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

Soft biometrics: emerging traits and applications

CARMEN BISOGNI

SUPERVISOR: **PROF. MICHELE NAPPI**

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

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

Contents

The machine does not isolate man
from the great problems of nature
but plunges him more deeply into them.

A rock pile ceases to be a rock pile
the moment a single man contemplates it,
bearing within him the image of a cathedral.

- Antoine de Saint-Exupery -

Abstract

Biometrics has been a thriving field of Pattern Recognition for long. Both Academia and Business have thus focused their attention in the practical use of biometrics to advance and promote varying applications in forensics, security and surveillance, health-care, mobility, human-computer interaction (HCI), safety and trust, and automation and robotics all of much interest to Government, Finance, Education etc. As the biometric problems that have to be solved have become ever more challenging and sophisticated, the techniques involved have found help and inspiration in human perception. As such, soft biometrics are traits that are naturally used by humans to distinguish their peers. Enhance as well both identification and re-identification but avoid impersonation and disinformation. Soft biometrics are physical and behavioral traits that capture human characteristics that go beyond appearance, e.g., age, gender, and gait. This thesis aims to advance the state-of-the art in varying applications on how to estimate the head pose of a subject and assist in her face recognition including tributaries such as attention, gait analysis to estimate the gender, and human behavior to meter the extent of cooperation and interest. Complete processing pipelines including data capture, preprocessing and feature extraction, adaptation and classification, and decision-making are presented and comparatively evaluated to show merit and advantage compared with current state-of-the art methods. Such methods are further integrated and successfully applied among others to the purposeful interplay between social engineering and social humanoid robots.

Chapter 1

Introduction

Biometrics is nowadays involved in everyday-use applications. It represent an ancient concept related to the way in which humans perceive themselves and the others. However, over the centuries, the applications of biometrics and the concept of biometry itself has been always examined from different points of views, following the needs of the moment. In this work we proposed several methods developed during the three years of PhD, focused on biometrics in general and soft biometrics in particular. As can be appreciated from the following Chapters and Sections, soft biometrics are nowadays involved in the most innovative use of biometrics traits, embracing very wide application fields.

1.1 Biometrics: definitions and uses

Following the Biometric Consortium, biometrics is the "automatic recognition of a person on the basis of discriminant characteristics". Those characteristics can be defined by the meaning of the word "Biometrics", composed by bios, "life", and metron, "measure". Each characteristic, physical or behavioural, that is measurable falls in the area of biometrics. The biometrics market is an increasing sector, due to the crescent use of biometrics algorithms in mobile devices (See Figure 1.1).

The history of biometrics, however, is very ancient. In 200

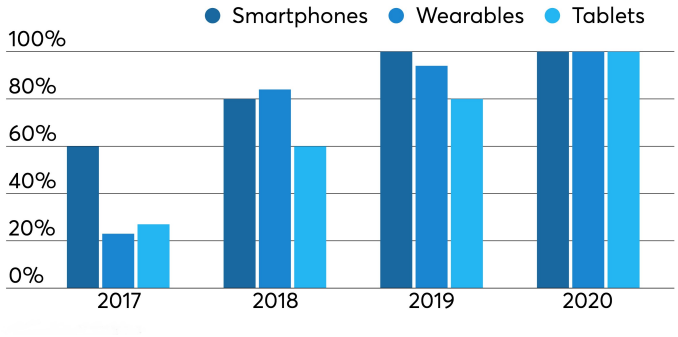

Figure 1.1: Involvement of biometrics on Mobile devices, evaluated in 2019. Data provided by [1].

A.D. in China, fingerprint was already used to authenticate official documents. But was only in 1870 with Alphonse Bertillon's studies, that the biometrics started to be used for recognition purposes, in a scientific way. The system he proposed was based on skeletal measures and was in use, with encouraging results until 1903. In that year, in fact, we have the first documented cases of misclassification in Biometrics. Two persons with the same name and very similar measures were misclassified leading to a problem in the identification of two different prisoners. For this reason, in the same year, the fingerprint was used in combination with those traits, by the study of Francis Galton and Edward Henry. From this historical moment, the use of biometrics has exploded leading to the market and the implications that influence our daily life. From a formal point of view, a Biometrics trait, to be such, should have the following characteristics:

- Universality: each person should have the trait.
- Distinctiveness: two person can not have the same set of characteristics for this trait.
- Permanence: the trait should be unchanged over time.
- Collectability: the trait must be measurable with numerical values and easy to collect.

- Performance: the automatic recognition of the trait should be advantageous in terms of computation resources and time.
- Acceptability: the biometrics acquisition of the trait should be well accepted by the most of the population.
- Circumvention: the biometrics trait should be robust to impersonation and fraud.

Not all the biometrics traits on the market are responsive to those requirements in the same way. As an example, for some of the most known biometrics traits, we can see how well they satisfy those characteristics, in Figure 1.2.

Biometric identifier / **Characteristics**	**Finger**	**Facial**	**Iris**	**Hand**	**Retina**	**Signature**
Universality	high	high	high	mid	high	low
Distinctiveness	high	low	high	mid	high	low
Permanence	high	mid	high	mid	mid	low
Collectability	mid	high	mid	high	low	high
Performance	high	low	high	mid	high	low
Acceptability	high	high	low	mid	low	high
Circumvention	mid	high	low	mid	low	high

Figure 1.2: How some of the biometrics traits responde to the biometrics characteristics. From the study in [2].

It is clear that, apart from the nature of each trait, the progress in methods and techniques lead to mless or more effective solutions. As a consequence, the market of those biometrics traits is constantly changing over time, to balance at the same time the costs of the recognition algorithms and acquisition devices, and the accuracy of the latter. As can be seen in Figure 1.3, from a 2018 research [3], the biometrics market is dominated by Fingerprint. This because of the low costs of fingerprint acquisition devices that can be also installed on a wide range of devices (doors, smartphones, notebooks, etc.). The second biometrics trait most used is the face. Also in this case we can find economical reasons, because the face can be captured by every camera on the market, even if with low resolution, due to the large surface of this trait.

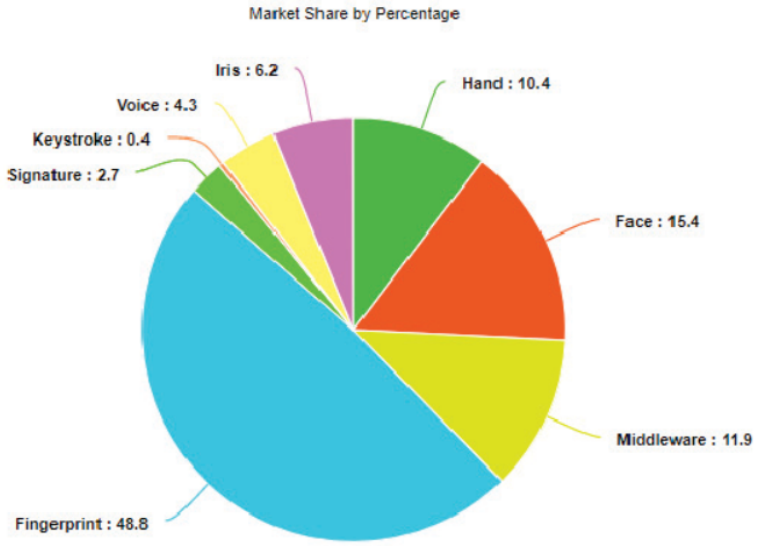

Figure 1.3: Biometric market in 2018, by involved technology.

Regardless of the biometrics system involved, the evaluation metrics to define the effectiveness of a biometrics framework are identified as three: False Acceptance Rate (FAR); False Rejection Rate (FRR) and Equal Error Rate (ERR). The FAR is defined as the percentage of identification instances in which unauthorised persons are incorrectly accepted. The FRR is defined, on the other hand, as the percentage of identification instances in which authorised persons are incorrectly rejected. In systems build for security purposes, it is clear that a higher FAR is more dangerous than a higher FRR. Those tho indices, however, impact on each other in an inverse relation. The equilibrium point in which the FAR is equal to FRR is called Equal Error Rate (EER).

In the following sections we will analyse the specific field of biometrics that we decided to investigate: soft biometrics. We will present the motivations that led us to choose them as the keypoint of this thesis and our point of view in conducting experiments and evaluations.

1.2 Contexts and Motivations

From the definition of *identity* from the Cambrige Dictionary, the identity is "who a person is, or the qualities of a person that make him or she different from others." In this sense, a soft biometrics is a trait naturally created by humans to distinguish their peers. Differently from the classical biometric traits, soft biometrics are not sufficient to distinguish exactly the identity of a person, however they extend the potentiality of biometrics, carrying with it some of their advantages. Formally, the definition of soft biometrics is, thus, "characteristics that provide some information about the individual, but lack the distinctiveness and permanence to sufficiently differentiate any two individuals". This shortfall, that can be seen as a defect, is instead the particularity that makes us able to apply soft biometrics in contexts where is not efficient or effective to proceed with biometrics. The transition to soft biometrics is further explained as follow.

