The algorithms presented use 2D RGB images so they can be captured by any type of camera; this choice is suitable for easily adapting these methods to video shooting in video surveillance contexts.To investigate further uses of the presented method, we performed

Table 2.5: Mean Absolute Error of Pitch, Yaw, and Roll Angles Across Different Methods over the AFLW2000 Dataset

Method	Yaw	Pitch	Roll	MAE
Multi-Loss ResNet50 [29]	6.470	6.559	5.436	6.155
Hyperface [30]	7.61	6.13	3.92	5.89
KEPLER [31]	6.45	5.85	8.75	7.01
3DDFA [36]	5.400	8.530	8.250	7.393
FAN [57]	6.358	12.277	8.714	9.116
QuatNet [26]	3.973	5.615	3.92	4.503
QT PYR [4]	7.6	7.6	7.17	7.45
QT PY+R [4]	9.33	17.84	3.44	10.20
WSM [48]	**3.11**	4.82	**2.25**	**3.39**
WSM-LR	3.88	**4.66**	2.50	3.68
WSM-BRR	3.82	4.67	2.49	3.66
WSM-LsR	3.86	4.69	2.58	3.71
WSM-LgR	4.31	5.34	2.62	4.09

Table 2.6: Mean Absolute Error of Pitch and Yaw Angles Across Different Methods over the Pointing'04 Dataset

Method	Yaw	Pitch	MAE
Stiefelhagen [21]	9.7	9.5	9.6
Gourier et al. [34]	12.1	7.3	9.7
SVR [56]	12.82	11.25	12.035
hGLLiM [32]	7.93	8.47	8.2
Probabilistic HDR [33]	8.70	8.85	8.775
Kong et al. [58]	10.98	9.71	10.345
WSM [48]	10.63	**6.34**	8.4
WSM-LR	5.61	7.73	6.67
WSM-BRR	5.60	7.68	6.64
WSM-LsR	5.61	7.61	6.61
WSM-LgR	**4.44**	7.55	**5.99**

tests on video sequences. The aim was to look for the one with the optimal pose in a sequence of frames. In this case the method

Table 2.7: QuatNet (training) vs WSM (training-free)

Method	GPU	Time	E_Pitch	E_Yaw	E_Roll	MAE
QuatNet [26]	NVIDIA GTX 1080	4.5h	4.32	3.93	2.59	3.61
WSM [48]	Intel HD Graphics 515	0.16h	4.82	3.11	2.25	3.39

was adapted to the purpose. Taking the video of an interview with an actress we asked the algorithm to extract a frame with the central pose (or the one with the pose closest to the one with coordinates 0° pitch, ° yaw, 0° roll), then we asked the algorithm to extract a certain pose from the video (or the closest one). In figure 2.19 the video of the interview, the query picture of the frame with the neutral pose (a) and the output frame in which the subject has the required pose (b). The second query requires to extract a frame with a specific pose (c), returns the frame in which the subject has the pose closest to the one requested (d).

Figure 2.19: Search for the desired pose in a sequence of frames extracted from a video interview (a) Search for an image that matches the frontal pose (the one with angle P: +00°, Y: +00°, R: +00°) (b) Front-most frame in the sequence (c) Search for an image that matches pose with angles: P +10°, Y +30°, R +05° (d) Frame more similar to the required pose.

A similar experiment was done on the Gotcha-I dataset. Given a video, the frame with the most frontal face was extracted, shown in fig. 2.17.

Chapter 3

Gait analysis

The work presented in this Chapter was carried out in collaboration with the Universidad de Las Palmas de Gran Canaria. In this chapter we delve into several aspects of biometrics that affect the way a person walks: the gait. Gait analysis is the study of human motion; this biometry has advanced with the rise of photography and cinematography that has allowed its acquisition. In recent decades it has been studied in medicine, sports and for biometric analysis. In medicine, gait abnormalities can be symptoms of diseases such as Cerebral Palsy or stroke. In sport, the study of gait can determine the choice of athletes' shoes, in order to optimize their performance. In biometric analysis, gait has the advantage that it can be acquired non-invasively by a video surveillance camera. Analysis of this biometry was conducted to explore what information can be acquired about walking subjects. In the course of this chapter we will see that thanks to the creation of a specific dataset it was possible to classify the genre and the cooperativeness of the subjects.

3.1 The state-of-the-art on gait analysis

In this section we provide an overview of the recent state of the art of the research on gait analysis. The gait is acquired both through sensors placed on the body of the subjects, called wearables, and through computer vision. Wearable devices acquisition in gait analysis are widely used in diagnosis or monitoring in the medical setting. For example, in the work in [59] the authors want to optimize gait acquisition by looking for the optimal positioning of sensors on the foot, varying five different positions. Internet of Things wearables [60] also performs gait recognition based on walking speed using the accelerometer. Walking speed has also been used in the medical field following disabilities caused by cerebellar ataxia [61]. In this case, the gait was acquired through sensors placed on the stomach and legs. Wearable gait also provides information on the health of the elderly, the authors in [62] implemented a method which, based on the age, gender and gait of the individual, determines whether the subject is healthy. Machine learning techniques used with wearable sensors involve Support Vector Machine (SVM) as in [63], K-Nearest Neighbor (KNN) as in [64, 65]. But the most popular methods concern neural networks and in particular Convolutional Neural Networks (CNN) [66, 67].

The wearable devices described above require the cooperation and awareness of the acquisition subject. In the context of video surveillance it is not possible to request the collaboration of all the subjects. Therefore it is preferable to analyze the video image of the subject, with the aid of computer vision techniques. In wild environments, such as in video surveillance, a person's gait can also be acquired from behind and still be meaningful [68]. Furthermore, the acquisition can take place through a normal RGB camera or from specific cameras to add depth information [69]. Depth information can be provided by depth cameras, multiple cameras or via Kinect [70]. When we acquire a person's gait, we cannot expect him to be in a front pose; as for the HPE, the

estimate of the pose [71] is also studied for the gait.

Figure 3.1 shows two types of gait that can be extracted from a 2d camera: Binary Silhouettes and Human Poses [72].

There are many studies to extract and analyze the binary silhouettes [73]. The Gait Energy Image (GEI) can be constructed from the human silhouette. GEI has been studied involving both small neural networks such as [74] and deep neural networks [75].

The human pose 3.1-b) extracted can be 2D or 3D and consists of creating the points of the human skeleton.

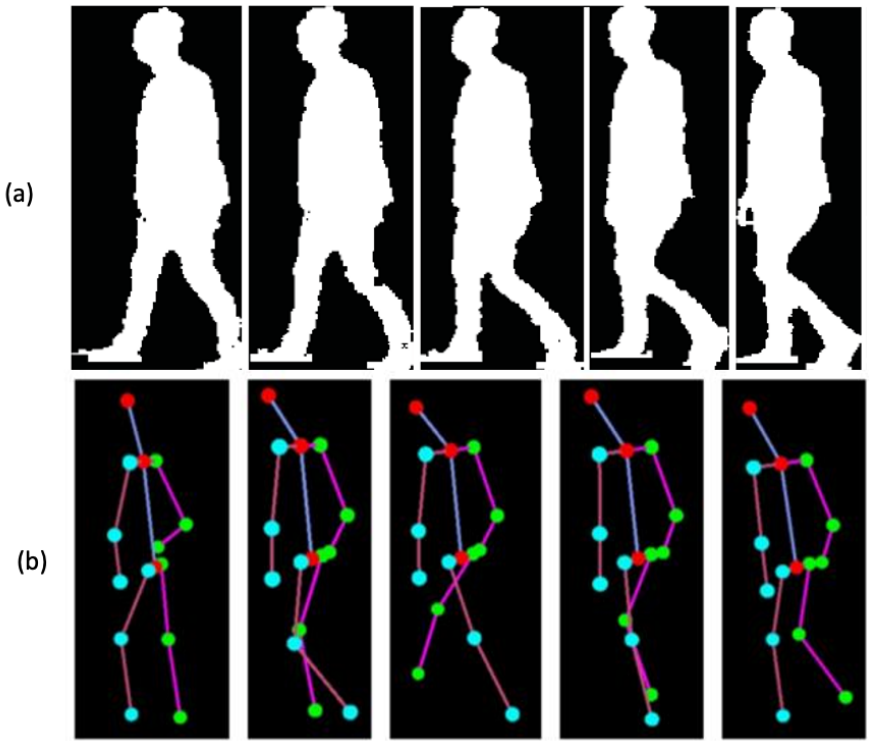


Figure 3.1: An examples of vision-based gait features. (a) the binary silhouettes and (b) the 2D human skeleton, extracted using OpenPose [76].

The software used in this study to extract the 2D human pose features is **OpenPose** [76]. This type of acquisition is less strong

in recognition than in identifying people by behavioral traits.

The human pose of OpenPose can be used in different areas. In the medical field: to prevent the fall of elderly people with senile dementia [77]; or assessing Parkinson's disease by gait as the second most dangerous neurological disorder [78]; or to detect differences between different issues by gender [79]. In the field of video surveillance, gender can be recognized for security purposes [80, 6, 7]. It is possible to classify the age of an individual based on the way he/she walks [81, 82]. In addition, the gait in video surveillance also allows us to classify [83, 84, 85] shares; acquired from video [86] or even from images transmitted by drones [87, 88]. As actions, intentions and behavior can also be analyzed through the gait; the study in [8] deepens the recognition of the cooperativeness and non-cooperativeness towards the camera of the framed subjects.

To carry out an in-depth analysis on these issues, it was necessary to create a special Gotcha-I dataset.

3.2 GOTCHA-I Dataset

The GOTCHA-I [**?**] dataset is a multi-environment dataset. This dataset has been partially presented in section 2.2.4. The Gotcha-I dataset was acquired at the University of Salerno, and was created to study the acquisition of features in people walking in different environmental contexts. Gotcha-I dataset contains 62 subjects: 15 women and 47 men. Thanks to the presence of both men and women it was possible to carry out a gender recognition on the gait [7] and on the human skeleton [6]. Dataset videos were captured in different ways: indoor with the flash camera, indoor with the artificial lights on and outdoor. Each of these contexts has also been acquired in cooperative and non-cooperative mode. In cooperative mode, the subjects were asked to walk normally without feeling bothered by the camera shot. In the non-cooperative mode the subjects were asked to act as if they felt annoyed by the presence of the camera, thus trying to evade the gaze. The dataset

contains a sample of data extracted from reality without changing its spontaneity, as the subjects were asked to keep glasses, hats or scarves if they were already clothing them. The main feature of the Gotcha-I dataset is that the videos have been acquired taking inspiration from body worn cameras; the intention is to simulate the wearable cameras used by police officers from different countries around the world.

All the videos of the dataset can be summarized as follows:

- (1) indoor with artificial light - cooperative mode;
- (2) indoor with artificial light - non cooperative mode;
- (3) indoor without any lights but the camera flash - cooperative mode;
- (4) indoor without any lights but the camera flash - non cooperative mode;
- (5) outdoor with sunlight - cooperative mode;
- (6) outdoor with sunlight - non cooperative mode;
- (7) 180°head video;
- (8) stairs outdoor - cooperative mode;
- (9) stairs outdoor - non cooperative mode;
- (10) path outdoor - cooperative mode ;
- (11) path outdoor - non cooperative mode;

Figure 3.6 shows the main differences between each video regarding the environment and the cooperation of the subjects.

Furthermore, the 3D model of people faces was extracted from the 180° videos (fig. 3.3) using photogrammetry techniques, from which the faces in all poses were synthetically extracted, so as to be able to study the HPE (fig. 3.4).

Table 3.1: Comparison of the contents of the datasets in the literature and of the Gotcha-I dataset: the number of subjects (Subj.), the analyzed biometries (Biom.), the type of environment (Envir.), the device used for the acquisition and information about non-cooperativeness (NC).

Dataset	Subj	Biom	Envir	Device	NC
COMPACT [89]	108	Face	Indoor	Camera	no
UBEAR [90]	126	Ear	Indoor	Camera	no
Quis-Campi [12]	320	Body	Outdoor	Camera	no
Droneface [91]	11	Face	Outdoor	Camera	yes
Mubidius-I [92]	80	Multi	Multi	Camera	no
Salsa [93]	18	Body	Multi	Multi	no
BIWI HeadPose [40]	20	Face	Indoor	Multi	no
GOTCHA-I [?]	62	Multi	Multi	Camera	yes

3.3 Gender from Gait

In this section we explore two studies to classify the gender from the human skeleton [6] and from the gait [7] of the subjects. This method consists of: the extraction of the body features with the OpenPose[76] algorithm; ii) the features creation starting from the OpenPose landmarks; iii) and finally the binary classification (male, female). In [6] the output of OpenPose (fig. 3.5), was studied to empirically establish which are the most important features for gender classification.

OpenPose estimates the position of 18 body landmarks in (x, y) coordinates. The body landmarks were extracted from all the videos present in the Gotcha-I dataset.

Features creation. The features have been created starting from the 18 OpenPose body landmarks (figure 3.6-a), with two different configurations: in the first configuration the distances between all the body landmarks (figure 3.6-b) were considered; in the second configuration, the distances between only the body landmarks of the upper part of the body (figure 3.6-c). The con-

Figure 3.2: Examples of frames extracted from the Gotcha-I dataset: outdoor with sunlight - cooperative mode (top left); indoor with artificial light - non coopeerative mode (top center); path outdoor - non cooperative mode (top right); indoor without any lights but the camera flash - non cooperative mode (bottom left); stairs outdoor - coopertive mode (bottom right).

figuration with the landmarks of the only upper part of the body was chosen for two reasons: because the landmarks of the whole body are not visible in all the frames, so often the lower part of the body is not captured by the camera; and because anthropomorphically the main difference between men and women is given by the different ratio between hip width and shoulder width [94].

We obtained 5,870 arrays of features, each representing a frame (3,970 arrays in cooperative mode and 1,830 arrays in non-cooperative mode). The number of arrays is unbalanced since it is not possible to extract landmarks from all the frames, because the whole body is not visible in all the frames.

Classification. The arrays with the distances were divided

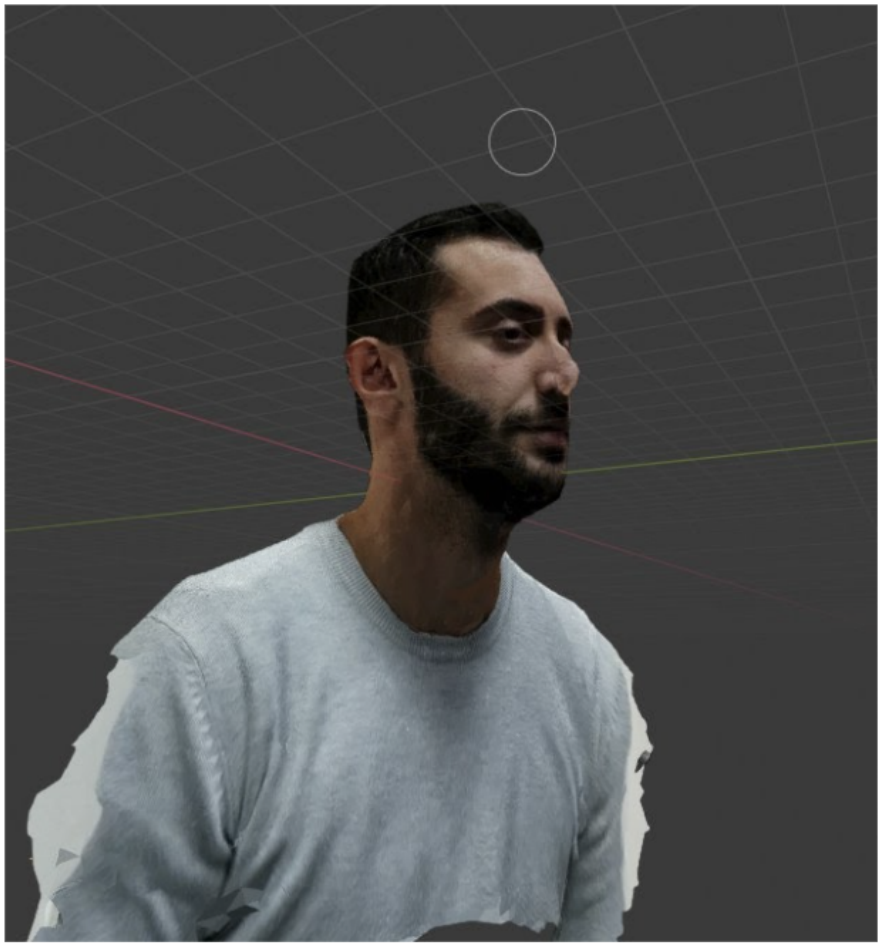


Figure 3.3: An example of the 3d model of the faces of the subjects in the 180° videos

into 70% for the training set and 30% for the test set and were given as input to a Random Forest (RF) classifier, with different depth configurations.

From the results in Table 3.2 the configuration with all the landmarks of the body gave the best results.

As it can be seen from Table 3.2, the possibility of using all the frames and of decreasing the computational time required to

Figure 3.4: The faces extracted in all poses from the 3D model of the face.

Table 3.2: Result of the first method based on Random Forest classifier.

RF Depth	**C. Acc.**	**Non-C. Acc.**	**C. and Non-C. Acc.**
Total body keypoints			
12	98.3%	55.4%	78.7%
Upper body keypoints			
4	83.3%	59.8%	73.7%

compute the distances, are paid with a less accuracy using only some upper body features. This preliminary study [7] was useful to empirically understand which configuration of features gives us better results in gender classification In the next study we created a method for identifying the gender from the gait; to this aim, we considered as input a video and not single frames. We examined the 18 body keypoints shown in figure 3.3 for 200 frames per video

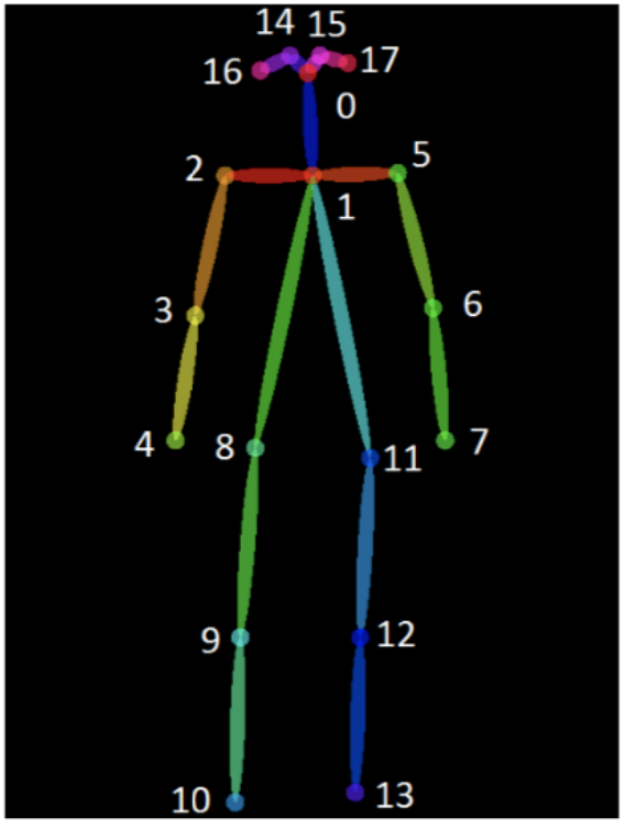
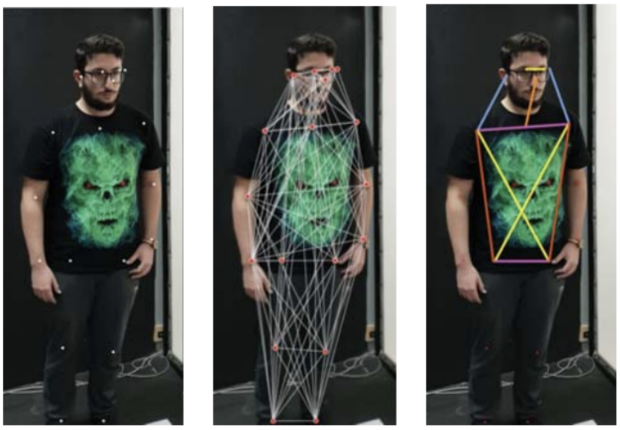


Figure 3.5: OpenPose is used to detect human body keypoints on single images. The output consists of 18 body keypoints estimation.

and calculated the 153 distances between all body keypoints; so for each video we obtained 36,000 features. For the gender classification we used 30 subjects (15 male and 15 female from the Gotcha-I dataset). As classifiers we used 4 different classifiers:

- Random Forest (RF): with 100 trees, entropy as a function to calculate the quality of the split in each phase and bootstrap samples during the construction phases of the tree.
- K-Nearest Neighbor (KNN): this algorithm is based on the similarity between the samples. K is the minimum number



(a) OpenPose data (b) Full body features (c) Upper-body features

Figure 3.6: a) The body landmarks of OpenPose on a subject of the Gotcha-I dataset. b) The distances between all the body landmarks. c) The distances of the body landmarks of the upper body.

of neighboring values to rank a sample, and a value of K = 5 was chosen.

- Support Vector Classifier (SVC): it create a hyperplane to separate the two classes involved.
- AdaBoost: converts the classification problem into simpler subproblems.

The 70% of the data was used for the training set, while 30% was used for the test set. Furthermore, the experiments were carried out in the cooperative and non-cooperative mode separately in order to highlight any differences between the two modes. The results are shown in table 3.3.

As shown by the results in the table, the non-cooperative mode always obtains lower results than the corresponding cooperative mode. The best results are given by the Random Forest classifier in cooperative mode - indoor with camera flash; for this mode the gender was recognized whit a success percentage of 82,5%.

Table 3.3: The gender recognition results using different classifiers, modalities and environments.

Classifier	Modality	Environment	Acc	Mean
RF	Cooperative	Indoor light	80.7%	
RF	Non-cooperative	Indoor light	75.5%	
RF	Cooperative	Indoor flash	**82.5%**	**75.45%**
RF	Non-Cooperative	Indoor flash	77.9%	
RF	Cooperative	Outdoor	68%	
RF	Non-Cooperative	Outdoor	68.1%	
KNN	Cooperative	Indoor light	69.1%	
KNN	Non-cooperative	Indoor light	65.8%	
KNN	Cooperative	Indoor flash	**74.1%**	67.36%
KNN	Non-Cooperative	Indoor flash	69.5%	
KNN	Cooperative	Outdoor	62.6%	
KNN	Non-Cooperative	Outdoor	63.1%	
SVC	Cooperative	Indoor light	74.1%	
SVC	Non-cooperative	Indoor light	66.6%	
SVC	Cooperative	Indoor flash	**77.7%**	69.06%
SVC	Non-Cooperative	Indoor flash	69.4%	
SVC	Cooperative	Outdoor	63.9%	
SVC	Non-Cooperative	Outdoor	62.7%	
AdaBoost	Cooperative	Indoor light	77.4%	
AdaBoost	Non-cooperative	Indoor light	72%	
AdaBoost	Cooperative	Indoor flash	**80.9%**	71.2%
AdaBoost	Non-Cooperative	Indoor flash	77.4%	
AdaBoost	Cooperative	Outdoor	59.7%	
AdaBoost	Non-Cooperative	Outdoor	60.2%	

3.4 Human cooperation detection

Given the information contained in the Gotcha-I dataset, it was possible to study cooperativeness detection: the problem was that of recognizing if a user was moving freely/naturally or trying to avoid the camera. The results obtained about this topic are also reported in the work in [8]. Figure 3.7 shows the pipeline of the

method for cooperativeness detection; it uses Recurrent Neural Networks, and its steps can be summarized as follows:

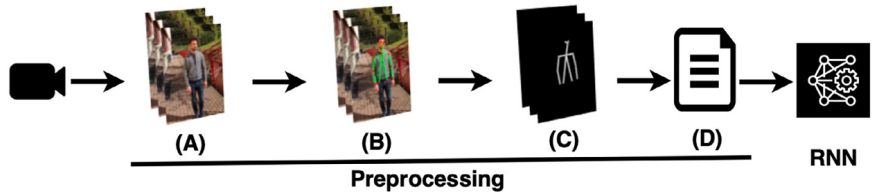


Figure 3.7: The pipeline of the cooperativeness detection method.

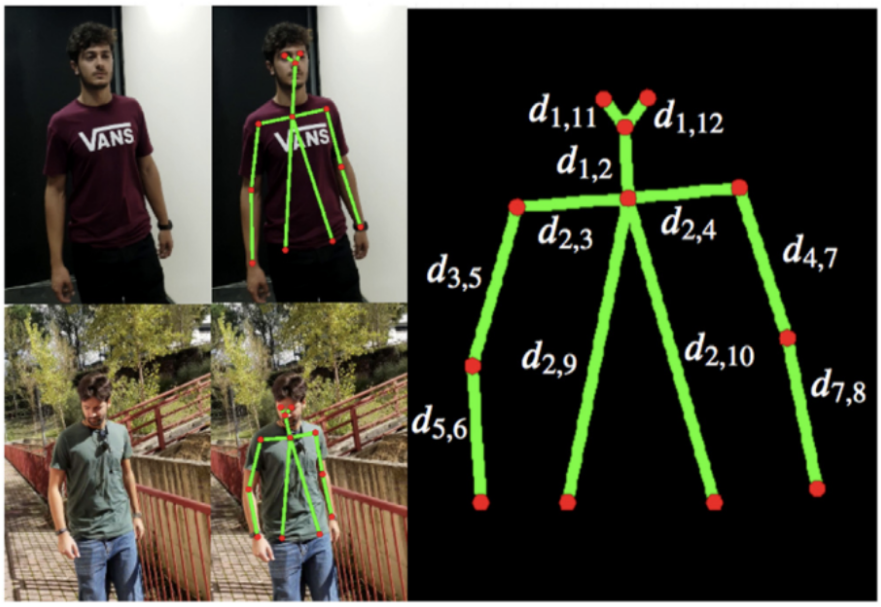


Figure 3.8: The features used refer only to the upper body.

- in the first step OpenPose skeleton representation is obtained: the features we are going to analyze are the body landmarks extracted with the OpenPose algorithm, as in the previous methods. In this case, however, we focus only on the upper body. As in the method for gender classification,

we start from the keypoints of the skeleton. In this case, as in the first gender analysis, we will only use the keypoints of the upper body, as shown in figure 3.8.

- in the second step the Bucket algorithm is invoked: this is the normalization phase of the data to be given in input to the network. This normalization phase serves to increase the training speed of the network.

- Attentive recurrent network: the Recurrent Neural Network (RNN) learns the long-term contextual dependencies. For the videos we solve this problem by using a multi-layer long short-term memory (LSTM). In this work we use a particular RNN combined with an LSTM, known as BiLSTM [95]. This network is bidirectional, so it can move both forward and backward. Usually not all human movements contribute in the same way, so we introduce a movement attention mechanism to capture the distinguished influence of the movement on cooperative/non-cooperative issues. If T denotes the number of time steps in the sequence, a_t the weights calculated at each time step t and h_t the hidden state vector, the attentions can be defined as

$$S = \sum_{i=1}^{T} a_t * h_t \tag{3.1}$$

the attention can be seen is a weighted average of h_t, and the resulting vector S is used to feed a fully connected layer to generate the final classification output.

The Gotcha-I dataset is the only video dataset in the literature that contains videos covering both cooperative and non-cooperative subjects. The training set and test set were randomly chosen and 5-fold validation was applied. Furthermore, the same subject was not used both in training and in the set to avoid creating bias in the model. To train the models, the metric chosen is the Matthews Correlation Coefficient (MCC), taking into account

true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). The MCC can be calculated as follows:

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TN + FN) * (TN + FP) * (TP + FN)}} \tag{3.2}$$

Unlike accuracy, in the case of MCC, if we have a binary class problem and it is balanced, the value will be 0, instead of 0.5, because MCC is between -1 and 1.

The experiments conducted take into account several variables. First of all, the number of buckets per video. A high number can reduce noise but, on the other hand, a low number can provide greater accuracy. Our architecture is compared to a basic LSTM, a CNN-LSTM, ConvLSTM (where the convolutional step is within the LSTM), BiLSTM and VGG16. The best results of our method were obtained for 100 buckets as reported in table 3.4 where we obtained for MCC a value of 0.952 with an accuracy of 97.58%. Also, with 50 and 150 buckets our method outperforms the other methods.

Table 3.4: Comparisons of Att-RNN with other methods.

Method	**MCC**	**Acc**
LSTM	0.909	95.43%
CNN-LSTM	0.877	93.82%
ConvLSTM	0.919	95.96%
BiLSTM	0.909	95.43%
VGG16	0.709	76.62%
Att-RNN (our)	**0.952**	**97.58%**

We also conducted experiments by fusing Att-RNN with other methods mentioned in the previous table. We got the results reported in table 3.5.

Table 3.5: Accuracies of the methods fused with Att-RNN.

Methods fused	**MCC**	**Acc**
Att-RNN ConvLSTM	0.957	97.84%
Att-RNN (100 buckets) Att-RNN (150 buckets) BiLSTM	0.963	98.12%
Att-RNN CNN-LSTM ConvLSTM (100 buckets) ConvLSTM (150 buckets)	0.969	98.39%

3.5 Conclusions

In this chapter we have analyzed the human gait. We presented the Gotcha-I dataset created specifically to study gait and HPE. From the Gotcha-I dataset we extracted the frames to study gender recognition on 2D human skeleton reaching 78 % accuracy thanks to the use of a Random Forest classifier. From the same dataset we extracted the information on 200 consecutive frames so as to be able to study gender from the gait. We achieved 82% accuracy with the Random Forest classifier for indoor cooperative videos with the lights off but the camera flash. We used more sophisticated techniques to solve the problem of recognizing the subjects' cooperativeness. We introduced ATT-RNN which has achieved approximately 98% accuracy by recognizing cooperative and non-cooperative users.

The information extracted could have a strong impact in the study of biometric recognition in video surveillance. This also opens the door to studies on action recognition.

Chapter 4

Face Recognition by facial features.

In this chapter we collect the work done in the year spent in the Softlab company, foreseen by the industrial doctorate. In this chapter we first make an introduction on biometric recognition systems in Section 4.1. We created two applications for biometric recognition. The first one consists of a software application that acquires facial features from people faces; on request there is also the possibility of grouping faces that share the same characteristics. The software application, presented in Section 4.2, can be used to tag the facial features of a large number of faces within a database. We then created a facial recognition application to recognize the identity of an individual from the dynamics of the face. This application was created for the control of outgoing personnel in a company; to enter the company building the person in question is required to pronunce a given sentence in front of the camera, the system will allow the entrance if it recognizes the subject's lip dynamics corresponding to that sentence (Section 4.3).

4.1 Biometric systems

In this section we examine in details the existing approaches and applications at the state-of-the-art of facial recognizers, both in

the case of recognition of facial features, and in the case of labial recognition.

The term "biometric system" refers to the set of technologies and applications (software) based on the use of biometrics and related characteristics. The main areas of interest of these technologies are the authentication and direct verification of personal identity and the indirect identification of a person by means of the biometric features available. An automatic recognition system is the set of methodologies and techniques to automatically identify objects and individuals, using information previously acquired and saved in a database. The results thus collected can then be analyzed and compared to carry out the recognition.