1.2.1 From strong to soft traits

Recognize the identity of a user is undoubtedly a matter of interest in several contexts. We can in particular split the applications in two categories: recognition and verification. In verification applications the user declare his/her identity, to access at a device or at a physical reserved space. In this case a biometrics method should verify if the trait provided by the user correspond to the trait stored in a model. Examples of verification can be found in airports, where the trait is compared with the one on the passport [4], to access at personal devices [5] [6], or to sign contracts [7]. In some cases, the trait is combined with other access key only known by the user, as we proposed in [8]. In this case we fused the face biometrics with a generated product of prime numbers to create a key that can benefit from the advantages of both cryptography and biometrics. On the other hand, the recognition task is more complex than verification, because the verification is a one-to-one check. The response to this check is binary, in other words the

subject is what he/she declared to be or not. In the case of recognition, the check to perform is one-to-many, for this reason the user to be identified can be any one of the ones stored in the model or, even, no one of them. Recognition task can be used when the user is non-cooperative, as in surveillance scenario [9], or in forensics [10]. In both verification, but especially in recognition, after the trait detection [11], the data are often affected by noise and need to be preprocessed in order to enhance the quality of the acquired images. We also analysed this further aspect of biometrics in the last years, in particular on Iris [12], but other researches in literature were also conducted on different traits often affected by this problem, as fingerprint [13], or voice [14]. In some cases, even if the denoising is applied, a single biometric trait is not sufficient to obtain the desired accuracy. For this reason, multibiometric systems were born in the field of recognition. Those systems have also the benefit to be harder to fool. We will present some multibiometrics applications in Chapter 4, in this case also involving soft biometrics, but have also focused in past years in the fusion of classical biometrics traits, face, ear and fingerprint in particular [15] [16], demonstrating that a multibiometric system improves the verification accuracy compared to a single biometrics.

In some cases, recognizing the identity of the user is not possible, regardless of the accuracy or the performances of the method involved. It can happen for two reasons in particular: the database of the users to check is so large so to compare an identity with each of them is too expensive in terms of time to make recognition useless; the identity to recognize is not in the database because it is an unknown subject. Both problems introduced can be solved using a physical soft biometrics. In the first case, to differentiate the population split them in classes (e.g. by the gender, some facial attribute, the hair color etc.), can significantly speed up searching the database. On the other hand, more often in security and surveillance scenario, it is not possible to obtain an enough quality of the traits to perform recognition, however a set of soft traits detected can be sufficient to help the purpose [17].

In the case of behavioural biometrics the motivations are even

more different than the classical biometrics. In this context in fact, we are not interested in all at the user identity, but at the users behaviours. This approach finds its application in contexts related to security or automatic surveillance. This is a very rich field connected to emotion recognition [18], action recognition [19], behaviour recognition in groups [20].

1.2.2 Contributions

Our contributions to the state of the art in the field of soft biometrics are related to their use in security. In particular we are interested in surveillance videos. The particularity of the surveillance videos is the fact that, in the best case, the user is non-cooperative, and in the worst, is anti-cooperative. On the other hand, surveillance videos may came from very different sources and cover hours and hours of recording. This is an aspect that requires the use of automatic algorithms able to work in real-time or near real-time. The methods presented here start from those assumptions, and, during their development, have also shown interesting application to other context security-related that we further explored with the applications to humanoid robots. The research covers apart from the framework to be applicable to robots, two main methods: head pose estimation and soft traits from gait.

1.3 Outline of the Thesis

The proposed work of thesis is organised as follow. In this first Chapter we introduced the fundamentals of biometrics and we gave an idea of our motivations and contributions. The next three Chapters (from Chapter 2 to Chapter 4) define the core of the thesis:

- Chapter 2: is focused on the Head Pose Estimation. We present three method built to solve this task using RGB images as initial data. The methods presented have different

characteristics and are focused on various aspects (robustness, independence from the dataset, real-time applications).

- Chapter 3; is focused on soft biometrics traits applicable to the gait. In particular we present a method to detect the gender from the gait skeleton and another method to detect the cooperativeness of a subject using space-temporal variable of the gait skeleton. Here we also introduce a dataset collected by us in the last years particularly suited for those purposes.

- Chapter 4: is focused on the application to robotics of the previous methods presented. In particular, the humanoid robot Pepper is used in two different frameworks with opposite purposes to highlight the contribution of soft biometrics in the social robot field.

We will drawn our conclusions in the last Chapter, Chapter 5, in which we also propose some works in progress and future developments about the above introduced methods.

Chapter 2

Head Pose Estimation: a multifunctional biometric

Face biometrics has a very wide literature and other face-related measures were taken under consideration in biometrics, as a consequence. One of the more recent field is undoubtedly the Head Pose Estimation (HPE). This particular biometric investigation can be classified at the same time as a Preprocessing step to support face recognition and a Soft Biometrics trait to detect attention, cooperation and so on. As can be seen in Figure 2.1, the advent of Machine Learning techniques, using training, strongly impacted the methods used in this field.

In this Section we will introduce in detail the recent advancements in HPE. The research conducted by the candidate in the last three year are strongly related to HPE and make a positive contribution to the state-of-the-art.

2.1 HPE background

As previously introduced, a biometric trait, should have some characteristics. We can choose a biometric trait instead of another also by the level they satisfy those characteristics. As an example, soft biometrics and classical biometrics can be distinguished by the fact that the firsts are by nature not strictly related to a

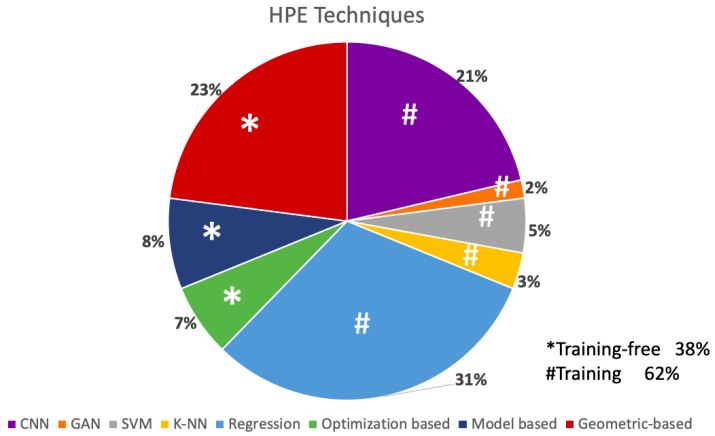

Figure 2.1: Techniques used in HPE, split in training and training-free.

specific individual. For this reason their power in recognition is indirect. Also HPE has this behaviour, so we can assimilate it to soft biometrics. However, it scores higher than most soft biometric traits in terms of Measurability, because we are able to estimate its numerical value. Also other properties are well afforded by HPE. If we think to Universality, the head rotation is observable in each individual and it is not dependent on his/her age. On the other hand, if we think to Acceptability, HPE can be considered the same as the acceptability of face or the ear because they are the only recognition traits involved. Based on this two biometrics trait involved, an impact in HPE is observable in individuals with strong differences along those traits. It is clear that the relevance of those differences depends on the particular characteristics of the used method. The main procedure to build a HPE technique, from the data to the evaluation of the errors can be summarized as follows:

- **Acquisition and Labeling**: During this step the trait involved is acquired and categorized. In general the acquisition of HP involvse the same techniques and devices of face acquisitions, as cameras, depth cameras, near infrared cameras

etc. For this reason, the acquisition results very simple to realize, however, on the other hand, Labeling is a more complex task. When depth data are not available, it is difficult to estimate manually the rotation angles of HP, because it is not human perceptible if it is small. The human involvement in this task is, thus, not recommended. This is a characteristic that differentiates HPE from other biometric traits like gender, age or facial features, that are well visible by an observer.

- **Preprocessing**: This step is necessary to normalize the data to be ready for the HPE model. It varies a lot among different architectures.

- **Pose estimation**: This is the core of the method in which the image, or the preprocessed data, is used to estimate HP. The input may be different, but the final output will always be the rotation in terms of angles.

- **Evaluation of errors**: In general in this field, the error is represented as the angular difference between the true label and the estimate.

The above mentioned procedure are also summarized in Figure 2.2.