A typical biometric system has two operational phases (fig. 4.4):

- **1. Enrollment** : the phase that involves the acquisition of the biometric traits of an individual, and their processing in order to extract a series of features for generating the templates used for subsequent identification/authentication operations. The extracted templates are stored in a database.

- **2. Recognition** the phase when the system acquires the biometric traits of an individual, extracts a new set of characteristics, generates a template that will be compared with those in the database.

Environmental conditions can have significant effects on the performance of the devices and on the stability of the biometric characteristics extracted; these should be ideal in terms of noise of the acquisition devices (good lighting for video acquisitions, temperature, humidity, background noise for audio acquisitions, etc.). The user plays an important role in the design choices concerning the system: an user can be cooperative, if it is in his interest to be recognized by the system; or non-cooperative, if he/she is indifferent to the recognition process. The choice of which biometrics to use is therefore dictated by the requirements of the application

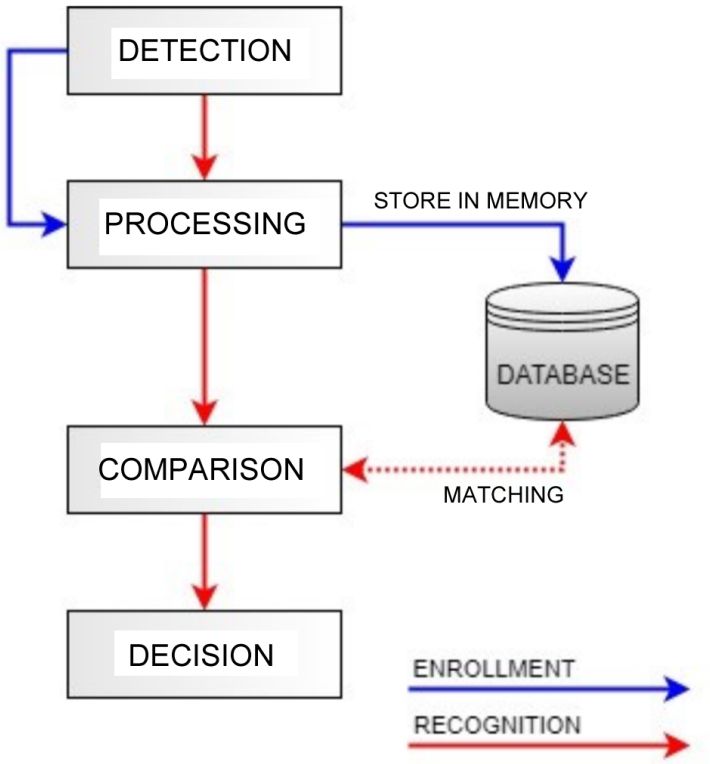


Figure 4.1: Operational Phases of a Biometric Recognition System.

in which it must be used; rarely one biometrics is optimal for any application context, but one biometrics may be more suitable than others for a specific case.

4.1.1 Facial recognition

Face recognition is a task that humans perform habitually and effortlessly in their daily life. The perception of faces, understood as the set of cognitive processes that induces the understanding and interpretation of the physiological characteristics of a face, is a ca-

Figure 4.2: In art, pareidolia has been widely exploited to create works that have multiple interpretations. For example, a landscape that looks like a face.

pacity that human beings develop from birth. It is a mechanism so rooted in the human brain that it often induces the phenomenon of Pareidolia: a subconscious illusion that tends to lead objects or profiles with a random shape to known shapes; this association is manifested especially towards human figures and faces (fig.

4.2). In Computer Vision, facial recognition becomes a method for identifying or verifying the identity of an individual using his face. Facial recognition systems can be used to identify people in photos, videos or in real time; these are non-intrusive methodologies that do not necessarily require the active participation of the subject. Despite the intrapersonal variations (the same subject can appear in several different ways) and inter-personal similarities (several subjects can resemble each other), the face as biometrics represents a good compromise between ease of acquisition (non-invasive) and performance (modern approaches with neural networks make recognition fast and accurate) as well as having excellent acceptance by users as it is the most natural method to associate identity with a subject. The most used approaches for facial recognition, according to [96], are based on the position and shape of facial attributes, such as the eyes, eyebrows, nose, lips and chin and the spatial relationships between them, or on the overall analysis of the face image representing a face as a weighted combination of a number of canonical faces. Facial recognition systems have the dual objective of identifying an individual or verifying his/her identity thanks to a series of discriminating biometric features present on the face. Therefore, they can operate in one or both operating modes, as in figure 4.3:

- **face verification (authentication):** implies a one-to-one correspondence that compares the image of the face in the query with the image in the database whose identity is claimed;
- **face identification:** implies a one-to-many correspondence that compares the image of the face in the query with all the images in the database in order to establish the identity of the subject to which the face belongs.

Another face recognition scenario involves a comparison with a watch list in which the face of the query is compared with a list of suspects (match one to many). In particular, a watch list can be of two types:

- **White list:** a list containing only the subjects admitted to

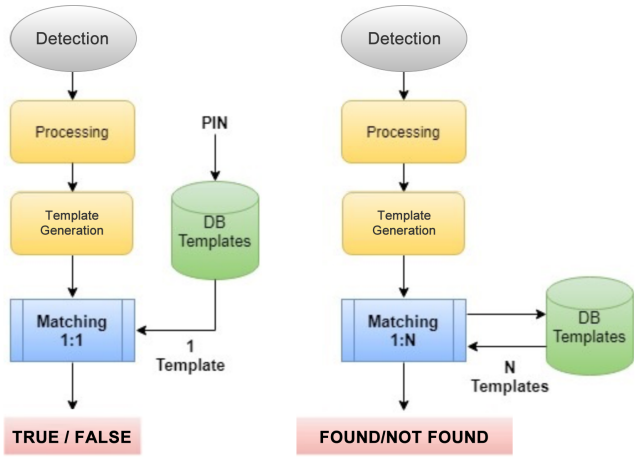


Figure 4.3: The two operting logics: a) authentication mode; b) identification mode.

the system. The membership of the individual on the list is verified before allowing access.

- **Black list:** a list in which the subjects excluded from the system are present. As soon as the system notices that an unauthorized individual is trying to access, an alarm is raised.

4.1.1.1 Facial attributes

Facial Attribute Analysis provides additional information where the face cannot be fully recognized. This may be due to an occlusion which may be voluntary such as wearing sunglasses, a hat or a scarf, or involuntarily due to poor lighting or environmental factors [97]. Being able to infer the identity of an occluded face is another vast field of research, the occlusions are also known as PIE (Pose, Illumination and Expression):

- Pose: the pose of the face concerns the inclination of the face

in relation to the camera. This problem has been addressed in chapter 2 and can lead to a distortion of the face and therefore to a misidentification.

- Lighting: capture lighting can create shadows or noise and create conditions that make it difficult to capture and locate facial features.

- Expressions: they reduce the possibility of a correct identification of the face because they distort its morphology.

- Occlusions: they partially occlude the face and can affect its recognition. Hats, glasses and scarves are occlusions.

With facial attributes we also mean all those characteristics that characterize the face such as: the color of the eyes, the color of the hair, the shape of the face, the absence or presence of make-up or beard and others. In section 4.2 we will deepen the study of these facial attributes within the CelebA dataset [98].

4.1.2 Structure of a Facial Recognition System

A facial recognition system generally consists of 4 modules: face detection and face alignment, features extraction and matching; detection and alignment are pre-processing phases performed before the actual recognition takes place, as shown in figure 4.4

- **1. Face detection:** the region containing only the face is extracted from the image. This region is segmented, associating a semantic meaning to each area of the face.

- **2. Face Alignment:** the alignment module is designed to obtain a more accurate localization and to normalize faces with respect to geometric properties (eg. size or pose) using a series of transformations (morphing). A further normalization can take place with respect to photometric properties, designed to alleviate the variability introduced by lighting or color (usually the images are shown in gray scale).

- **3. Feature Extraction:** once the face has been normalized, a series of distinctive features are extracted which allow the subjects to be effectively discriminated based on specific geometric or photometric variations. The extracted characteristics are collected in a data structure (vector of characteristics) capable of representing a specific subject.

- **4. Matching:** the vector of the extracted features is compared with those present in the database the identity of the face that matches with a certain degree of accuracy (if found) is output.

The performance of a facial recognition system is highly dependent on the characteristics that are extracted to represent the face model and on the classification methods used to distinguish between these models. The detection and normalization modules form the basis for a correct extraction of the characteristics.

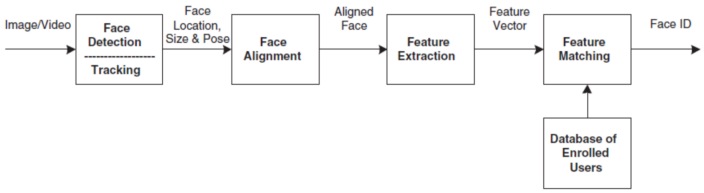


Figure 4.4: The four stages of a facial recognition system.

4.1.3 Performance evaluation

Unlike password-based systems, where a perfect match between two alphanumeric strings is required to allow access to a service, a biometric system rarely encounters two samples of the biometric trait extracted by the same user that translate exactly to the same set of features. This can be due to imperfect sensing conditions (e.g. noisy fingerprints due to sensor malfunction), alterations in the user's biometric characteristics (e.g. respiratory disorders affecting speech recognition), changes in environmental conditions

(e.g. inconsistent lighting levels in facial recognition) and/or variations in user interaction with the sensor (e.g. occluded iris or partial fingerprints). Consequently, it is rare to find two vectors of characteristics of the same subject that are perfectly identical [1]. In this case, a perfect match between two sets of features could indicate the possibility that a replay attack has been launched against the system. The observed variability in an individual's biometric characteristics is called intra-class variation, (see figure 4.5 for some examples), the variability between sets of characteristics from two different individuals is known as inter-class variation, in figure 4.6. A good set of features shows little intra-class variation and large inter-class variation.

In the following paragraphs we deepen some of the analyses used to evaluate the performance of the proposed system.

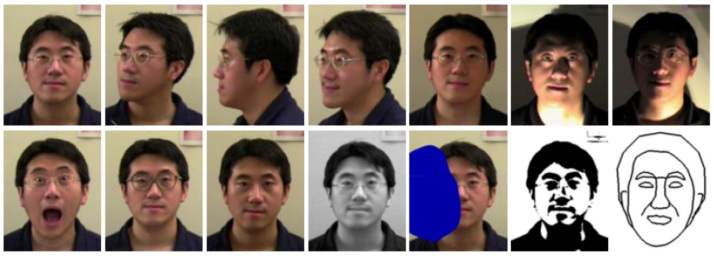


Figure 4.5: Intra-class variations of pose, lighting, expression, occlusion, color and brightness.[99]

4.1.3.1 FAR, FRR, EER

The goodness of the performance of a biometric recognition system is measured on the basis of two types of errors:

- FRR **(False Rejection Rate):** it represents the percentage of false rejections that leads the system to reject authorized users by mistake, failing to recognize;
- FAR **(False Acceptance Rate):** it represents the percentage of false acceptances: users who are not authorized are accepted by mistake.

Figure 4.6: Interclass variations: similarity of faces between twins and between a father and his son [99]

FRR and FAR are two inversely proportional quantities, as one decreases, the other increases. The FRR/FAR ratio is arbitrarily adjustable in any biometric system, of which, depending on the purpose and on the application context, it can be decided whether to increase or decrease its sensitivity. The degree of tolerance that one chooses to give to a system is defined through a threshold t designed to determine its goodness in terms of safety. As the degree of tolerance increases, there is an increase in the number of false acceptances, i.e. the FAR rises; with a low degree of tolerance there is a higher number of false rejections, i.e. the FRR rises. Once the variable t has been arbitrarily fixed, the functions $FAR(t)$ and $FRR(t)$ are constructed, which result to be respectively, as non-increasing monotone and non-decreasing monotone. Through these two functions it is possible to calculate the ERR (Equal Error Rate) which represents the intrinsic error of the system:

$$FAR(t') = FRR(t') = EER \tag{4.1}$$

that is, EER describes the point at which FAR and FRR assume the same value. t' represents the point where it is possible to adjust the ratio FRR/FAR, in fact, at point t' this ratio is equal to 1; for values $t > t'$ the ratio decreases, while for values $t < t'$ the ratio increases. Graphically it represents the point where the two monotonic functions (non-increasing for the FAR and non-

decreasing for the FRR) intersect; in fact, the EER also takes the name of CER (Crossover Error Rate) [100]. It follows that, depending on the specific case and the level of security required, the threshold is adjusted according to which of the two errors (FAR or FRR) is considered less serious or more acceptable than the other: for example, in a system for controlling access to a restricted area it is more prudent to maintain the risk of false refusals than to risk of allowing access to false positives. Generally false positives are considered more serious than false negatives; in real applications the tolerance threshold t often assumes a value lower than t', that is: $t < t'$ guaranteeing a reduced number of false acceptances.

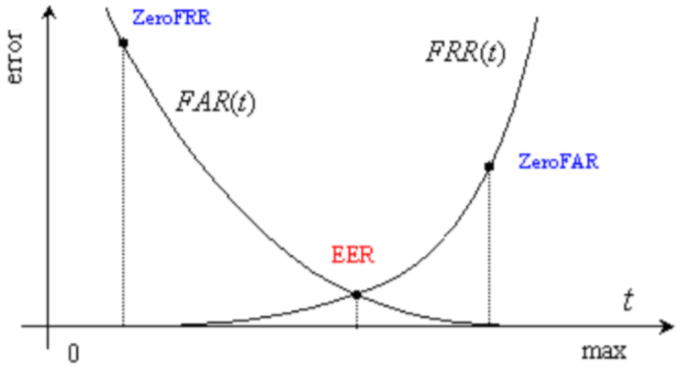


Figure 4.7: Graphic representation of the Equal Error Rate

4.1.3.2 Accuracy, precision and recall

For a more accurate assessment of biometric systems, other performance measures are also used such as accuracy, precision and recall that refer to the methodologies applied in the actual recognition phase. One of the most used metrics in performance evaluation is certainly the Accuracy. Informally, accuracy is the fraction of forecasts made correctly. Formally, it is defined as follows:

$$\textbf{Accuracy} = \frac{number\_of\_correct\_predictions}{total\_number\_of\_predictions} \tag{4.2}$$

Accuracy can also be calculated in terms of false/true positives and false/true negatives:

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.3}$$

Where:

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative

Accuracy alone is not enough to define the goodness of a 360-degree model, especially when dealing with an unbalanced data set, i.e. where there exists a disparity between the number of examples belonging to a class rather than to another one. It therefore becomes necessary to introduce two other measures: Precision and Recall; Precision attempts to answer the following question: What percentage of positive identifications is actually correct? It is formally defined as:

$$\textbf{Precision} = \frac{TP}{TP + FP} \tag{4.4}$$

Recall, on the other hand, tries to answer the question: What percentage of actual positives was correctly identified? Mathematically, the Recall is defined as follows:

$$\textbf{Recall} = \frac{TP}{TP + FN} \tag{4.5}$$

4.1.3.3 ROC curve and AUC

For a more complete evaluation of the performance of a system, we rely on the Receiver Operating Characteristic curve (ROC in the following), a graph showing the performance of a predictive model at all classification thresholds. Commonly used with ensemble, it plots the following two parameters:

- **TPR** (True Positive Rate), synonymous with Recall, is defined as follows:

$$\mathbf{TPR} = \frac{TP}{TP + FN} \tag{4.6}$$

- **FPR** (False Positive Rate) is the ratio of negative instances that are incorrectly classified as positive:

$$\mathbf{FPR} = \frac{FP}{FP + TN} \tag{4.7}$$

In other words, the relationships between "true alarms" and "false alarms" are studied. The higher the Recall or TPR, the more false positives FPR the classifier produces. The dashed line represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible (towards the upper left corner). One way to ascertain the accuracy of a classifier is to measure the area under the curve (AUC). A perfect classifier will have a AUC equal to 1, it is clear that the greater the area, the better the performance [101].

4.1.3.4 Confusion matrix

One of the most visually intuitive methods for evaluating the performance of a self-learning model is probably the confusion matrix. The general idea is to count the number of times the instances of class A are classified as class B [101]. Each row in the confusion matrix represents the effective class, while each column represents the predicted class, the parameters of true positive/negative and false positive/negative (respectively TP, TN, FP, FN).

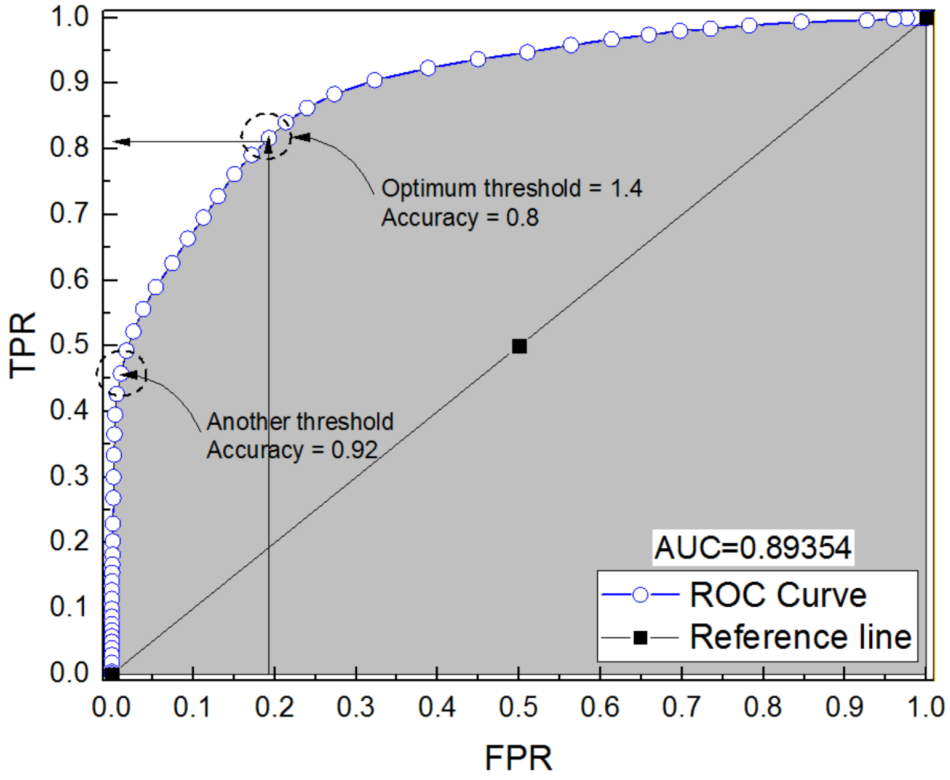


Figure 4.8: Example of ROC curves and AUC curves

4.2 Clustering Facial Features

We present a method of clustering of facial features, discussed in [102]. This method consists of a pre-trained neural network which is able to group people faces based on facial characteristics. The dataset on which the experiments have been performed is CelebA, presented later in this section

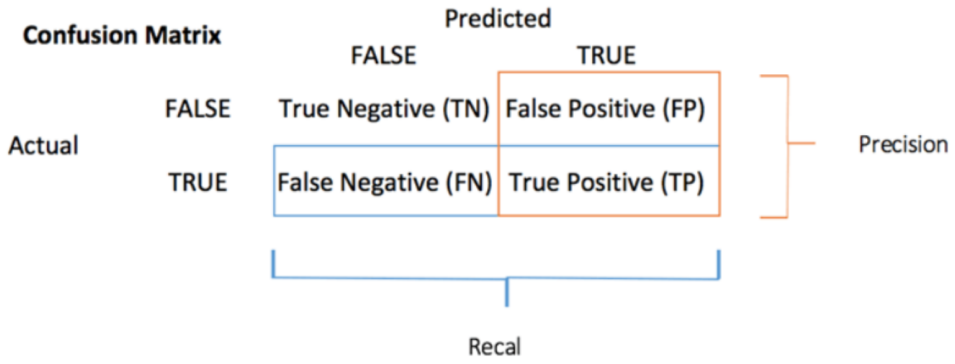


Figure 4.9: Structure of a confusion matrix

4.2.1 State-of-the-art in clustering facial features

In this section various aspects of the state-of-the-art of facial feature clustering are discussed. The related areas that are explored concern the prediction of facial attributes, clustering methods and transfer learning. These are the main issues addressed in this part.

4.2.1.1 Attribute prediction

Attribute prediction leads facial recognition in case of missing information or non-recognition. From the structure of the face, it is possible to reconstruct the missing information thanks to the geometric proportion of the facial features [103]. The authors in [**?**] have proposed a deep learning framework for predicting facial attributes in the wild; this framework uses two CNNs in cascade. These two networks are pre-trained differently: one locates the face and the other one predicts its facial attributes. The works in [104, 105], tackle the problem by learning the discriminating representation of the face. The authors in [105] implement a CNN to study angularly discriminative features. The idea is to exploit this type of features to study the intra-class and inter-class distances of the faces. In [104] it is presented FaceNet, a system

that synthesizes each image in a Euclidean space to which a similarity measure of the face belongs. This method facilitates the recognition and clustering of faces. Both of these methods offer a synthetic representation of facial attributes, known as face embedding.

4.2.1.2 Clustering methods

The clustering techniques considered in this study are K-means, Agglomerative Clustering and DBSCAN. K-Means [106] creates n groups of equal variance by minimizing the distance between the points of the same cluster. The number of clusters to be formed is decided a priori; it is therefore established empirically. The agglomerative clustering [107] approach creates clusters by building a hierarchical tree structure (dendrogram), in fact this type of method falls within the hierarchical clusters. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [108], unlike K-Means, arbitrarily decides the number of clusters and creates clusters based on density. Any point reachable from a point in the cluster must belong to that cluster. The goal of this work is to choose a priori the number of clusters based on the number of facial features we want to extract. The choice falls inevitably on the K-Means algorithm, because it is the only one that allows us to choose the number of clusters a priori.

4.2.1.3 Transfer learning

The complexity of the Machine Learning tasks increases as the research goes further. Consequently, the model architectures tend to become particularly big together with both time and computing demanding. In turn, this implies the need for enormous processing power and longer training time duration. This point is particularly true for recent Convolutional Neural Networks models [109], which require a huge amount of data and computational power. Thanks to the ImageNet classification challenge [110], the submitted AlexNet model [109] marked a turning point in 2012 for deep learning in computer vision. The models that followed, like

VGGNet [111], InceptionNet [111], and ResNet [112] are examples nowadays very used and useful as solvers for a wide range of computer vision tasks. The success and the accuracy achieved by these models have assessed, over the years, the tendency of using them as feature extractors, rather than as a solution for classification or regression problems. The Transfer Learning [113] has so achieved huge consideration, for the benefit of using pre-trained models like off-the-shelf solutions which do not required to be trained from scratch. Thus, recycling a model trained for a specific task on a new similar task reduces significantly the overall training time to cope with the new problem. In this work, this technique is exploited to fine-tune the proposed model.

4.2.2 Our approach

The following work can be summarized as follows:

- **preprocessing:** data on 37 facial features taken from the CelebA [98] dataset are preprocessed to be fed to the model;
- **clustering:** the output from the model above is used to label the clusters that will be created by the clustering algorithm;
- **analysis and results:** the attributes results from each cluster are used to calculate the accuracy of the method and the percentage of membership to a cluster of each face.

4.2.3 The CelebA dataset

The CelebA dataset [114] was chosen for the experiments. This dataset contains more than 200,000 celebrity face images tagged with 40 different facial features; some examples are shown in figure 4.11. CelebA, unlike similar LFW [115] or UTKFace [116] datasets, is very demanding due to the diversity of the features it contains. As it can be seen from figure 4.10, the dataset is vast but not balanced and there is a wide variety of environmental factors

and face types for pose, age, gender, facial expressions, occlusions, and so on. One third of the attributes are extremely rare facial features (10 % frequency or less), and only a couple of them are very common (they occur in more than 70 % of cases). The imbalance just mentioned negatively affects many loss functions significantly if adopted during training.

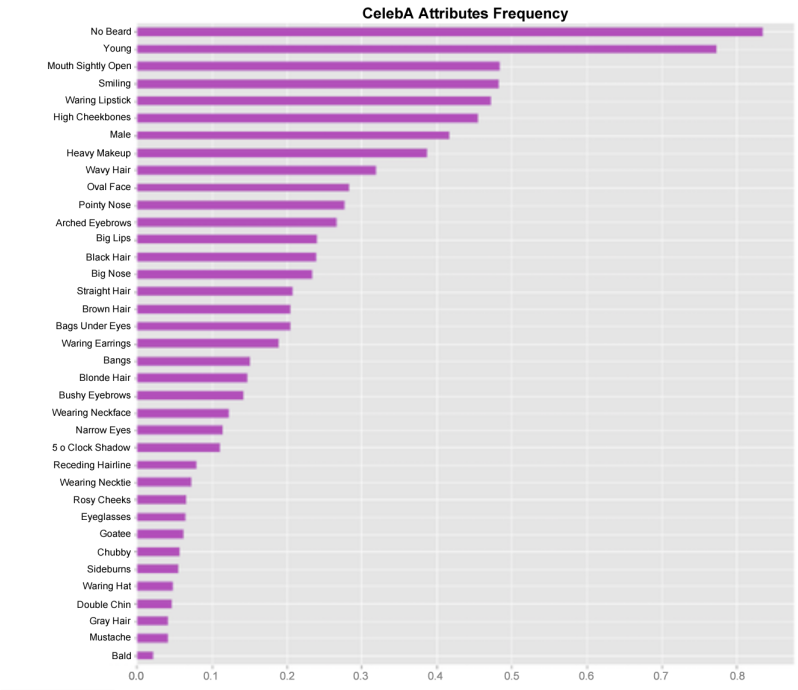


Figure 4.10: The percentage of images for each label.

4.2.4 The fine-tuning of the model

Our model performs a transfer learning from the MobileNetV2 [117] network, chosen empirically after experimenting with different architectures. It turned out that MobileNet2 performs better at the cost of slower training. The proposed model achieves a test accuracy of 90.95 %. Transfer learning was implemented by removing the top ranking layers (see figure 4.12), adding new top

Figure 4.11: Examples of images for label.

layers for classification adapted to the problem to be solved (see figure 4.13) and adjusting the network weights by a quick training phase.

Input	Operator	t	c	n	s
$224^2 \times 3$	conv2d	-	32	1	2
$112^2 \times 32$	bottleneck	1	16	1	1
$112^2 \times 16$	bottleneck	6	24	2	2
$56^2 \times 24$	bottleneck	6	32	3	2
$28^2 \times 32$	bottleneck	6	64	4	2
$14^2 \times 64$	bottleneck	6	96	3	1
$14^2 \times 96$	bottleneck	6	160	3	2
$7^2 \times 160$	bottleneck	6	320	1	1
$7^2 \times 320$	conv2d 1x1	-	1280	1	1
$7^2 \times 1280$	avgpool 7x7	-	-	1	-
$1 \times 1 \times 1280$	conv2d 1x1	-	k	-	

Figure 4.12: MobileNetV2 architecture.

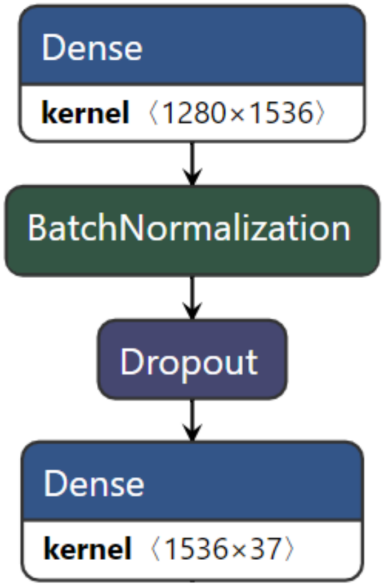


Figure 4.13: From the upper layer: Top Layers: a Dense Layer with 1536 neurons; standardization by Batch-Normalization; regularization of neurons by applying a dropout on 30 % of the connections; the last dense layer outputs 37 neurons to classify the 37 facial features.).

The output of the proposed model consists of a binary vector with 37 values; compared to the dataset, three tags have been eliminated because they are considered too subjective (light skin and attractiveness) or not inherent to the subject face (blurred image). In figure 4.13 the last layers added to the final layer of the MobileNetV2 model are shown. A Data augmentation technique was used to overcome the problem of the unbalanced dataset and to obtain a higher level of generalization of the results. New image samples have been introduced starting from those available by applying:

- maximum rotation of 20 degrees of the image;
- pixel shift for rows and columns for a maximum of 20% of

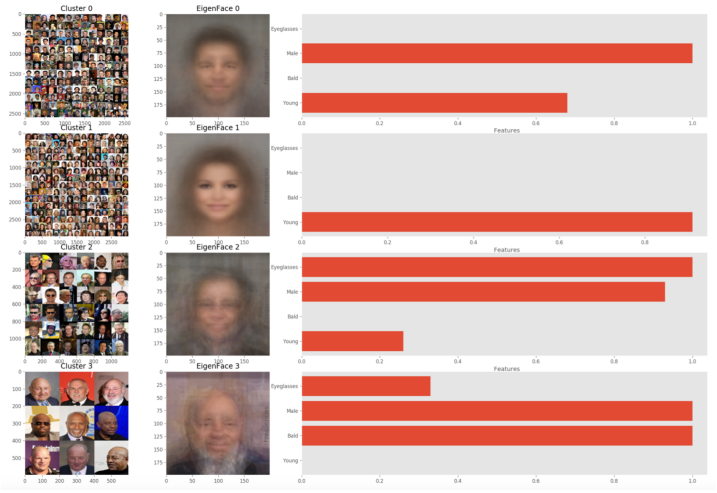


Figure 4.14: The output of the method on a sample of images; selecting the attributes "Eyeglasses", "Male", "Bald" and "Young" four clusters were extracted.

	5 Shadow	Arched Eyebrows	Bags Under Eyes	Bald	Bangs	Big Lips	Big Nose	Black Hair	Blond Hair	Brown Hair	Bushy Eyebrows	Chubby	Double Chin	Eyeglasses	Goatee	Gray Hair	Heavy Makeup	High Cheekbones	Male
Our approach	95	81	85	99	96	71	84	90	96	89	92	96	96	99	97	98	91	87	98

	Mouth Slightly open	Mustache	Narrow Eyes	No Beard	Oval Face	Pointy Nose	Receding Hairline	Rosy Cheeks	Sideburns	Smiling	Straight Hair	Wavy Hair	Wearing Earrings	Wearing Hat	Wearing Lipstick	Wearing Necklace	Wearing Necktie	Young	**Average**
Our approach	94	97	87	96	76	76	93	95	98	93	83	82	90	99	91	88	97	88	**91**

Figure 4.15: Performance of our attribute prediction models.

the entire length/height of the image;

- random shear;
- zoom, maximum 20% of the image size;
- horizontal flipping of images.