As above-mentioned, the Head Pose variation is measured in rotation angles. The center of the rotation is, ideally, the center of the head, however it is more often the nose, because 3D data are not always available. The center will have coordinates $O(0, 0, 0)$, and it is the only fixed point of the rotation. The possible angle of rotation are 3, since the head is a 3-dimensional object, by nature. In this system, the axes are, by convention, represented by the Motion Imagery Standards Board (MISB) [21] as pitch, yaw and roll. **Pitch** axes is also called the transverse or lateral axes, and the rotation about this axes is called pitch. **Yaw** axes is also

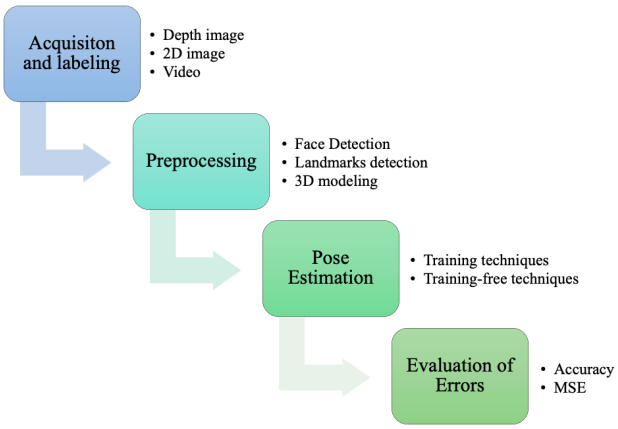

Figure 2.2: The main steps of an HPE framework.

called the normal or vertical axes, and the rotation about this axes is called yaw. **Roll** axes is also called the longitudinal axes, and the rotation about this axes is called roll. In figure 2.3 the conventional direction of pitch, yaw and roll can be appreciated. It is clear that not all the possible rotation along three axis can be taken under consideration, since some of them are not possible to perform by an human head (e.g. $\pm 180°$ in yaw) Although there is individual variation, most people are able to turn their head $\pm 90°$ in yaw, $\pm 45°$ in roll and $\pm 30°$ in pitch.

Despite the amount of way to represent a 3D rotation, the most popular among the HPE datasets and algorithm are the Euler angles, the rotation matrix and the quaternions.

2.1.1 Euler angles

The Euler angles, firstly introduced by Leonhard Euler to describe the orientation of a rigid body in space, has two categories: Proper Euler angles and Tait-Bryan angles. The HP rotation follows, as above-mentioned the rules of MISB. It implies that the Tait-Bryan angles are used to describe the rotations. We define as x,y and z the original axes and X,Y and Z the axes after the rotation. The lines that represent the intersection between plan xy and YZ is

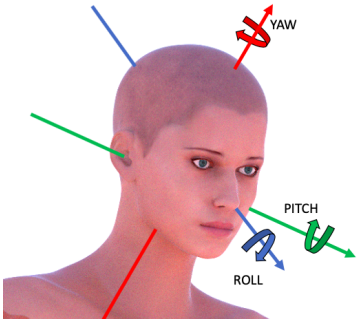

Figure 2.3: Pitch, Yaw and Roll

called the line of nodes N. With this conventions we can define the Euler angles as:

- ϕ the rotation angle between x and N, covering a range of 2π
- θ the rotation angle between z and Z, covering a range of π
- ψ the rotation angle between N and X, covering a range of 2π

Each increment on one of the defined angles is called an Euler Rotation. Rotations are not commutative, for this reason their sequence and their combinations lead to different conventions. The convention used to define Pitch, Yaw and Roll in Tait-Bryan angles is the intrinsic rotation z, y',x'', where y' represent the position of the axes y after the first rotation and x'' represent the position of the axes x after the first and the second rotations. Since the rotation is intrinsic, by definition each following rotation after the first one is considered around the new position of the axes.

2.1.2 Rotation matrix

The particularity of the rotation expressed by a rotation matrix, is the use of a single angle θ. A rotation is defined as:

$$R_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix} \quad R_y(\theta) = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix}$$
$$R_z(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.1}$$

where $R_x(\theta)$ is the rotation about the x axes of θ degrees, $R_y(\theta)$ is the rotation about the y axes of θ degrees and $R_z(\theta)$ is the rotation about the z axes of θ degrees. In terms of pitch, yaw and roll, by convention, $R_z(\theta)$ is called yaw, $R_y(\theta)$ is called pitch and $R_x(\theta)$ is called roll. A simple multiplication between matrices can give us a final rotation matrix, where we can distinguish yaw, pitch and roll, assuming they are α, β and γ respectively. The final matrix will have the form:

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} =$$
$$\begin{bmatrix} \cos\alpha\cos\beta & \cos\alpha\sin\beta\sin\gamma - \sin\alpha\cos\gamma & \cos\alpha\sin\beta\cos\gamma + \sin\alpha\sin\gamma \\ \sin\alpha\cos\beta & \sin\alpha\sin\beta\sin\gamma + \cos\alpha\cos\gamma & \sin\alpha\sin\beta\cos\gamma - \cos\alpha\sin\gamma \\ -\sin\beta & \cos\beta\sin\gamma & \cos\beta\cos\gamma \end{bmatrix} \tag{2.2}$$

2.1.3 Quaternions

Quaternions, firstly introduced by William Rowan Hamilton, are often known as versors. They are nowadays mainly involved in videogames developing. Basing on the concept of complex number and complex plane, we can introduce the general form to express quaternions as

$$q = s + xi + yj + zk \tag{2.3}$$

where $s, x, y, z \in \mathbb{R}$ and i, j, z are imaginary number that follow the rules

$$i^2 = j^2 = k^2 = ijk = -1 \tag{2.4}$$

and

$$ij = k, jk = i, ki = j, ji = -k, kj = -i, ik = -j \tag{2.5}$$

The quaternions are more simpler to compose, respect to Euler angles, and have a more compact representation, compared to rotation matrix. The quaternions represent the rotation using four terms that can be defined as

$$q_0 = e_x \sin\frac{\theta}{2},\ \ q_1 = e_y \sin\frac{\theta}{2},\ \ q_2 = e_z \sin\frac{\theta}{2},\ \ q_3 = \cos\frac{\theta}{2} \tag{2.6}$$

where (e_x, e_y, e_z) is the principal axis and θ is the principal angle. Quaternions are related to pitch, yaw and roll by the following formulas:

$$\begin{aligned} yaw &= \tan^{-1}\frac{2(q_0q_1 + q_3q_2)}{q_3^2 - q_2^2 - q_1^2 + q_0^2} \\ pitch &= \sin^{-1}(-2(q_0q_2 - q_1q_3)) \\ roll &= \tan^{-1}\frac{2(q_1q_2 + q_0q_3)}{q_3^2 + q_2^2 - q_1^2 - q_0^2} \end{aligned} \tag{2.7}$$

There is not a standard by the angles formulation among different datasets. For this reason, the HPE method developer usually chooses a representation and, by the transformation formulas they normalize the label of the dataset accordingly.

On the other hand, the errors related to HPE are evaluated in the same ways. In particular, the errors in yaw, pitch and roll, are represented by the angular values of the differences between the estimated pitch, yaw and roll and the true pitch, yaw and roll for each head in the data. Those are next combined to obtain a mean error test for each axes, using a simple mean operation:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|\theta_j - \hat{\theta}_j| \tag{2.8}$$

where θ_j is the ground truth, i.e the true angular value and $\hat{\theta}_j$ is the prediction, i.e the predicted angular value. The same formula is often used to compute also the total MAE along the three axis together.

Another concept to keep well in mind to perform HPE, is that the system of reference, we called x, y, z axes, is not available. This because, as anticipated, the facial traits of a subject can be strongly different. Each algorithm chooses how to calculate the system of reference before estimating HP. It is clear that if this initial estimation is wrong, each following HPE will be affected.

The facial shape has also an impact on facial movements [22]. This differentiates the HP, and method operating on 2D images can be particularly affected by it.

2.2 Data and devices

Since the HP is a biometric trait strictly related to face, techniques and devices used to collect it are more or less the same used for the face. A high discriminant component, compared to face, is the need for pose labels, yaw, pitch and roll. The identity of a subject or his/her soft biometric traits in the face, are easily recognisable from who collect the dataset. However, as previously introduced, the exact rotation angles of the head are not easy to classify by human. This problem can be solved using devices able to collect depth information. Differently from a traditional RGB image, that has three channels of information about the color, a depth camera has an additional information that provides the distance of the subject from the camera. In particular, for HPE are usually preferred RGBD image, which have both 3 color channels and depth channel. The depth information can be captured using different techniques. Which one pick will depend on the environment. Three kinds of depth cameras can be distinguished: struc-

tured light; time of flight; stereo camera. A graphical summary of how these camera work can be found in Figure 2.4.