4.2.5 Experimental results

The dataset was divided into training set, validation set and test set. The model was trained with 160,000 samples of size 224 x 224 pixels. The validation phase took place on 20,000 samples, the batch size was set to 64 images. The output gives us information on: the number of clusters found, the attributes contained in each cluster, a graphical view of the clusters and an eigenface (all the facial features of the cluster summarized in a single face). In figure 4.14 an example of the output is shown: four clusters have been identified, the content of the clusters is shown in the first image, the eigenface is shown in the second image and the percentage of occurrences of each attribute in the cluster is shown in the third image.

The clustering method used is K-Means, the number of clusters to be input is decided by the DBSCAN method. Figure 4.15 shows the forecast accuracy percentage of each cluster attribute.

4.3 Lip-based video surveillance system

In this section we analyze how a system is created to recognize the identity of an individual from the dynamics of the subject's face, framed by an RGB camera, when he/she pronounces phrases that induce sub-facial movements; these dynamics can be seen as a kind of signature that uniquely and unmistakably identifies an individual. An authentication system is therefore developed which extracts a series of facial geometric characteristics, given a continuous video stream; the geometric characteristics are then subjected to a normalization procedure and given as input to a machine learning model that returns the class of membership of the subject. The work presents a series of experimental results using four different machine learning models: artificial neural networks, SVM, Decision Tree and Random Forest. Performance is estimated through metrics such as accuracy, precision and recall, ROC and AUC curves, EER and confusion matrices. The system shows a good percentage of accuracy, which on the best config-

urations, regarding neural networks, SVM and Random Forest, exceeds 90% reaching a maximum threshold of 99% in more controlled conditions and with low ambiguity. On the other hand, the proposed method shows a fair sensitivity to noise as well as a decline in performance in the presence of noises induced by environmental factors (unfavorable lights and shadows) that have a negative impact on the extraction of facial features. The low complexity of calculation of the features makes the extraction method suitable for a real-time authentication procedure in an application context of a small/medium-sized company that wants to increase the security level of any sensitive area or gate, to a limited number of employees. The ability to deliver an output with an average of 6 seconds represents an excellent compromise between performance and response time.

4.3.1 State-of-the-art of labial recognition

Facial recognition research today is strongly motivated by the numerous practical applications to which this biometrics lends itself very efficiently, both in terms of ease of acquisition and performance. Thanks to the rapid advancements of technologies such as digital cameras, Internet, mobile devices and the growing security needs, face recognition has aroused increasing interest to become one of the most important and used biometric recognition technologies. Let us now focus our attentions on the discriminating power that the dynamism of a face offers. The idea behind the proposed study is based on two research areas, and on the analysis of the changes that occur on the face when a certain expression or phrase is pronounced: facial expression recognition and visual speech recognition.

Facial expression recognition: the recognition of facial expressions is part of the AFEA (Automatic Facial Expression Analysis) systems which aim to automatically analyze and recognize facial movements and changes in facial features with respect to visual information, [96]. In 1978, a first attempt to automatically study expressions was presented by Suwa et al. [118], who au-

tomatically analyzed facial expressions by tracing the movement of 20 points located on the face, on a sequence of images. Since then, progresses have been made in building systems that aim to understand and use facial expressions [119, 120, 121].

Visual speech recognition: the principal results are represented by audio-visual speech recognition systems (AVSR) [122, 123] designed to automate the lip reading process. These systems differ according to the purpose: Automated Lips Reading (ARL) [124, 125] and Lip Motion Recognition Systems (LMR) aim to recognize, respectively, the words spoken by an individual and the lip movement using only the visual signal produced during the speech, excluding any auditory signals. These can be divided into two categories, based on the type of features they use: geometric features and appearance-based features.

4.3.1.1 Geometric features

The geometric features explicitly analyze facial features and the geometric relationships that exist between them, thus describing the shape of the face and its components, such as mouth, nose or cheekbones. An example of geometric features are the facial landmarks detailed in section 4.3.1.1. Geometry-based facial recognition algorithms exploit key points located on the face and the distances between them to obtain a representative descriptor of the face. Therefore, the main operations are the localization and tracing of a dense set of facial points (landmarks). J. Zhang et al. in [126] present a study on modern face geometry based recognition techniques (eigenface, elastic matching and neural networks) starting from the foundations laid in 1992 by the work of A. Samal and P.A. Iyengar [127]. In [128] the landmarks are used to determine a series of measurements derived from three key points of the human face: the two eyes and the center of the mouth (SDAM method, Simple Direct Appearance Mode [129]). In [130] a method is proposed that measures the Procrustes distance [131] between two sets of facial landmarks. Another geometric approach, this time dynamic, for the extraction of the features is developed by

Petajan [124], whose Lip Reading system makes use of measures such as height, width, perimeter and area of the mouth, obtained separately from binary facial images. The extraction process uses a simple technique based on thresholding, while Dynamic Time Warping (DTW) is used for the recognition phase. The prototype proposed by Werda et al. [132]: Automatic Lip Feature Extraction (ALiFE), on the other hand, includes a lip localization module, which exploits the geometric characteristics relating to height, width and area of the mouth.

4.3.1.2 Appearance-based features.

Other recognition techniques use features based on appearance to describe the facial texture and how it is modified following an expression. The best known methods are Eigenfaces [133] and PCA (Principal Component Analysis) [96]. Given a set of normalized images of human faces, a projection is made in a subspace in which the salient features are highlighted, excluding information that is not relevant. The structure of the face is then broken down into a combination of uncorrelated orthogonal components (eigenfaces). Then, each image is represented as a weighted sum (vector of features) of these eigenfaces. The comparison between the images is done simply by evaluating the distance between these vectors of local characteristics. On the basis of [133], various studies have been carried out that have led not only to drastically improve their performance, but also to the expansion of the fields of application of these methods. In 1995, Belongie and Weber introduced a lip-reading system that exploits the optical flow and a gradient-based filtering technique for the extraction of features. These are encoded in the form of a 1D wave and further processed by a PCA. The performances of traditional approaches are sensitive to unconditional or uncontrolled facial changes. These are changes in brightness, changes in poses, masking and management of the different possible expressions. These issues are analyzed in the work of M.Pantic and L.J.M. Rothkrantz [134] An improvement in performance can be found in Gabor [135] and LBP (Local Binary

Pattern) [136], and even more in their multi-dimensional extensions [137, 138, 139] thanks to the use of some invariance properties related to local filtering. The binary local pattern method is also exploited for the recognition of the periocular area: in [140] a system is proposed that detects the iris and pupil region within an image of the eye and extracts the characteristics using LBP. The matching is then carried out by means of bit-shifting.

4.3.1.3 Approaches with neural networks

The growing interest and development of machine learning models has made it possible to tackle several intrinsically complex problems, such as gender classification and expression recognition. In [141] a system for face recognition and verification and for the analysis of facial expression, called WISARD, is developed; it uses a single-level adaptive network. A classifier is built for each subject in the dataset, and the result is obtained by choosing the one that gives the highest recognition accuracy value for a given input image. A further push towards self-learning was given in 2012 with the advent of Deep Learning (DL) and AlexNet [109]. DL includes a set of algorithms that solve a large class of problems using Machine Learning algorithms in multi-level Neural Networks, each corresponding to a different degree of abstraction. With Convolutional Neural Networks (CNN), features are automatically extracted directly from the network. Although the DL methods show a strong invariance to unconditional changes, the use of facial descriptors has some advantages: a very large dataset is not necessary, the analysis of the features is much simpler and the computational complexity decreases both in terms of machine power and execution times. In the proposed work, a geometric approach based on landmarks is exploited, with the addition of a dynamic component, linked to the temporal aspect, which contributes to making biometrics more robust and secure. Furthermore, the method of extracting the landmarks of the face is very efficient in terms of time complexity. The choice of a biometry that links facial features (facial recognition techniques) to the fa-

cial dynamics triggered during phonation depended not only on the considerations made just now but also on a series of encouraging results obtained in the works [142, 143], in which the dynamism of the face was a very strong discriminating factor.

4.3.2 The proposed system

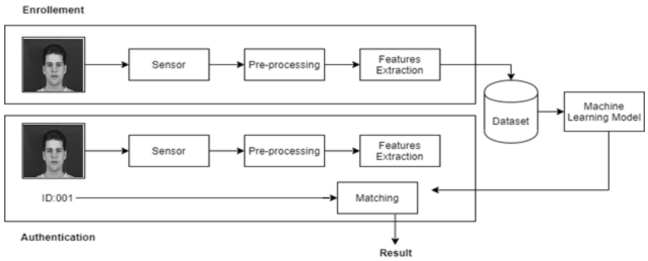


Figure 4.16: Architecture of the proposed system.

In this section the goal is, therefore, to develop a video surveillance system that exploits facial dynamics for staff authentication.

4.3.2.1 System requirements

The fundamental requirements were performance, in terms of recognition accuracy, algorithm response time and privacy. In particular, the time taken by the system from the identification of the face in the video stream to a response (access denied or consented) must be a few tens of seconds. Furthermore, the transformation of a face into a numerical vector of characteristics allows full respect for privacy [144] as the identity of the subject is associated with a series of distance measures, regardless of the images of the face. The system is, moreover, tolerant to partial occlusions of the face, which, in facial recognition systems, negatively affect the acquisition process, hindering it or completely preventing it. The proposed solution, at the end of the design phase, provided for a system that takes as input a continuous video stream and validates the access of an individual based on the pronunciation of a

sentence. The conclusion of the design phase consists of the definition of the functional and non-functional requirements of the system.

Figure 4.17: Examples of facial occlusions

4.3.3 System implementation

The implementation of the proposed authentication system is divided into the two operational phases of Enrollment and Recognition [145], as in figure 4.18.

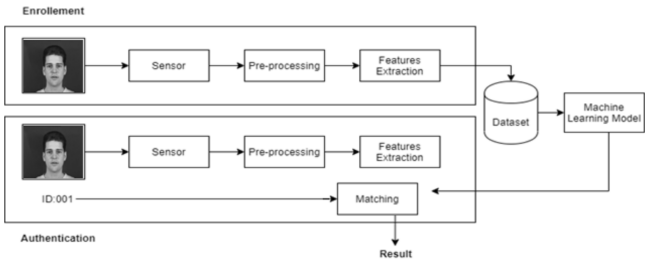


Figure 4.18: Architecture of the proposed system.

Figure 4.19: Face Detector (Dlib) is the result of applying a set of random regression trees, used to extract 194 landmarks of a face in an image.

4.3.3.1 Enrollment phase

The Enrollment phase involves the implementation of three distinct modules. The first module consists of the pre-processing phase and extraction of the characteristics from the previously collected videos. As a first step, a face detection algorithm implemented in the DLib library is applied to the video sequence. This algorithm carries out the face alignment, that is the identification of the geometric structure of the human face. Given the position and size of the face, the shape of the components, such as eyes and nose, is automatically detected. Once the area containing the face has been located, the feature calculation process begins, which identifies and tracks the landmarks, key points for

the calculation and extraction of features. A total of 59 landmarks has been identified, some of which are connected in pairs by a segment representing their distance; the variation of the expression will lead to a variation of these distances. The trend of these distances allows to obtain a time series that summarizes the variations frame by frame, thus representing the dynamism of the subject's expression. Formally, each segment connecting two landmarks represents a geometric feature f_i characterized by a length l_i and a weight w_i, with $1 \leq i \leq K$ where K is the number of features. Let:

$$T_{si} = (l_{i1}w_i, l_{i2}w_2, ..., l_{ij}w_i, ..., l_{iN}w_i) \tag{4.8}$$

be the i-th time series, related to the variation of a given feature in all N frames of the video. The final "Dynamic Facial Feature" vector is obtained as the sequence of all time series:

$$DEF = (T_{s1}, T_{s2}, ..., T_{si}, ..., T_{sK}) \tag{4.9}$$

The dimension of the DEF vector is equal to $K * N$, that is the product of the number of features by the number of frames. The second module, takes the previously extracted characteristics as input, applies a normalization procedure that reorganizes the data and makes them structurally homogeneous in order to be able to input them to the classifier in an adequate way. In fact, the same number of frames is taken into consideration for each video in order to make all DEF of equal size. The procedure also provides for a reduction in the set of features from 59 to the most important 14. This importance has been attributed to each feature in terms of weight, through the use of Random Forest. Random Forest allow us to get information about the contribution of each feature during the training process. The results showed that 14 features (all located on the lower part of the face) had a weight greater than the others.

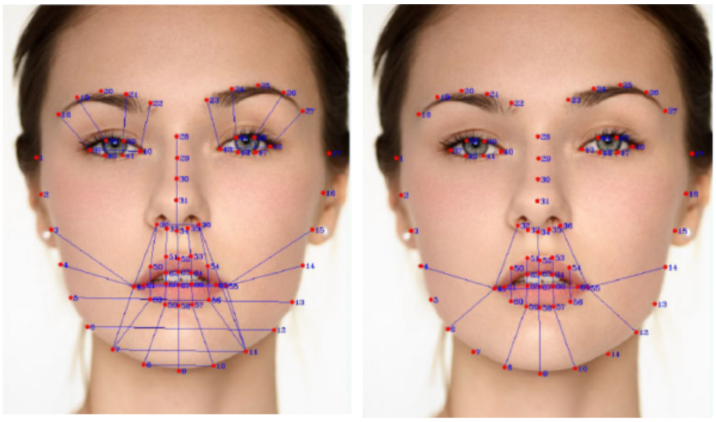


Figure 4.20: Comparison between the complete set of 59 features and the optimized one reduced to 14.

4.3.3.2 Model training phase

In the third and last module, the set of data previously processed is divided into two subsets: one for the training of the automatic learning model and the other for the verification. These two subsets are called training set and test set and comprise respectively 80% and 20% of the data of the entire dataset. The training set is used to train the classifier to recognize subjects, while the test set is used to verify the correctness of the classification. The goodness of the classifier was measured through the use of appropriate performance evaluation metrics presented in detail in the system validation phase. Once trained, the model is exported and used for the subsequent verification and recognition procedure.

4.3.3.3 Real-time recognition

The single procedure for real-time recognition involved the integration of the first two modules in a sequential manner and the querying of the previously trained model. More specifically, given a continuous video stream, the subject who wishes to authenticate is identified through a procedure which, in addition to the detection of the face, captures the lip movement in order to acquire the

frames related to the dynamics of speech. Once a minimum number of pre-established frames has been collected (61 in our case), the feature extraction procedure is applied and the new vector of characteristics is created. Finally, through the prediction function, the array of features is given as input to the pre-trained classifier which returns the class to which the subject in question belongs. If the declared identity coincides with the aforementioned identity, access is granted, otherwise it is denied.

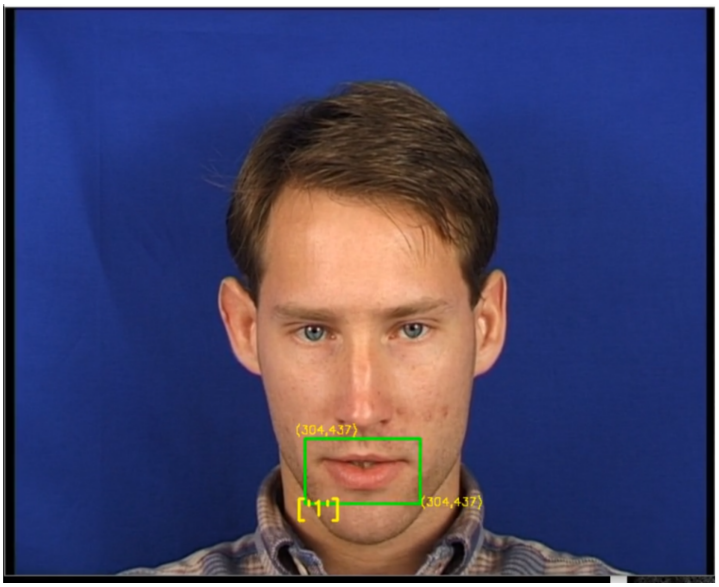


Figure 4.21: The subject is correctly associated with ID 1

4.3.4 Response time

Once the system detects a face and detects a lip movement (indicating that there is a person ready to authenticate), it acquires a certain number of frames from which it extracts the geometric facial features that will be collected in a vector for the newly acquired subject (DEF). The machine learning module provides the subject's identity in output. As this is a real-time authentication system, all the steps described above must take place in a time

interval ranging from 0 to a maximum of 10 seconds. The construction of the vector of characteristics involves the simple processing of a series of distances calculated starting from the points identified on the face, so the complexity in terms of time becomes minimal, allowing an average response time of about 6 seconds.

4.3.5 Experiments and results

The validation of the system proposed in this work gave rise to a series of experiments that allowed both to evaluate its performance and to understand which configuration (number of frames captured - characteristics extracted) was the most suitable for the purpose. All tests were carried out on a Lenovo Legion Y540, equipped with an Intel Core i7-9750H processor (4.5 GHz with 6-core Turbo Boost), NVIDIA GTX1660Ti GPU and 16 GB of RAM. On the software side Python in version 3.6.10 was used, and for the test and tuning of the network the Tensorflow framework in its version for GPU, with the help of the Keras library. The classifiers were created with the help of libraries such as Scikit-learn, SciPy, NumPy and Pandas, while for the part related to Computer Vision (face detection and alignment) OpenCV2 and DLib were used.

4.3.5.1 The XM2VTS dataset

Privately developed by the University of Surrey in England as part of the M2VTS project (Multimodal Verification for Remote Assistance and Security Services) XM2VTS is a multimodal database that collects digital videos relating to 295 subjects. The acquisitions are divided into 4 sessions and perpetuated over time for a period of 4 months (one session per month). Videos are recorded in an indoor environment with controlled lighting conditions using a Sony VX1000E digital video camera in .avi format, 725x576 resolution with 25 fps and 32GHz audio sampling rate. [146] The database also presents a fair intra-class variability for different subjects in which it is possible to find changes such as hair, beard,

glasses, etc. The recordings see the subjects utter three sentences:

- 1. ordered sequence of numbers: "zero, one, two, three, four, five, six, seven, eight, nine, ten";
- 2. unordered sequence of numbers: "five, zero, six, nine, two, eight, one, three, seven, four";
- 3. a sentence: "Joe took fathers green shoe bench out".

There are also recordings, for possible stress tests, in which the participants do not speak, but rotate their heads to the right and left. Given the large number of subjects, the sessions acquired at different times and the consequent intra-class variability, the XM2VTS dataset lends itself perfectly to the experiments of this work.

4.3.5.2 The tests on the XM2VTS dataset

The experiments carried out starting from the XM2VTS database are divided:

- **by number of subjects**, 294 subjects were initially considered in order to test the effectiveness of the recognition system based on the dynamics of geometric facial features; the number of subjects was then reduced, first to 50 and then to 10 in order to simulate an application context relating to a real-time authentication system in which only some employees are authorized to access (column "Subj" in table 4.1).
- **the number of facial landmarks** taken into consideration (59 for the complete feature-set, 14 weighted features) (column "Feat" in table 4.1).
- **the number of frames captured by the videos**, from 61 frames (UPSAMPLED) to 10 (DOWNSAMPLED), (column "Frames" in table 4.1).

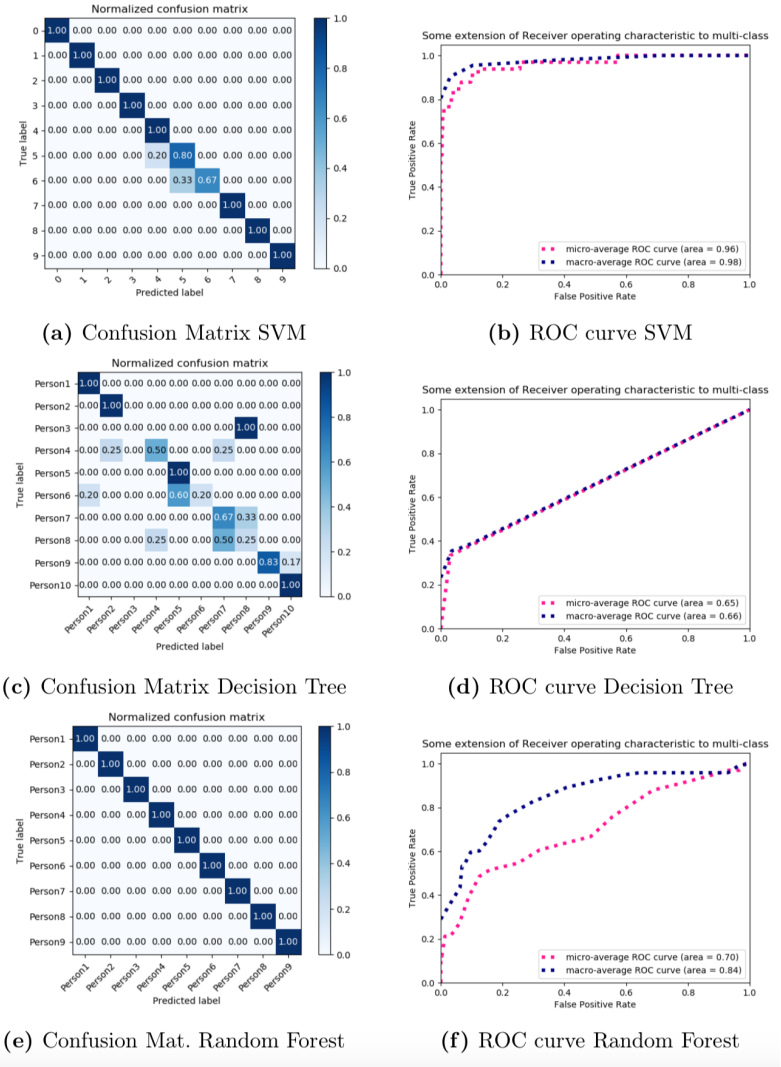
Having to deal with a video surveillance system placed near an opening in which real time acquisitions take place in an uncontrolled environment, the tests were conducted taking into consideration both the videos in which the subjects pronounce the sentence in front of the camera and those in which they rotate their heads left and right.

The best results were obtained on the subset of 10 subjects, with the use of 59 features and the number of frames collected around 61 (UPSAMPLED). The experiments were carried out with the help of three different types of classifiers (SVM, Decision Tree and Random Forest). Below are the results obtained on four different configurations through the use of metrics such as confusion matrix, accuracy, precision, recall and ROC curves.

For the small set of 10 subjects, there was a notable increase in performance of a general nature, due to a lowering of inter-class variability. The best results in terms of accuracy were obtained through the use of the Random Forest classifier, which reaches an accuracy of 99% in the configuration with 59 features for the maximum number of frames (59 UPSAMPLED) showing an ideal confusion matrix (Fig.4.22). Excellent performance is also achieved by the SVM classifier which not only has a better ROC curve than the Random Forest for 59 UPSAMPLED (Fig.4.22b and 4.22f) but on the configuration with 14 features for the maximum number of frames (14 UPSAMPLED) it presents a slightly higher accuracy.

Table 4.1: The results of the experiments carried out with all the combinations of features described.

Subj	Feat	Frames	Method	Acc	Prec	Rec
294	59	UPSAMP	ANN	0,83	-	-
294	14	UPSAMP	ANN	0,70	-	-
50	59	UPSAMP	SVM	0,910	0,889	0,901
50	59	UPSAMP	DT	0,542	0,614	0,542
50	59	UPSAMP	RF	0,937	0,954	0,937
50	59	UPSAMP	ANN	0,85	-	-
50	59	DOWNSAMP	SVM	0,910	0,889	0,901
50	59	DOWNSAMP	DT	0,582	0,634	0,582
50	59	DOWNSAMP	RF	0,951	0,962	0,952
50	59	DOWNSAMP	ANN	0,91	-	-
50	14	UPSAMP	SVM	0,870	0,866	0,871
50	14	UPSAMP	DT	0,466	0,523	0,466
50	14	UPSAMP	RF	0,826	0,877	0,826
50	14	UPSAMP	ANN	0,85	-	-
50	14	DOWNSAMP	SVM	0,862	0,838	0,861
50	14	DOWNSAMP	DT	0,386	0,440	0,386
50	14	DOWNSAMP	RF	0,831	0,878	0,831
50	14	DOWNSAMP	ANN	0,80	-	-
10	59	UPSAMP	SVM	0,955	0,967	0,949
10	59	UPSAMP	DT	0,689	0,871	0,689
10	59	UPSAMP	RF	**0,999**	**0,999**	**0,999**
10	59	DOWNSAMP	SVM	0,954	0,967	0,9498
10	59	DOWNSAMP	DT	0,636	0,690	0,636
10	59	DOWNSAMP	RF	0,957	0,970	0,957
10	14	UPSAMP	SVM	0,919	0,944	0,907
10	14	UPSAMP	DT	0,729	0,822	0,729
10	14	UPSAMP	RF	0,909	0,947	0,909
10	14	DOWNSAMP	SVM	0,869	0,912	0,871
10	14	DOWNSAMP	DT	0,544	0,501	0,544
10	14	DOWNSAMP	RF	0,954	0,977	0,954



(a) Confusion Matrix SVM **(b)** ROC curve SVM

(c) Confusion Matrix Decision Tree **(d)** ROC curve Decision Tree

(e) Confusion Mat. Random Forest **(f)** ROC curve Random Forest

Figure 4.22: ROC curves and confusion matrix related to video experiments with 59 features with the maximum number of frames for 10 subjects, which are the experiments that achieved the best results.

Chapter 5

Conclusions

In this thesis we have presented our results in terms of biometric systems in the field of Homeland Security, developed in the last three years. We explored both physical and behavioral traits:

- head pose estimation, to find out the frontal pose of a person from a video sequence or to extract a required pose. In this context, the use of regression has given better results than classification. This method, used on high resolution images, can help in building a database of frontal faces;
- gender recognition, to understand the gender of a person based on how he/she walks. It has given us excellent results on the Gotcha-I dataset purposely built in a controlled context. In the future we plan to experiment with these methods in contexts in the wild;
- facial features that can be derived from the face of an individual; a future contribution consists in the creation of a platform capable of automatically labeling all the information concerning the shape of the face, eyebrows, gender, age, eye color, hair color, etc. from a face. This tool can be of support to the police for the construction of the identikit of wanted persons;
- cooperativeness, to identify the subject's attitude towards

the camera. This biometrics, associated with context recognition techniques, can return information useful in video-surveillance applications;

- labial, as identification of an individual based on his/her facial expressions during the pronunciation of a certain sentence. Such a tool could be of interest in access control applications.

The biometric data considered are not to be intended as an alternative to other types of biometrics for recognition, but as additional sources of information. The same biometrics can be used in different contexts to extract information for different tasks.

These studies are aimed at the constitution of research results useful for an in-depth analysis of the action recognition. Information on **who the person is** and **what he/she is doing** in a video surveillance context is important for the creation of increasingly advanced video surveillance systems that minimize the need for human intervention.

Bibliography

[1] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*, 1st ed. Springer Publishing Company, Incorporated, 2010. [Online]. Available: https://doi.org/10.1007/978-0-387-71041-9

[2] M. Nappi and D. Riccio, *Moderne tecniche di elaborazione di immagini e biometria*. CUA - Coop. Univ. Athena, 2008. [Online]. Available: https://books.google.it/books?id=LRL8PAAACAAJ

[3] A. E. K. Ghalleb and N. E. B. Amara, "Remote person authentication in different scenarios based on gait and face in front view," *2017 14th International Multi-Conference on Systems, Signals Devices (SSD), DOI: 10.1109/SSD.2017.8167008*, pp. 486–491, 2017. [Online]. Available: https://doi.org/10.1109/SSD.2017.8167008

[4] P. Barra, C. Bisogni, M. Nappi, and S. Ricciardi, *Fast QuadTree-Based Pose Estimation for Security Applications Using Face Biometric*, 08 2018, pp. 160–173. [Online]. Available: https://doi.org/10.1007/978-3-030-02744-5_12

[5] A. F. Abate, P. Barra, C. Bisogni, M. Nappi, and S. Ricciardi, "Near real-time three axis head pose estimation without training," *IEEE Access*, vol. 7, pp. 64 256–64 265, 2019. [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2917451

[6] P. Barra, C. Bisogni, M. Nappi, D. Freire-Obregon, and M. Castrillon-Santana, "Gender classification on 2d human skeleton," in *Proceedings of the 2019 3rd International Conference on Bio-engineering for Smart Technologies (BioSMART), DOI:*

10.1109/BIOSMART.2019.8734198, 2019, pp. 1–4. [Online]. Available: https://doi.org/10.1109/BIOSMART.2019.8734198

[7] P. Barra, C. Bisogni, M. Nappi, D. Freire-Obregon, and M. Castrillon-Santana, "Gait analysis for gender classification in forensics," *Communications in Computer and Information Science, DOI: 10.1007/978-981-15-1304-6_15*, vol. 1123 CCIS, 2019.

[8] D. Freire-Obregon, M. Castrillon-Santana, P. Barra, C. Bisogni, and M. Nappi, "An attention recurrent model for human cooperation detection," *Computer Vision and Image Understanding, DOI:10.1016/j.cviu.2020.102991*, vol. 197-198, p. 102991, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S107731422030062X

[9] I. Maria De Marsico (Sapienza University of Rome and I. Michele Nappi (University of Salerno, "Face recognition in adverse conditions: A look at achieved advancements," *Face Recognition in Adverse Conditions 2014 —Pages: 26 ISBN13: 9781466659667—ISBN10: 1466659661—EISBN13: 9781466659674 DOI: 10.4018/978-1-4666-5966-7.ch018*.

[10] M. De Marsico, M. Nappi, D. Riccio, and H. Wechsler, "Robust face recognition for uncontrolled pose and illumination changes," *IEEE Transactions on Systems, Man, and Cybernetics: Systems. DOI: 10.1109/TSMCA.2012.2192427*, vol. 43, no. 1, pp. 149–163, 2013.

[11] Y. Cho and K. Yoon, "Pamm: Pose-aware multi-shot matching for improving person re-identification," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3739–3752, 2018. [Online]. Available: https://doi.org/10.1109/TIP.2018.2815840

[12] J. C. Neves, G. Santos, S. Filipe, E. Grancho, S. Barra, F. Narducci, and H. Proença, "Quis-campi: Extending in the wild biometric recognition to surveillance environments," in *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops*, V. Murino, E. Puppo, D. Sona, M. Cristani, and C. Sansone, Eds. Cham: Springer International Publishing, 2015, pp. 59–68. [Online]. Available: https://doi.org/10.1007/978-3-319-23222-5_8

[13] N. F. B. S. e. a. Neves, J., "Biometric recognition in surveillance scenarios: a survey." *Artif Intell Rev 46, 515â541 (2016).* [Online]. Available: https://doi.org/10.1007/s10462-016-9474-x

[14] M. Ding and G. Fan, "Articulated and generalized gaussian kernel correlation for human pose estimation," *IEEE Transactions on Image Processing. DOI: 10.1109/TIP.2015.2507445*, vol. 25, no. 2, pp. 776–789, 2016.

[15] J. Chen, S. Nie, and Q. Ji, "Data-free prior model for upper body pose estimation and tracking," *IEEE Transactions on Image Processing. DOI:10.1109/TIP.2013.2274748*, vol. 22, no. 12, pp. 4627–4639, 2013.