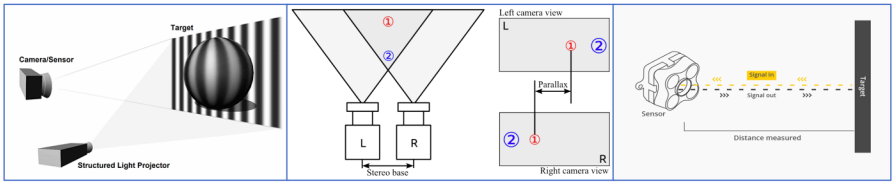

Figure 2.4: How depth camera work: from left to right, structured light, stereo camera, time of flight.

Structured light and time of flight are the most used techniques to acquire HP images. We can find the structured light camera in the popular Kinect 1. They use an emitter in the scene, for example infrared light. A known pattern is projected in the scene, and depending on the deformation of this pattern, the distance of the surface from the camera can be calculated. Those cameras has a relatively low cost [23]. On the other hand, time of flight cameras represent the more recent technique in this sense, and we can found it on Kinect 2. In Time of Flight camera, the device emit some kind of light, the exact type depend on the device. This light sweeps over the scene and the time the light needs to reach the surface and go back to a sensor on the camera, will give the essential information to extract depth. In other words, differently from a structured light camera in which is known the pattern to project, in this camera the speed of light is the known information. The advantage of these cameras is that they are able to measure the depth also at a long distance, using the power and the wavelength of the light.

Two datasets captured by Kinect 1 and Kinect 2, are Biwi and SASE, respectively.

The Biwi Kinect Head Pose Database (BIWI) [24] contains 24 sequences for a total of over 15K images. People captured in the datasets are 6 females and 14 males, for a total of 20 people in which 4 of them are recorded twice. For each RGB image captured, for a dimension of 640x480 pixels, also the depth image is captured,

with the same size. The variation of the head pose is between $-75°$ and $+75°$ in yaw and $-60°$ and $+60°$ in pitch. To annotate the ground truth of the pose, they used the center of the head in 3D and the rotation angle of the head. To this purpose, they used an automatic system of Faceshift, a startup specialized in capturing and transferring facial expressions on 3D models in real-time.

SASE database [25] contains 16-bit depth frame of 424x512 pixels and RGB frames of 1080x1920 pixels. The subjects collected are 50, 32 males and 18 females from 7 to 35 years old. The mean amount of frames per subject is 600. They provide Pitch, Yaw and Roll using five blue stickers placed on the chin, the tip of the nose, two on cheeks and one between eyebrows. SASE cover a range between $-75°$ and $+75°$ in yaw and $-45°$ and $+45°$ in pitch.

Even if collected with a recent technology, SASE is not very popular.

If we consider the recent literature about HPE using depth images, a dataset called ICT-3DHP is the most popular after BIWI. ICT-3DHP [26] is a dataset of head poses collected with the Kinect 1. The dataset is composed of 10 RGB-D videos for a total of about 1400 frames. The image resolution is 640 x 480 pixels. The labels of this dataset were obtained using a Polhemus FASTRACK. This device can provide an accurate head tracking in real time, does not require user calibration and operates at a distance of about five feet from the standard source.

Despite the fact that depth images provide more accurate ground truth labeling, the real-case study is represented by RGB images. HPE from a single RGB image represents the more challenging problem to solve. Some depth datasets can be also used in this task, as BIWI, however with a very controlled environment. A very popular RGB image dataset with HPE annotations is AFLW.

The Annotated Facial Landmark in the Wild (AFLW) [27] is a dataset of about 25K images collected from the web and, as a consequence, they have a large variation in pose, expressions, age, gender, ethnicity. The faces are annotated with 21 landmarks for each face and have different resolutions. This dataset provides face rectangle and the face ellipses, used in a POSIT algorithm to

estimate the head pose. Because of the heterogeneity of the data, this is one of the most used and challenging dataset in 2D HPE.

As a more accurate version of AFLW respect to the HP, we can find AFLW2000 that includes the first 2000 images of AFLW annotated using a 3DMM fitting and can be downloaded at [28].

As in the case of Depth Datasets, not always the recent most used dataset are the ones with an accurate ground truth or a higher number of frames/subjects.

Is the case of Pointing'04. Pointing '04 Head Pose Image Database [29] was collected using 15 subjects by the PRIMA Lab. For each subject there are 2 series of 93 images, for a total of 2790 images. The images have a resolution of 384x288 pixels and roll rotation angle is not contemplated. The dataset has a limited number of poses, in particular 9 for pitch and 13 for yaw, and their combination between $-90°$ and $+90°$ degree. To obtain pose with known labels, the authors have put markers in a room and ask to stare at the 93 post-it notes without moving his eyes. As it is clear, these annotation are not precise by construction and in addition some users move the eyes instead of the head during the experiments. Despite of these characteristics, Pointing'04 is very popular also in recent applications.

On the other hand, there is the 300W_lp Dataset [30] which include 68 landmarks localization. 300W_lp collect different datasets, in particular AFW, LFPW, HELEN, IBUG and XM2VTS. The large pose variations and annotations are available for each of this datasets but not for XM2VTS. In particular there are 61225 images, 1786 from IBUG, 5207 from AFW, 16556 from LFPW and 37676 from HELEN. The pose annotation are in Euler angles in radiant and the resolution of the images are different due to the different datasets on which they belong. It is better to underline that in this dataset the majority of head rotation are artificially obtained and lead to deformation in the facial structure.

As a natural extension of 2D images, the application of HPE to video has the aim to use multiple frames to understand the user behaviour from the pose. The video datasets have not gained a lot of popularity in HPE domain. However, it is not absurd

to expect a future increase in their use, since the hardware and machine learning techniques are making HPE faster and faster. The most recent video annotated datasets are Gotcha-I, UPNA and UBIPOSE.

GOTCHA-I [31] was collected as different video sequences of 62 subjects in 11 different environment, for a total of 682 videos containing both faces and bodies. From a 180° video, the authors build for each subject a 3D model and then obtain 137826 labeled frames with 2223 HP per 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 UPNA Head Pose Database [32] is composed of 120 videos of 10 subjects, 6 males and 4 females. Since this dataset is born for head tracking and pose estimation, the authors collected 6 guided-movement sequences and 6 free-movement sequences. The guided movement represent the pure translations in pitch, yaw and roll, and the free movements the combination of the latter. There are 300 frames per video, each associated to a ground truth for the pose and 54 landmarks. They used the initial frame of frontal face as a keypoint to label the head pose. They claim an error with the original ground truth of 0.83°, 0.86°, and 1.05° in roll, yaw, and pitch respectively.

Finally, the UBIPose dataset [33] was collected using a Kinect to obtain labels. The videos are 32 simulating a reception desk environment, but only 22 of the 32 video are annotated. The head pose is available for about 10K frames. The resolution is from the Kinect and it is of 640x480 pixels.

In Table 2.1 we can find the main characteristics of the above-mentioned datasets. The popularity represents the number of recent HPE papers, that, to the best of our knowledge, used the datasets in the last five years.

Since the more challenging HPE came from a single RGB image, as can be seen in the amount of literature on this matter, we will introduce some preprocessing techniques that are common in more than one HPE method.

Table 2.1: HPE Datasets that contain Pose Annotation. The Popularity of each dataset is calculated using the amount of recent depth HPE method that use it, to the best of our knowledge (last five years).nd is "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

2.2.1 Preprocessing steps

There are three main steps that are commonly used as preprocessing. They are often sequential but not all of them are necessary.

The first step is the face detection. This step, performed with different techniques, is common at every HPE algorithm. This because, once the face is localized, other noisy information as background or other body parts, can be excluded. One of the most used technique over the years is undoubtedly the Viola-Jones method [34]. Firstly introduced to detect objects, Viola-Jones method was successfully adapted to detect faces. It takes advantages from the fact that all faces have some similar properties like light and dark zone on the face. These features, called Haar, are evaluated by a representation of the image called integral image, that works in constant time. To optimize the number of features to evaluate, an Adaboost training is performed, building a strong classifier composed from simpler classifiers. This is followed by a cascade architecture to progressively select the image part to evaluate. Viola-Jones algorithm is wide used because it is fast and robust,

however if the head pose is extreme (more than 60°), this method would not succeed in find the face location. More recent, are algorithms involving Support Vector Machine (SVM), as Histogram of Oriented Gradients (HOG). In this case, the SVM finds the function that linearly splits positive and negative samples. Then, using a sliding window technique, at each window HOG features are computed [35].