[16] N. M. . R. D. De Marsico, M., "Face authentication with undercontrolled pose and illumination." *SIViP 5, 401 (2011). https://doi.org/10.1007/s11760-011-0244-6.*

[17] R. Valenti, N. Sebe, and T. Gevers, "Combining head pose and eye location information for gaze estimation," *IEEE Transactions on Image Processing. DOI: 10.1109/TIP.2011.2162740*, vol. 21, no. 2, pp. 802–815, 2012.

[18] M. De Marsico, M. Nappi, and D. Riccio, "Measuring measures for face sample quality," in *Proceedings of the 3rd International ACM Workshop on Multimedia in Forensics and Intelligence*, ser. MiFor '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 7â12. [Online]. Available: https://doi.org/10.1145/2072521.2072524

[19] T. Jantunen, J. Mesch, A. Puupponen, and J. Laaksonen, "On the rhythm of head movements in finnish and swedish sign language sentences," in *Speech Prosody 2016*, 2016, pp. 850–853. [Online]. Available: http://dx.doi.org/10.21437/SpeechProsody.2016-174

[20] M. De Marsico, M. Nappi, and D. Riccio, "Measuring sample distortions in face recognition," in *Proceedings of the 2nd ACM Workshop on Multimedia in Forensics, Security and Intelligence*, ser. MiFor '10. New York, NY, USA: Association

for Computing Machinery, 2010, p. 83â88. [Online]. Available: https://doi.org/10.1145/1877972.1877994

[21] R. Stiefelhagen, "Estimating head pose with neural networks-results on the pointing04 icpr workshop evaluation data," in *Proc. Pointing 2004 Workshop: Visual Observation of Deictic Gestures*, vol. 1, no. 5, 2004, pp. 21–24. [Online]. Available: https://cvhci.anthropomatik.kit.edu/~stiefel/papers/pointing2004_final.pdf

[22] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, and Y.-Y. Chuang, "Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1087–1096. [Online]. Available: https://www.csie.ntu.edu.tw/~cyy/publications/papers/Yang2019FSA.pdf

[23] Y. Wang, W. Liang, J. Shen, Y. Jia, and L.-F. Yu, "A deep coarse-to-fine network for head pose estimation from synthetic data," *Pattern Recognition*, vol. 94, pp. 196 – 206, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320319302018

[24] M. Raza, Z. Chen, S.-U. Rehman, P. Wang, and P. Bao, "Appearance based pedestriansâ head pose and body orientation estimation using deep learning," *Neurocomputing*, vol. 272, pp. 647 – 659, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231217312869

[25] I. Chamveha, Y. Sugano, D. Sugimura, T. Siriteerakul, T. Okabe, Y. Sato, and A. Sugimoto, "Appearance-based head pose estimation with scene-specific adaptation," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011, pp. 1713–1720.

[26] H.-W. Hsu, T.-Y. Wu, S. Wan, W. H. Wong, and C.-Y. Lee, "Quatnet: Quaternion-based head pose estimation with multiregression loss," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 1035–1046, 2018. [Online]. Available: https://doi.org/10.1109/TMM.2018.2866770

[27] H. Liu and L. Ma, "Online person orientation estimation based on classifier update," in *2015 IEEE International Conference on Image Processing (ICIP). DOI: 10.1109/ICIP.2015.7351064*. IEEE, 2015, pp. 1568–1572.

[28] K. Pawelczyk and M. Kawulok, "Head pose estimation relying on appearance-based nose region analysis," in *Computer Vision and Graphics. ISBN: 978-3-319-11331-9*, L. J. Chmielewski, R. Kozera, B.-S. Shin, and K. Wojciechowski, Eds. Cham: Springer International Publishing, 2014, pp. 510–517.

[29] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2018, pp. 2155–215 509. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPRW.2018.00281

[30] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2019. [Online]. Available: https://doi.org/10.1109/TPAMI.2017.2781233

[31] A. Kumar, A. Alavi, and R. Chellappa, "Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors," in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, 2017, pp. 258–265. [Online]. Available: https://doi.org/10.1109/FG.2017.149

[32] V. Drouard, R. Horaud, A. Deleforge, S. Ba, and G. Evangelidis, "Robust head-pose estimation based on partially-latent mixture of linear regressions," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1428–1440, 2017. [Online]. Available: https://doi.org/10.1109/TIP.2017.2654165

[33] V. Drouard, S. Ba, G. Evangelidis, A. Deleforge, and R. Horaud, "Head pose estimation via probabilistic high-dimensional regression," in *Proceedings of the 2015 IEEE*

International Conference on Image Processing (ICIP). IEEE, 2015, pp. 4624–4628. [Online]. Available: https://doi.org/10.1109/ICIP.2015.7351683

[34] N. Gourier, J. Maisonnasse, D. Hall, and J. L. Crowley, "Head pose estimation on low resolution images," in *Proceedings of the International Evaluation Workshop on Classification of Events, Activities and Relationships*. Springer, 2006, pp. 270–280. [Online]. Available: https://doi.org/10.1007/978-3-540-69568-4_24

[35] G. Fanelli, J. Gall, and L. Van Gool, "Real time head pose estimation with random regression forests," in *CVPR 2011. DOI:10.1109/CVPR.2011.5995458*, 2011, pp. 617–624.

[36] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3d total solution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 78–92, Jan 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2017.2778152

[37] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*, 2017.

[38] J. M. Diaz Barros, B. Mirbach, F. Garcia, K. Varanasi, and D. Stricker, *Real-Time Head Pose Estimation by Tracking and Detection of Keypoints and Facial Landmarks*, 07 2019, pp. 326–349. [Online]. Available: https://doi.org/10.1007/978-3-030-26756-8_16

[39] G. A. PelAez C., F. GarcIa, A. de la Escalera, and J. M. Armingol, "Driver monitoring based on low-cost 3-d sensors," *IEEE Transactions on Intelligent Transportation Systems. DOI: 10.1109/TITS.2014.2332613*, vol. 15, no. 4, pp. 1855–1860, 2014.

[40] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3d face analysis," *Int. J. Comput. Vision*, vol. 101, no. 3, pp. 437–458, February 2013. [Online]. Available: https://doi.org/10.1007/s11263-012-0549-0

[41] T. Weise, S. Bouaziz, H. Li, and M. Pauly, "Realtime performance-based facial animation," in *ACM SIGGRAPH 2011 Papers*, ser. SIGGRAPH '11. New York, NY, USA: Association for Computing Machinery, 2011. [Online]. Available: https://doi.org/10.1145/1964921.1964972DOI:10.1145/1964921.1964972

[42] T. Weise, T. Wismer, B. Leibe, and L. V. Gool, "Online loop closure for real-time interactive 3d scanning," *Comput. Vis. Image Underst.*, vol. 115, no. 5, p. 635â648, May 2011. [Online]. Available: https://doi.org/10.1016/j.cviu.2010.11.023

[43] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance. DOI: 10.1109/AVSS.2009.58*, 2009, pp. 296–301.

[44] H. Li, B. Adams, L. J. Guibas, and M. Pauly, "Robust single-view geometry and motion reconstruction," *ACM Trans. Graph.*, vol. 28, no. 5, pp. 1 – 10, Dec. 2009. [Online]. Available: https://doi.org/10.1145/1618452.1618521

[45] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence. DOI: 10.1109/34.121791*, vol. 14, no. 2, pp. 239–256, 1992.

[46] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," 03 2018. [Online]. Available: https://doi.org/10.1007/978-3-030-01264-9_33

[47] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1867–1874. [Online]. Available: https://doi.org/10.1109/CVPR.2014.241

[48] P. Barra, S. Barra, C. Bisogni, M. De Marsico, and M. Nappi, "Web-shaped model for head pose estimation: An approach

for best exemplar selection," *IEEE Transactions on Image Processing*, vol. 29, pp. 5457–5468, 2020. [Online]. Available: https://doi.org/10.1109/TIP.2020.2984373

[49] M. Fiorucci, M. Khoroshiltseva, M. Pontil, A. Traviglia, A. Del Bue, and S. James, "Machine learning for cultural heritage: A survey," *Pattern Recognition Letters*, vol. 133, pp. 102–108, 2020. [Online]. Available: https://doi.org/10.1016/j.patrec.2020.02.017

[50] C. M. Bishop, *Pattern recognition and machine learning. ISBN: 978-0-387-31073-2*. Springer, 2006.

[51] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of machine learning research*, vol. 1, no. Jun, pp. 211–244, 2001. [Online]. Available: https://doi.org/10.1162/15324430152748236

[52] D. J. MacKay, "Bayesian interpolation," *Neural computation. DOI : https://doi.org/10.1007/978-94-017-2219-3_3*, vol. 4, no. 3, pp. 415–447, 1992.

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research.*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://scikit-learn.org/stable/

[54] C. E. Rasmussen, *Gaussian Processes in Machine Learning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 63–71. [Online]. Available: https://doi.org/10.1007/978-3-540-28650-9_4

[55] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Subspace, Latent Structure and Feature Selection*, C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 34–51. [Online]. Available: https://doi.org/10.1007/11752790_2

[56] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, Aug 2004. [Online]. Available: https://doi.org/10.1023/B:STCO.0000035301.49549.88

[57] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. [Online]. Available: https://doi.org/10.1109/ICCV.2017.116

[58] S. G. Kong and R. O. Mbouna, "Head pose estimation from a 2d face image using 3d face morphing with depth parameters," *IEEE Transactions on Image Processing*, vol. 24, no. 6, pp. 1801–1808, 2015. [Online]. Available: https://doi.org/10.1109/TIP.2015.2405483

[59] A. R. Anwary, H. Yu, and M. Vassallo, "Optimal foot location for placing wearable imu sensors and automatic feature extraction for gait analysis," *IEEE Sensors Journal*, vol. 18, no. 6, pp. 2555–2567, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8246577

[60] F. Sun, C. Mao, X. Fan, and Y. Li, "Accelerometer-based speed-adaptive gait authentication method for wearable iot devices," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 820–830, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8421575

[61] D. Phan, N. Nguyen, P. N. Pathirana, M. Horne, L. Power, and D. Szmulewicz, "A random forest approach for quantifying gait ataxia with truncal and peripheral measurements using multiple wearable sensors," *IEEE Sensors Journal*, vol. 20, no. 2, pp. 723–734, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8850039

[62] S. Majumder, T. Mondal, and M. J. Deen, "A simple, low-cost and efficient gait analyzer for wearable healthcare applications," *IEEE Sensors Journal*, vol. 19, no. 6, pp. 2320–2329, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8561201

[63] F. Juefei-Xu, C. Bhagavatula, A. Jaech, U. Prasad, and M. Savvides, "Gait-id on the move: Pace independent human identification using cell phone accelerometer dynamics," in *Proceedings of the 2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2012, pp. 8–15. [Online]. Available: https://doi.org/10.1109/BTAS.2012.6374552

[64] C. Nickel, T. Wirtl, and C. Busch, "Authentication of smartphone users based on the way they walk using k-nn algorithm," in *Proceedings of the 2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2012, pp. 16–20. [Online]. Available: https://doi.org/10.1109/IIH-MSP.2012.11

[65] M. Nowlan, "Human recognition via gait identification using accelerometer gyro forces," 2009. [Online]. Available: http://root81.com/pub/mfn_gait_id.pdf

[66] M. Gadaleta and M. Rossi, "Idnet: Smartphone-based gait recognition with convolutional neural networks," *Pattern Recognition*, vol. 74, pp. 25 – 37, 2018. [Online]. Available: https://doi.org/10.1016/j.patcog.2017.09.005

[67] G. Giorgi, F. Martinelli, A. Saracino, and M. Sheikhalishahi, "Try walking in my shoes, if you can: Accurate gait recognition through deep learning," 09 2017, pp. 384–395. [Online]. Available: https://doi.org/10.1007/978-3-319-66284-8_32

[68] C. Meena, R. Kumar, and N. Mittal, "Recent developments in human gait research: parameters, approaches, applications, machine learning techniques, datasets and challenges," *Artificial Intelligence Review*, vol. 49, pp. 1–40, 01 2018. [Online]. Available: https://cerc.rloew.eu/proceedings/CERC2020_Proceedings.pdf

[69] Y. Li, P. Zhang, Y. Zhang, and K. Miyazaki, "Gait analysis using stereo camera in daily environment," in *Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 1471–1475. [Online]. Available: https://doi.org/10.1109/EMBC.2019.8857494

[70] W. Kim, Y. Kim, and K. Y. Lee, "Human gait recognition based on integrated gait features using kinect depth cameras," in *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2020, pp. 328–333. [Online]. Available: https://doi.org/10.1109/COMPSAC48688.2020.0-225

[71] Y. Guo, F. Deligianni, X. Gu, and G. Yang, "3-d canonical pose estimation and abnormal gait recognition with a single rgb-d camera," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3617–3624, 2019. [Online]. Available: https://doi.org/10.1109/LRA.2019.2928775

[72] A. Sokolova and A. Konushin, "Methods of gait recognition in video," *Programming and Computer Software*, vol. 45, pp. 213–220, 07 2019. [Online]. Available: https://doi.org/10.1134/S0361768819040091

[73] S. Yu, H. Chen, Q. Wang, L. Shen, and Y. Huang, "Invariant feature extraction for gait recognition using only one uniform model," *Neurocomputing*, vol. 239, pp. 81 – 93, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S092523121730276X

[74] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Geinet: View-invariant gait recognition using a convolutional neural network," in *Proceedings of the 2016 International Conference on Biometrics (ICB)*, 2016, pp. 1–8. [Online]. Available: https://doi.org/10.1109/ICB.2016.7550060

[75] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep cnns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 2, pp. 209–226, 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2545669

[76] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. [Online]. Available: https://doi.org/10.1109/TPAMI.2019.2929257

[77] K. D. Ng, S. Mehdizadeh, A. Iaboni, A. Mansfield, A. Flint, and B. Taati, "Measuring gait variables using computer vision to assess mobility and fall risk in older adults with dementia," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 8, pp. 1–9, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9103018

[78] N. Kour, Sunanda, and S. Arora, "Computer-vision based diagnosis of parkinsonâs disease via gait: A survey," *IEEE Access*, vol. 7, pp. 156 620–156 645, 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8884146

[79] A. Phinyomark, S. Osis, B. Hettinga, D. Kobsar, and R. Ferber, "Gender differences in gait kinematics for patients with knee osteoarthritis," *BMC Musculoskeletal Disorders*, vol. 17, 04 2016. [Online]. Available: https://doi.org/10.1186/s12891-016-1013-z

[80] M. H. Ahmed and A. T. Sabir, "Human gender classification based on gait features using kinect sensor," in *Proceedings of the 2017 3rd IEEE International Conference on Cybernetics (CYBCONF)*, 2017, pp. 1–5. [Online]. Available: https://doi.org/10.1109/CYBConf.2017.7985782

[81] C. Xu, Y. Makihara, Y. Yagi, and J. Lu, "Gait-based age progression/regression: a baseline and performance evaluation by age group classification and cross-age gait identification," *Machine Vision and Applications, DOI : https://doi.org/10.1007/s00138-019-01015-x*, vol. 30, pp. 629–644, 2019.

[82] B. Abirami, T. Subashini, and V. Mahavaishnavi, "Automatic age-group estimation from gait energy images," *Materials Today: Proceedings, doi: https://doi.org/10.1016/j.matpr.2020.08.298*, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2214785320361708

[83] C. Dhiman and D. K. Vishwakarma, "A robust framework for abnormal human action recognition using -transform and zernike moments in depth videos," *IEEE Sensors Journal. , DOI: 10.1109/JSEN.2019.2903645*, vol. 19, no. 13, pp. 5195–5203, 2019.

[84] M. A. Khan, T. Akram, M. Sharif, N. Muhammad, M. Y. Javed, and S. R. Naqvi, "Improved strategy for human action recognition; experiencing a cascaded design," *IET Image Processing. DOI:10.1049/iet-ipr.2018.5769*, vol. 14, no. 5, pp. 818–829, 2020.