Right after face detection, some algorithms perform the landmark prediction. The landmarks are defined as facial keypoints that locate particular features in the face. The number of detected landmarks can be very different depending on the aim of the algorithm involved. They can be 5 or 6 that give the center position of the nose, eyes, mouth [36], or more than 60. The algorithm in [37] is one of the most used landmark detector. In this method, a cascade of regressors is used to estimate 68 landmarks. In particular, this is not a detector, but a predictor. This means that in any case, as occlusions, low quality images etc, the landmarks will be 68. This method has been trained on a part of the above mentioned 300_lp dataset, iBuG in particular. As a consequence, it can be used for 2D RGB images. This kind of methods differ from RGB to depth images. For depth images, a popular landmark detector is the Shape Regression with Incomplete Local Features (SRILF)[38].

The landmark detection produces a results similar to the one in Figure 2.5. The data are expressed in terms of spatial coordinates.

2.3 Recent advances in HPE

In this section we will explore the recent advances in the state of the art. Recently methods reached accuracy so high that they focused more on the application fields and the performance, to differentiate a method to another. The contribution reached with the involvement of the author of this thesis, will be extensively discussed in Section 2.3.2.

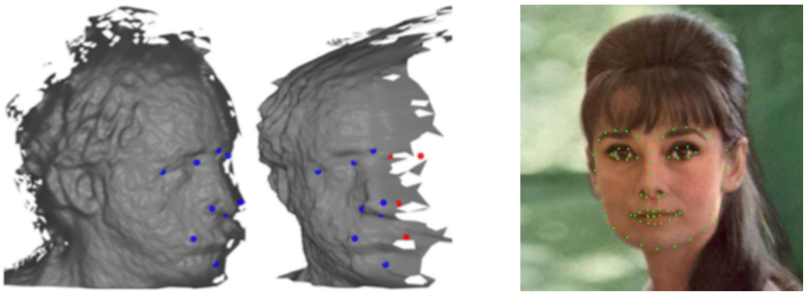

Figure 2.5: On the left 10 keypoint detected on a depth image [38], on the right 68 landmarks detected on a 2D RGB image [37].

2.3.1 State-of-the-art

The attention of the state of the art about HPE methods differentiate itself by the initial data. Different kinds of data are useful for different purposes.

It is clear that the use of depth image is undoubtedly an advantage to solve HPE problem. But differently from methods using both RGB than Depth images to perform HPE, the recent works focused on the use of Depth images only. The method in [39] is called POSEidon and works in real time a 30 fps. A initial CNN extracts the face from the depth image, than, this architecture reconstruct the appearance-like face from depth. Other CNNs are used to estimate the pose in Pitch, Yaw and Roll, without the use of facial features. They present their results on Biwi, ICT-3DHP dataset and their homemade dataset Pandora. The characteristics of this method are justified to its application field. In fact, the authors focused on the detection of driver attention, since the pose they estimate is also used for the shoulders.

This method and application field was recently confirmed in the development of POSEidon+ [40]. In this case, the authors used 3 depth images to feed a CNN composed by some steps that are similar the previous version but iterated over the three images. The three head images provided by the initial CNN are merged using a Face-From-Depth architecture that reconstructs a grey-level face using the depth images, as in [39] and a Motion im-

age as support, since the captured depth images are consecutive. As in the previous work, another CNN uses those information to estimate the HP angles. They also present a Deterministic Conditional GAN (Generative Adversarial Network) to hallucinate a face from the depth images. The experiments were conducted on Biwi, ICT-3DHP and Pandora, all of them with the driver attention purpose.

As we can see, those methods involves both training algorithms, that learn how to classify head pose using a large set of data as reference. This trend is also observable in 2D RGB images HPE. In some cases, training techniques are used in combination with Histogram Of Gradient (HOG) features. In others, regression networks or CNN are preferred. The methods that prefer HOGs can perform the following steps in different ways. In [41] HOG are used to optimize the alignment to avoid errors derived from a wrong face localization and the compensation of offset errors is made by an iterative prediction using Gaussian locally-linear mapping (GLLiM). In [42] HOGs are combined with a Support Vector Regression (SVR) trained with extremely low resolution images. HOGs can be also combined with generalized discriminating common vectors and continuous local regression, as in [43], with other features techniques as the Uniform Pattern of Local Binary Pattern (UP-LBP), as in [44], or with Haar features and speed up robust features (SURF) as in [45]. As anticipated, other methods prefer regressions learning algorithm, often in combination with other techniques. In [46] a CNN, VGG-16, is used in combination with a Gaussian mixture of linear inverse regressions that can work also with relatively small datasets. In [47], more than one regression is used to detect landmarks and to estimate the head pose called Coupled Cascade Regression (CCR). A multi-loss network is also used is [48], where Euler angles are directly estimated from face image intensities using ResNet50. Again, Multi-regression is also the strategy of [49], that focused on the loss of the regression net. In [50], Heatmap-CNN trained by 3D-pose, face's visibility and fiducials are used to learn regressors and to obtain key-points estimation and pose prediction. Other architectures are more fo-

cused on CNNs, as in [51] in which a GNet is trained to obtain the face location,a preliminary pose and few landmarks. Then, an LNet refines this work learning local CNN features and predicting the final head pose. The CNN used as basis can be popular ones, as AlexNet, born for other purposes, as used in [52]. Other authors, even if starting from a standard CNN architecture, focus their attention to how to generalize the CNNs, training and testing the network using different datasets, as in [53]. Or, again, fusing the hidden layers of a first CNN by means of a second CNN along with a multi-task learning algorithm operating on the fused features, as in [54].

Finally, if we consider video applications, as the best frame selection, or attention detection, we will find in literature works using the temporal variable. Starting from an RGB video, the aim may be the subject tracking using the HPE, as in [55]. Here, the video are captured using a limited resources device as a tablet. The pose is estimated using a POSIT algorithm and the use of consecutive frame is necessary to speed up the method on those devices. The application to a mobile device was also explored in [56], were also the facial features was captured. The low computational requirements lead to the use of an Haar-based detection and the use of four different classifier to estimate the HP. On the other hand, if we start from a RGB-D video, we can find works as in [57]. In this case nine frames from the video are used and an online face template reconstruction is made. Here, the tracking is not the aim, but a tool to minimize the HPE errors.

2.3.2 Our contribution to literature

In the last three years, we developed three different techniques to perform Head Pose Estimation. Two of them are based on the distribution of the landmarks, 68 facial features point coordinates as mentioned in Section 2.2.1. The last one is based on the pixel values and on the concept of self-similarity. All of them work without a training step, being training-free methods.

2.3.2.1 WSM: The Web-Shaped Model

The first method we present can be found in [58]. This method is focused on both the minimization of the HPE errors than the real-time applications. This method is composed of three main steps executed in cascade:

- The face is detected together with the position of 68 landmarks representing facial features points;
- A spider-web model is applied to evaluate the distribution of the landmarks over the face, using three main parameters: the distance of each landmark from the nose, the distance of each landmark from the vertical axes, the relative position of each landmark respect to the nose (bottom-right, bottom-left, top-right, top-left).
- The resulting model is compared to the ones stored, to classify the HP.

Those steps are depicted in Figure 2.6.

Figure 2.6: The main steps of WSM.

In more detail, the facial landmarks are extracted using the method in [37]. The extracted coordinate of this methods are ordered depending on the facial feature, as in Figure 2.7. This method is chosen because it results robust to head pose variations.

Subsequently, to build the web-shaped model, the nose tip, corresponding to the 33th landmark was chosen as the center of

Figure 2.7: 68 numbered landmark and their positions.

the spiderweb $O = (x_{33}, y_{33})$. After this assumption, the other points, can be defined as $P_j = (x_j, y_j)$ with $j = 1, .., 68$ and $j \neq 33$. To find the spider dimension, we define the distance from the nose as $r = d(O, P_i)$, and then, the farthest landmark has index $i = argmax_j(O, P_j)$. As previously introduced, we use 3 parameters to define the landmark location in the spiderweb. The distance of each landmark from the nose represents the *circle* to whom the landmark belongs, the relative position respect to the nose represents the *quarter* where the landmark is located and the distance of each landmark from the vertical axes represents the *slice* to whom the landmark belongs. To better explain how these parameters are obtained, we will suppose that we want to build a spiderweb composed of n equidistant circles and m slices per quarter. To obtain the circle, we consider the nose as the center of the spiderweb, O and r_i the radius of the i-th circle, with $i = 1, ..., n$ from the outermost to the innermost. Considering the distance between the nose and the landmark, the latter belongs to the smallest circle C_i containing it. In other words, to the ring limited by C_i and C_{i+1}. The quarter is obtained considering, as previously defined, the relative position of the landmark respect to the nose, and it can have value 1, 2, 3 or 4. And, finally, the slice is obtained though the angular coefficient. Let α be the angle

between the segment $\bar{OP}$ and the vertical axes, and $\theta = 90°/m$ the width of each slices in a quarter, the point will belong to the slice $s = \lceil \alpha/\theta \rceil$. The subdivision can be seen in Figure 2.8. It is also visible the numbering chosen to refer to the sectors.

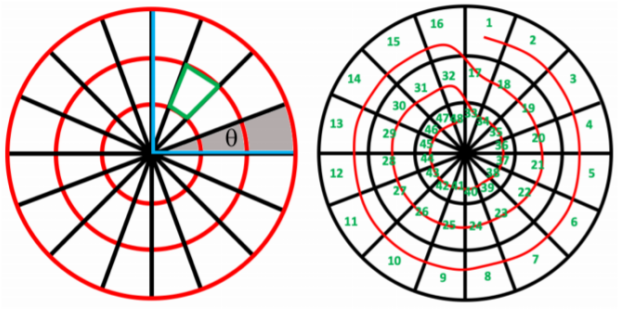

Figure 2.8: On the left the parameters of the spiderweb, on the right the used numbering.

Once we have the sectors of each landmarks, the features vector is simply built counting how many landmarks belong to each sector. An example of the array generation can be seen in Figure 2.9, where there are 4 circles and 4 slices per quarters.

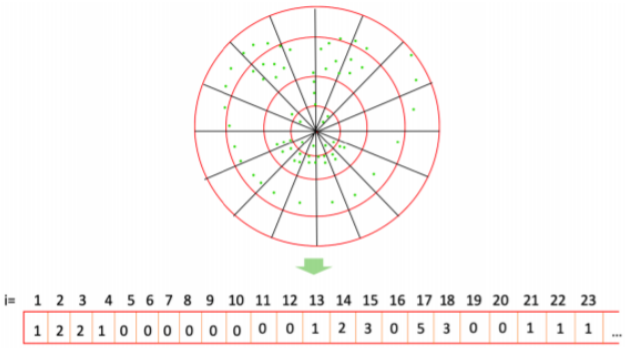

Figure 2.9: On the top the spiderweb, on the bottom the generated array.

Using this techniques, for each pose we can generate the corresponding features array. To obtain a high pose variation, independent from the dataset used to test the method, we used a synthetic model as in Figure 2.10. 2223 head poses with a step of 5 degrees in pitch, yaw and roll, were computed into the above mentioned array. The obtained 2223 arrays are subsequently used as a reference model to obtain the HP. The rotation ranges of the model, and, as a consequence, of the used test set, are the following:

- pitch: from -30° to +30°
- yaw: from -45° to +45°
- roll: from -20° to +20°

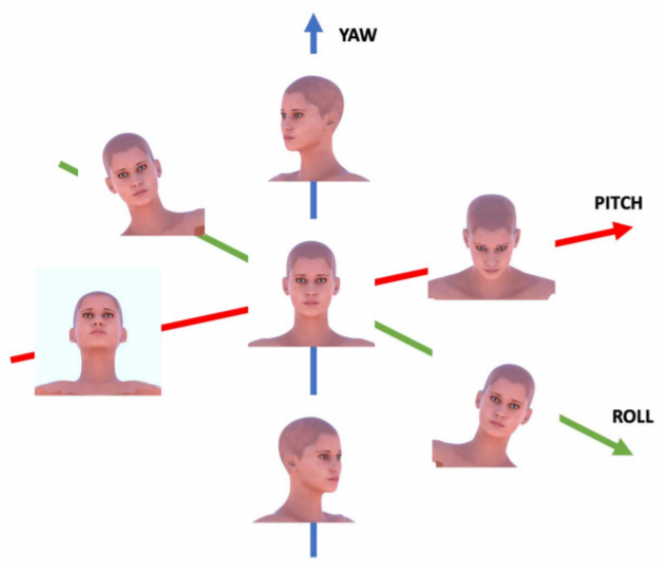

Figure 2.10: The synthetic model rotations, an example.

Once we have the model and an input image to be evaluated, the method used to obtain the comparisons is the Hamming distance. To compare two strings of the same length, e.g. n, the Hamming distance is defined as:

$$d(s,t) = \sum_{i=1}^{n} \delta(s_i, t_i) \tag{2.9}$$

where s and t are the strings to compare having length n and $\delta(s_i, t_i)$ is the following function

$$\delta(s_i, t_i) = \begin{cases} 1, & \text{if } s_i \neq t_i \\ 0 & \text{if } s_i = t_i \end{cases} \tag{2.10}$$

This metric is particularly fast because the length of the strings is the number of sectors, that, as subsequently demonstrated, does not require to be high to reach a good precision in HPE.

The only constraint of this method is that, to properly use the Hamming distance, the spider-web must have the same dimension both for the model than the input images. It is also clear that, varying the configurations of the spiderweb, by increasing or decreasing the number of sectors, we can study the different behaviour of the errors. From the previous two claims, we understand that it is necessary to choose in the initial phase of the algorithm the number of circles and slices, defining the sectors, that could give us the best results. On the other hand, since the spider-web model is invariant with respect to the initial image dimensions, for how it was built, once the number of circles and slices is defined, all the features arrays will be comparable.

A preliminary experiment conducted on the number of circles and slices, with each circle equidistant from the others, highlights that the best configuration is composed by 4 circles and 4 slices. For this reason, an additional set of experiments were conducted with the configurations in Table 2.2, represented by the same number of sectors obtained varying the distance between the circles.

From those experiments, the best configurations result to be $4C\_4S\_var4$, that obtained the lowest mean error computed as:

$$MAE = \frac{1}{n} \sum_{j=1}^{n} |y_j - \hat{y_j}| \tag{2.11}$$

Table 2.2: Model configurations with the optimal number of sectors. R represents the ratio.

Config	**Cir**	**Sl**	**Ratio of each circle**	**Sec**
4C_4S_var1	4	4	1/15*R; 3/15*R; 7/15*R; R	64
4C_4S_var2	4	4	8/15*R; 12/15*R; 14/15*R; R	64
4C_4S_var3	4	4	1/10*R; 3/10*R; 6/10*R; R	64
4C_4S_var4	4	4	4/10*R; 7/10*R; 9/10*R; R	64

where y_j is the ground truth, in our case the true angular value and $\hat{y_j}$ is the prediction, in our case the predicted angular value. We calculated the MAE for Pitch, Yaw and Roll separately and also an overall MAE of the errors along the three axes.

The datasets used to perform the test are BIWI, Ponting'04 and AFLW introduced in Section 2.2. The resulting errors obtained are shown in Table 2.3

Table 2.3: Results of the WSM model per Datasets. Pointing'04 does not contain 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

The performance of the method were also tested in competitive conditions. The GOTCHA Video Dataset was used to test the ability of WSM to estimate the pose in a surveillance video environment. To test the ability of the method in low resolution contexts, three different kinds of perturbation were added to AFLW, obtaining AFLW_blur with a Gaussian filter with a standard deviation of 7, AFLW_motion with a filter simulating a 9-pixel horizontal moving, and AFLW_noise with a Gaussian filter of mean 0 and variance 0.15. The tests were conducted both with 4 circles, and 5 circles, revealing, as can be seen in Table 2.4, that in some cases the 5 circles configuration performs better.

Table 2.4: The results of WSM on GOTCHA and AFLW perturbations.

Conf	Dataset	E_yaw	E_pitch	E_roll	MAE
4S	GOTCHA	11.8	9	7	9.26
4S	AFLW_blur	9.95	8.33	3.81	7.06
4S	AFLW_motion	9.78	7.17	3.91	6.95
4S	AFLW_noise	18.75	5.62	6.87	10.41
5S	GOTCHA	12.17	10.65	6.52	9.78
5S	AFLW_blur	7.14	6.42	4.76	6.11
5S	AFLW_motion	8.26	8.48	3.91	6.88
5S	AFLW_noise	18.75	5.62	5	9.79

The comparisons of WSM with the state of the art, is presented at the end of this Section, together with the results of the other two methods that we will introduce.

Finally, also the mean time required to perform the operations is considered, and it is slightly dependent of the number of sectors, as can be seen in Figure 2.11. The overall algorithm is, anyway, quite fast, with 0.36 seconds necessary to perform all the operations in the best cases, of which only 0.108 seconds are necessary to the operation of the WSM algorithm.

2.3.2.2 QT PYR: The Quad-Tree approach

We firstly proposed the Quad-Tree based approach to estimate HP in [59], demonstrating the encouraging preliminary results. The most performing version of the method we want to present is in [60], where we obtained results comparable to the state of the art, due to the improvements we are going to talk about below.