[85] H. Wang and L. Wang, "Learning content and style: Joint action recognition and person identification from human skeletons," *Pattern Recognition*, vol. 81, pp. 23 – 35, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320318301195

[86] F. Becattini, T. Uricchio, L. Ballan, L. Seidenari, and A. Del Bimbo, "Am i done? predicting action progress in videos," 05 2017. [Online]. Available: https://doi.org/10.1145/3402447

[87] J. P. T. Sien, K. H. Lim, and P.-I. Au, "Deep learning in gait recognition for drone surveillance system," *IOP Conference Series: Materials Science and Engineering*, vol. 495, p. 012031, jun 2019. [Online]. Available: https://doi.org/10.1088%2F1757-899x%2F495%2F1%2F012031

[88] J. Choi, G. Sharma, M. Chandraker, and J.-B. Huang, "Unsupervised and semi-supervised domain adaptation for action recognition from drones," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). DOI: 10.1109/WACV45572.2020.9093511*, March 2020.

[89] M. Wlodarczyk, D. Kacperski, and W. S. andKamil Grabowski, "Compact: Biometric dataset of face images acquired in uncontrolled indoor environment," 2019. [Online]. Available: https://doi.org/10.7494/csci.2019.20.1.3020

[90] R. Raposo, E. Hoyle, A. Peixinho, and H. ProenÃ§a, "Ubear: A dataset of ear images captured on-the-move in uncontrolled conditions," in *2011 IEEE Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, 2011, pp. 84–90. [Online]. Available: https://ieeexplore.ieee.org/document/5949208

[91] H.-J. Hsu and K.-T. Chen, "Droneface: An open dataset for drone research," in *Proceedings of the 8th ACM on Multimedia*

Systems Conference, ser. MMSys'17. New York, NY, USA: Association for Computing Machinery, 2017, p. 187â192. [Online]. Available: https://doi.org/10.1145/3083187.3083214

[92] L. De Maio, R. Distasi, and M. Nappi, "Mubidus-i: A multibiometric and multipurpose dataset," in *2019 15th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, 2019, pp. 748–753. [Online]. Available: https://ieeexplore.ieee.org/document/9084058

[93] X. Alameda-Pineda, R. Subramanian, E. Ricci, O. Lanz, and N. Sebe, "Chapter 14 - salsa: A multimodal dataset for the automated analysis of free-standing social interactions," in *Group and Crowd Behavior for Computer Vision*, V. Murino, M. Cristani, S. Shah, and S. Savarese, Eds. Academic Press, 2017, pp. 321 – 340. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128092767000175

[94] A. K. Rajivan, "Measurement of gender differences using anthropometry," *Economic and Political Weekly*, pp. WS58–WS62, 1996. [Online]. Available: https://www.jstor.org/stable/4404709?seq=1

[95] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997. [Online]. Available: https://doi.org/10.1109/78.650093

[96] S. Z. Li and A. K. Jain, "Handbook of face recognition," *Springer-Verlag New York. DOI: 10.1007/b138828*, 2005.

[97] H. Proenca, J. C. Neves, S. Barra, T. Marques, and J. C. Moreno, "Joint head pose/soft label estimation for human recognitionin-the-wild," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 12, pp. 2444–2456, 2016.

[98] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV). DOI: 10.1109/ICCV.2015.425*, December 2015.

[99] A. Jain and R.-L. Hsu, "Face detection and modeling for recognition," 2002. [Online]. Available: http://biometrics.cse.msu.edu/Publications/Thesis/Reserved/VincentHsu_FaceDetection_PhD02.pdf

[100] E. Conrad, S. Misenar, and J. Feldman, "Chapter 5 - domain 5: Identity and access management (controlling access and managing identity)," *Eric Conrad, Seth Misenar, and Joshua Feldman, editors, Eleventh Hour CISSPâR (Third Edition)*, p. 117â134, 2017.

[101] A. Gron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. ISBN :1491962291*, 1st ed. O'Reilly Media, Inc., 2017.

[102] A. F. Abate, P. Barra, S. Barra, C. Molinari, M. Nappi, and F. Narducci, "Clustering facial attributes: Narrowing the path from soft to hard biometrics," *IEEE Access. DOI: 10.1109/ACCESS.2019.2962010*, vol. 8, pp. 9037–9045, 2020.

[103] A. K. Singh and G. C. Nandi, "Face recognition using facial symmetry," in *Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology*, ser. CCSEIT '12. New York, NY, USA: Association for Computing Machinery, 2012, pp. 550 – 554. [Online]. Available: https://doi.org/10.1145/2393216.2393308

[104] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). DOI: 10.1109/CVPR.2015.7298682*, June 2015.

[105] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). DOI: 10.1109/CVPR.2017.713*, 2017.

[106] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii, "Scalable k-means++," *Proc. VLDB Endow.*,

vol. 5, no. 7, pp. 622 – 633, Mar. 2012. [Online]. Available: https://doi.org/10.14778/2180912.2180915

[107] I. Davidson and S. S. Ravi, "Agglomerative hierarchical clustering with constraints: Theoretical and empirical results," in *Proceedings of the 9th European Conference on European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. DOI: 10.5555/3121445.3121458*, ser. ECMLPKDD'05. Berlin, Heidelberg: Springer-Verlag, 2005, p. 59â70.

[108] B. Khalil and C. Ali, "Density-based spatial clustering of application with noise algorithm for the classification of solar radiation time series," in *2016 8th International Conference on Modelling, Identification and Control (ICMIC). DOI: 10.1109/ICMIC.2016.7804123*, 2016, pp. 279–283.

[109] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, p. 84â90, May 2017. [Online]. Available: https://doi.org/10.1145/3065386

[110] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV). DOI: 10.1007/s11263-015-0816-y*, vol. 115, no. 3, pp. 211–252, 2015.

[111] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[112] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). DOI: 10.1109/CVPR.2016.90*, 2016, pp. 770–778.

[113] "S231n convolutional neural networks for visual recognition." [Online]. Available: http://cs231n.github.io/transfer-learning#tf

[114] S. Yang, P. Luo, C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *2015 IEEE International Conference on Computer Vision (ICCV). DOI: 10.1109/ICCV.2015.419*, 2015, pp. 3676–3684.

[115] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007. [Online]. Available: https://people.cs.umass.edu/~elm/papers/lfw.pdf

[116] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). DOI: 10.1109/CVPR.2017.463*, 2017, pp. 4352–4360.

[117] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. DOI: 10.1109/CVPR.2018.00474*, 2018, pp. 4510–4520.

[118] M. SUWA, "A preliminary note on pattern recognition of human emotional expression," *Proc. of The 4th International Joint Conference on Pattern Recognition*, pp. 408–410, 1978. [Online]. Available: https://ci.nii.ac.jp/naid/10006751528/en/

[119] J. J.-J. Lien, T. Kanade, J. F. Cohn, and C.-C. Li, "Detection, tracking, and classification of action units in facial expression," *Robotics and Autonomous Systems*, vol. 31, no. 3, pp. 131 – 146, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921889099001037

[120] Y. . I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence. DOI: 10.1109/34.908962*, vol. 23, no. 2, pp. 97–115, 2001.

[121] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803 – 816, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885608001844

[122] A. B. A. Hassanat, "Visual speech recognition," *ArXiv. DOI: 10.5772/19361*, vol. abs/1409.1411, 2014.

[123] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia. DOI: 10.1109/6046.865479*, vol. 2, no. 3, pp. 141–151, 2000.

[124] E. D. Petajan, "Automatic lipreading to enhance speech recognition (speech reading)," *Ph.D. dissertation, University of Illinois at Urbana-Champaign. DOI: 10.5555/911713*, 1984, aAI8502266.

[125] G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," *IEEE Transactions on Multimedia. DOI: 10.1109/TMM.2009.2030637*, vol. 11, no. 7, pp. 1254–1265, 2009.

[126] Jun Zhang, Yong Yan, and M. Lades, "Face recognition: eigenface, elastic matching, and neural nets," *Proceedings of the IEEE. DOI: 10.1109/5.628712*, vol. 85, no. 9, pp. 1423–1435, 1997.

[127] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey," *Pattern Recogn.*, vol. 25, no. 1, p. 65â77, Jan. 1992. [Online]. Available: https://doi.org/10.1016/0031-3203(92)90007-6

[128] Yubo Wang, Haizhou Ai, Bo Wu, and Chang Huang, "Real time facial expression recognition with adaboost," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. DOI: 10.1109/ICPR.2004.1334680*, vol. 3, 2004, pp. 926–929 Vol.3.

[129] T. Wang, H. Ai, and G. Huang, "A two-stage approach to automatic face alignment," in *Third International Symposium on Multispectral Image Processing and Pattern Recognition*, H. Lu and T. Zhang, Eds., vol. 5286, International Society for Optics

and Photonics. SPIE, 2003, pp. 558 – 563. [Online]. Available: https://doi.org/10.1117/12.539038

[130] J. Shi, A. Samal, and D. Marx, "How effective are landmarks and their geometry for face recognition?" *Computer Vision and Image Understanding*, vol. 102, no. 2, pp. 117 – 133, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314205001761

[131] I. L. Dryden and K. V. Mardia, "Statistical shape analysis: With applications in r," *ISBN: 978-0-470-69962-1*, 2016.

[132] S. Werda, W. Mahdi, M. Tmar, and A. Ben Hamadou, "Alife: Automatic lip feature extraction: A new approach for speech recognition application," in *2006 2nd International Conference on Information Communication Technologies. DOI: 10.1109/ICTTA.2006.1684886*, vol. 2, 2006, pp. 2963–2968.

[133] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neuroscience*, vol. 3, no. 1, p. 71â86, Jan. 1991. [Online]. Available: https://doi.org/10.1162/jocn.1991.3.1.71

[134] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence. DOI: 10.1109/34.895976*, vol. 22, no. 12, pp. 1424–1445, 2000.

[135] Z. Zheng, J. Zhao, and J. Yang, "Gabor feature based face recognition using supervised locality preserving projection," in *Advanced Concepts for Intelligent Vision Systems. ISBN: 978-3-540-44632-3*, J. Blanc-Talon, W. Philips, D. Popescu, and P. Scheunders, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 644–653.

[136] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence. DOI: 10.1109/TPAMI.2006.244*, vol. 28, no. 12, pp. 2037–2041, 2006.

[137] Wenchao Zhang, Shiguang Shan, Wen Gao, Xilin Chen, and Hongming Zhang, "Local gabor binary pattern histogram sequence (lgbphs): a novel non-statistical model for face representation and recognition," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. DOI: 10.1109/ICCV.2005.147*, vol. 1, 2005, pp. 786–791 Vol. 1.

[138] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *2013 IEEE Conference on Computer Vision and Pattern Recognition. DOI: 10.1109/CVPR.2013.389*, 2013, pp. 3025–3032.

[139] W. Deng, J. Hu, and J. Guo, "Compressive binary patterns: Designing a robust binary face descriptor with random-field eigenfilters," *IEEE Transactions on Pattern Analysis and Machine Intelligence. DOI: 10.1109/TPAMI.2018.2800008*, vol. 41, no. 3, pp. 758–767, 2019.

[140] S. R. Cho, G. P. Nam, K. Y. Shin, D. T. Nguyen, and K. R. Park, "Periocular recognition based on lbp method and matching by bit-shifting," in *Advanced Multimedia and Ubiquitous Engineering*, J. J. J. H. Park, H.-C. Chao, H. Arabnia, and N. Y. Yen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 99–104. [Online]. Available: https://doi.org/10.1007/978-3-662-47487-7_15

[141] K. Pontes Cotta, R. Sena Ferreira, and F. M. G. FranÃ§a, "Weightless neural network wisard applied to online recommender systems," in *2018 7th Brazilian Conference on Intelligent Systems (BRACIS). DOI: 10.1109/BRACIS.2018.00067*, 2018, pp. 348–353.

[142] A. Castiglione, G. Grazioli, S. Iengo, M. Nappi, and S. Ricciardi, "Dependable person recognition by means of local descriptors of dynamic facial features," in *Dependability in Sensor, Cloud, and Big Data Systems and Applications. ISBN : 978-981-15-1304-6*, G. Wang, M. Z. A. Bhuiyan, S. De Capitani di Vimercati, and Y. Ren, Eds. Singapore: Springer Singapore, 2019, pp. 247–261.

[143] S.-L. Wang and A. W.-C. Liew, "Physiological and behavioral lip biometrics: A comprehensive study of their discriminative power," *Pattern Recognition*, vol. 45, no. 9, pp. 3328 – 3335, 2012, best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320312000787

[144] C. P., "Security and privacy in biometrics: Towards a holistic approach." *Campisi P. (eds) Security and Privacy in Biometrics. Springer, London.*, 2013. [Online]. Available: http://doi-org-443.webvpn.fjmu.edu.cn/10.1007/978-1-4471-5230-9_1

[145] F. Cascetta and M. D. Luccia, *Sistemi di identificazione personale*, 2004. [Online]. Available: http://archivio-mondodigitale.aicanet.net/Rivista/04_numero_due/Cascetta_p.44-55.pdf

[146] K. Messer, J. Kittler, M. Sadeghi, S. Marcel, C. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, J. Czyz, L. Vandendorpe, S. Srisuk, M. Petrou, W. Kurutach, A. Kadyrov, R. Paredes, B. Kepenekci, F. B. Tek, G. B. Akar, F. Deravi, and N. Mavity, "Face verification competition on the xm2vts database," in *Proceedings of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication. ISBN: 3540403027*, ser. AVBPA'03. Berlin, Heidelberg: Springer-Verlag, 2003, pp. 964 – 974.

Acknowledgements

Eccoci arrivati alla pagina piú informale della tesi: i ringraziamenti.

Questo percorso é stato molto stimolante; continuare ad approfondire e studiare un campo che mi appassiona mi ha fatto solo capire quanto poco ne sapessi e quanto ancora ho da imparare. Dovrei ringraziare una lista infinita di persone perché in un modo o nell'altro in tanti mi hanno supportata e spinta a migliorarmi ogni giorno in questo percorso di studi.

Il prof. Michele Nappi che mi ha sempre dato fiducia e lanciata nelle sfide piú grandi, insegnandomi che si impara facendo e che si puó sempre migliorare un risultato raggiunto.

Mia madre e mio padre, presenti in tutti i traguardi che ho raggiunto, sempre pronti a sostenermi e consigliarmi in tutte le scelte della mia vita.

Marco, il mio compagno di vita. Un amore che mi dá soluzioni e non problemi, sicurezza e non paura, fiducia e non dubbi. (cit. P. Coelho)

Mio fratello Silvio presente quando necessario, nonostante tutto.

Carmen che ha condiviso con me l'intero percorso di dottorato, una donna tenace e con un animo buono. Non solo una collega, ma anche un'amica.

Tutto il gruppo di ricerca Biplab & Co. nessuno escluso, in questi

anni hanno rappresentato per me una seconda famiglia e un punto di riferimento.

Augusto, Sissi, Anna, Lidia e Domenico: amici cari e sinceri che hanno saputo dirmi le cose giuste nel momento giusto.

...e tu, che in tutta la tesi hai cercato la pagina dei ringraziamenti per vedere se c'é il tuo nome. Magari non l'hai trovato ma se sei qui ti ringrazio con tutto il cuore per essermi stato/a accanto.

La borsa di dottorato è stata cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 (CCI 2014IT16M2OP005),
Fondo Sociale Europeo, Azione I.1 “Dottorati Innovativi con caratterizzazione Industriale”

UNIONE EUROPEA
Fondo Sociale Europeo