As in WSM, also here the preliminary step is composed by the face detection and the landmark prediction. The difference is represented by the core of the method. Using the Quad-Tree decomposition, the image containing the landmarks is subsequently decomposed based on the amount of information in it, in our case the presence of the landmark. The initial image is the root of the

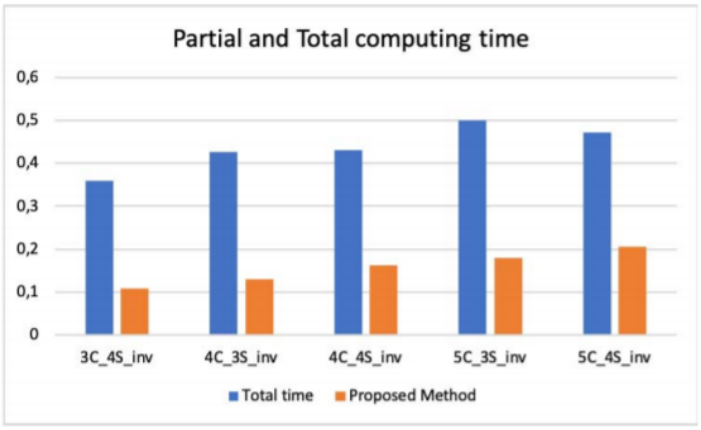

Figure 2.11: The overall and relative time required to the WSM framework.

tree, since it contain 68 landmarks, it is split into 4 sub-images. Then, each sub-image is split into 4 sub-sub-images, according to the presence of the landmarks. This subdivision will continue until at least one of the quadrants reached a pixel size of 4x4.

To obtain a pose array starting from the Quad-Tree decomposition, at each iteration, for each cut, four 1 will be inserted in the array, for each non-cut, four 0 will be inserted in the array. This process can be better explained using the workflow in Figure 2.12.

Following the steps of this decomposition, it results clear that chose to add elements in the array both in the presence of landmarks and in the case of their absence, is necessary to guarantee the fixed length of the array, regardless of the tree decomposition. The decomposition information, in fact, is not observable in the length of the decomposition, but in the distribution of 0 and 1 in the tree array. This distribution represents the imbalance of the tree that is the subject of study in this method.

Those arrays were created for each head pose in the model. Also in this case, the synthetic model solves this task and the resulting tree arrays will be 2223 as in WSM. To test an input image, the latter is decomposed by the quad-tree decomposition

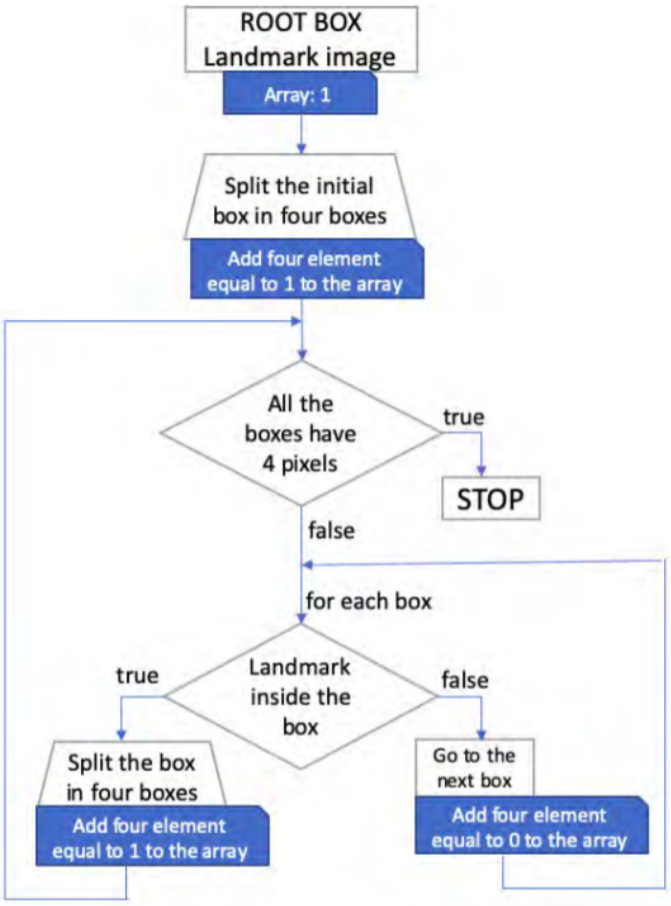

Figure 2.12: The quad tree generation workflow.

and the tree array is obtained. Then, this array is compared to the ones in the model using the Hamming distance, introduced for the previous method.

Considering the initial dimension of the face extracted, that became fixed in the preprocessing steps, and the 4x4 pixels lower limit, we will obtain tree arrays of dimension 1365. If we want to compare this dimension to the one obtained in WSM, this one is about 21 times bigger. To solve the related computational problem that affects the efficiency of the method, we exploit the fact

that, differently from the WSM features array, the quad-tree array is binary. In addition, the values in the arrays have a decreasing weight. This property is directly inherited from the method, since two landmarks distributions are as much different as sooner the difference in the decomposition in observable. The array is ordered following an horizontal order of decomposition of the tree, and from this, in conclusion, the concept of correspondence between ordered arrays and similar poses is evidenced. Once the 2223 arrays are ordered, the way to perform the comparisons can follow the style of a dictionary search. Instead of performing 2223 comparisons for each images to be tested, the comparisons will be at most 8 ($log_2 2223$). This allows to have a near-real time algorithm, that perform the operation of the core, QT comparisons, in 0.044 seconds.

The overall framework of this method can be found in Figure 2.13

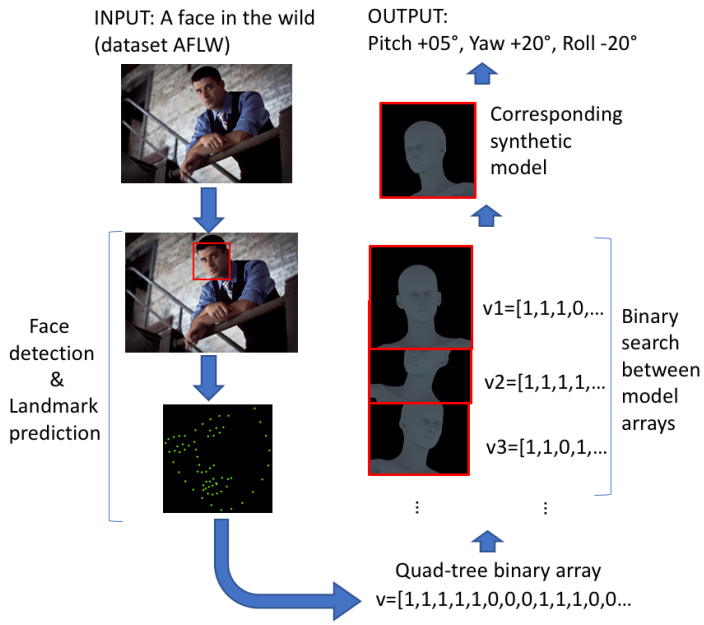

Figure 2.13: The QT PYR workflow.

The results of this method are presented for BIWI and AFLW.

In addition, another set of experiments were performed on the same datasets, called QT PY+R. In this case the image is previously normalized in roll and then, without the roll component, pitch and yaw are computed using a model of 207 images. The normalization of the roll is made starting by measuring the angular coefficient of the straight line passing through the two points represented by the external corners of the eyes. As can be seen in Table 2.5, the configurations with a more homogeneous behaviour over the three axes is QT_PYR. We can also observe that evaluating the roll separately improves the error along this axis. Even if more expensive in terms of computational time, an ideal solution will be to detect the roll and proceed with PYR, ignoring the roll result of the PYR method.

Table 2.5: Results of QT method 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

As in WSM, also the comparisons with the state of the art of QT_PYR will be discussed at the end of this Section.

2.3.2.3 HP^{2}IFS: The Fractal Encoding approach

Differently from the previous methods, based on the landmark distribution, HP^{2}IFS considers the image appearance by the pixels values. The method, presented in [61], uses several mathematics concepts related to the fractal decomposition of an image. Fractal encoding is usually applied in image compression. Here, we will use the first step of this algorithm to obtain an encoding of the image, exploring the concept that a part of an image can be approximated by a transformed and down-sampled version of another part of the same image, using a property called self-similarity.

Given a metric space M, with metric d, a contraction mapping on (M, d) is a function f from M to itself, with the property:

$$d(f(x), f(y)) \leq kd(x, y) \tag{2.12}$$

for all x and y in M where k is a non negative real number between 0 included and 1 excluded. If the concept of contractive function is merged with the Fixed Point Theorem defined as follow:

Fixed Point Theorem In a complete metric space (M, d) if $f : M \to M$ is a contractive transformation with parameter k, then exist and it is unique, a fixed point $x_i \in M$ such that

$$f(x_i) = x_i \tag{2.13}$$

and for any point x in M is also true

$$\lim_{n\to\infty} f^n(x) = \lim_{n\to\infty} \underbrace{f(f(f(...(x))))}_{n\ times} = x_i \tag{2.14}$$

we can claim that, if our fixed point is the image itself, our aim is to find a set of contractive function that works as affine transformations over this image. A set of such transformations is called a Iterated Function System (IFS), defined as follow:

$$F(X) = \bigcup_{i=1}^{N} f_i(X) = X \tag{2.15}$$

where F is a set of contractive affine transformations $f_1, ..., f_N$, which is itself a contractive transformation, and x is the fixed point.

This theoretical background was initially used by Arnaud Jacquin to define Partitioned Iterated Function System (PIFS), to perform image compression as follow:

- Let M the metric space, $D_i \in M$ a collection of sub-domains and f_i a collection of contractive maps.
- The image to be encoded is partitioned in R_i, non overlapping range blocks.

- The image is then partitioned in larger non overlapping blocks D_j called domain blocks.
- For every range block R_i, a domain block D_{R_i} is found such that a contractive affine trasformation f, transforms this Domain block to a good approximation of the range block.

The such obtained sequences of Domain Block D_{R_i} represents the encoding of our method.

After the face detection and landmark prediction, with the same above mentioned methods, a Facial Mask of the initial image is created starting from the boundary landmarks detected. Then, the image is resized to 256x256 pixels and the encoding is applied with a Domain of 8x8 pixels and a range of 4x4 pixels. The resulting codec matrix is of 256 rows and 6 columns where each row represents a block. In the first two columns there are the coordinates of the block, in the third the affine value of the inversion, in the fourth the affine value of the rotation and in the last two brightness and contrast respectively. The matrix is then converted in a 1536 entries array to perform the comparisons. The latter are performed using the Hamming distance previously introduced.

The overall workflow of the method is depicted in Figure 2.14.

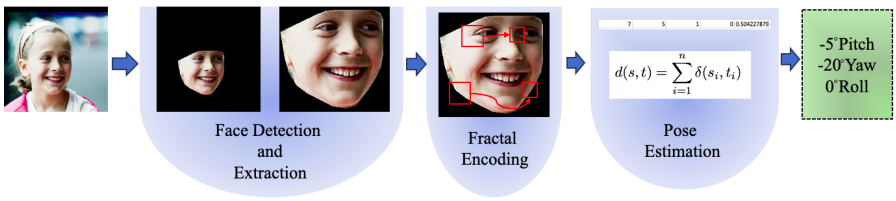

Figure 2.14: The HP^{2}IFS workflow.

Differently from the previous two methods in which it was possible to use a synthetic model as reference, because only the landmarks were used, here has been necessary to use real images to build the model. In particular, each dataset used to test the method was split in 80% of elements to build the model and the remaining to perform tests.

The results were presented for BIWI and AFLW2000, and can be seen in Table 2.6.

Table 2.6: Errors of HP2IFS on BIWI and AFLW2000.

Dataset	**E_yaw**	**E_pitch**	**E_roll**	**MAE**
BIWI	4.05	6.23	3.30	4.52
AFLW2000	6.28	7.46	5.53	6.42

In addition, has been studied the behaviour of the error, respect to the angular pose, visible in Figure 2.15 and 2.16. As can be seen, the error grows when the rotation angle becomes bigger. This is appreciable for all the three axes, however this behaviour is particularly strong for pitch in the BIWI dataset. If we consider the percentage of images falling in the error classes, for BIWI we have that about 35% of images have no error in Pitch, 77% of images have an error equal or less than 5°, 95% have an error equal or less than 10 ° and almost all images has an error less than 15 °. In terms of Yaw, about 55% of images have no error, 90% of images have an error equal or less than 5°, 97% have an error equal or less than 10° and almost all images has an error less than 15°. In Roll, 71% of images have no error, 97% of images have an error equal or less than 5°, and almost all images have an error equal or less than 10°. On the other hand, on AFLW2000 for both Pitch, Yaw and Roll, about 28% of images have no error. Then, in Pitch, 60% of images have an error equal or less than 5°, 81% have an error equal or less than 10°, 92% have an error equal or less than 15°, 96% have an error equal or less than 20° and almost all images has an error less than 25°. In Yaw, 66% of images have an error equal or less than 5°, 87% have an error equal or less than 10°, 95% have an error equal or less than 15°, 97% have an error equal or less than 20° and almost all images has an error less than 25°. In Roll, 72% of images have an error equal or less than 5°, 92% have an error equal or less than 10°, 98% have an error equal or less than 15°, and almost all images have an error less than 20°.

In the following Section, the results of the three proposed method are compared to the state of the art.

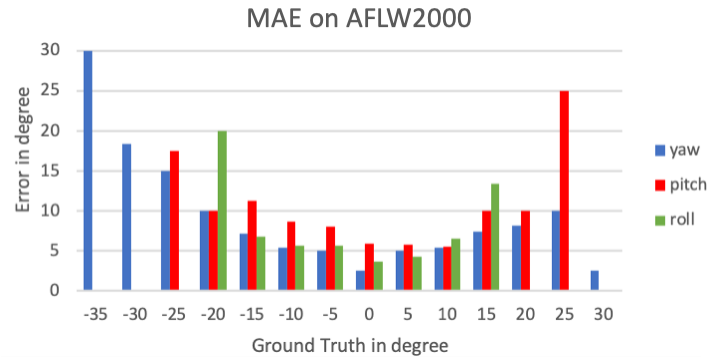

Figure 2.15: The error behaviour in HP^{2}IFS on AFLW2000.

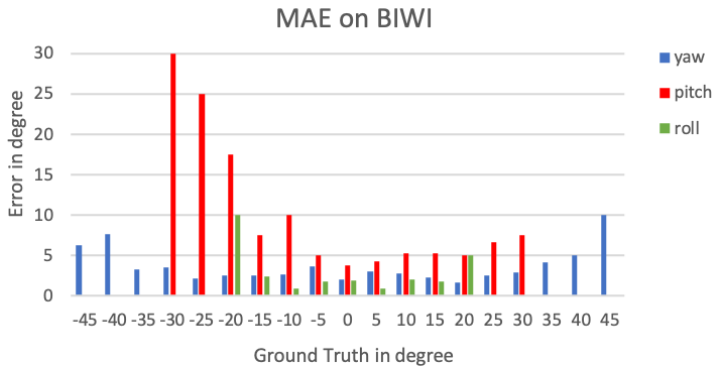

Figure 2.16: The error behaviour in HP^{2}IFS on BIWI.

2.3.2.4 Comparisons with the state of the art

The recent techniques to estimate the Head Pose are very performing. In particular, due to the advent of the Deep Neural Networks, also the HPE benefit from it in terms of errors. Before introducing our results, we have to highlight that no one of the method we developed and presented in the previous three sections uses Machine Learning approaches. This choice was made to differentiate the proposed model from the most classical and well known CNN and RNN architectures and, in particular, to offer methods that are dataset-independent. This is possible because

our proposed method, not having a training step, does not need any further training to be adapted to different environment or new test images.

The result of our methods are in the last three rows of Table 2.7. As can be seen, the better is WSM for pitch and HP^{2}IFS for yaw, roll and MAE. However the recent methods using DNNs, QuatNet and FSA-Net, outperform our methods in yaw and roll.

Table 2.7: Mean Absolute Error of Pitch, Yaw, and Roll Angles Across Different Methods over the Biwi Dataset

Method	Yaw	Pitch	Roll	MAE
Multi-Loss ResNet50 [48]	5.17	6.97	3.39	5.177
GPR [62]	7.72	9.64	6.01	7.79
PLS [63]	7.35	7.87	6.11	7.11
SVR [64]	6.98	7.77	5.14	6.63
hGLLiM [65]	6.06	7.65	5.62	6.44
FSA-Net [66]	4.27	4.96	**2.76**	**3.99**
Coarse-to-Fine [67]	4.76	5.48	4.29	4.84
QuatNet [68]	**4.01**	5.49	2.93	4.14
WSM [58]	6.21	**3.95**	4.16	4.77
QT PYR [60]	5.41	12.80	6.33	8.18
HP^{2}IFS	4.05	6.23	3.30	4.52

In Table 2.8, there are the results of the recent literature on the AFLW/AFLW2000 dataset. As can be seen, in this case, other than the best of the three methods proposed, WSM outperform also the state of the art along all the axis, beating also the DNN based methods. Since AFLW is a dataset more competitive than BIWI, because of the heterogeneous configurations in it, this result well demonstrating the ability of the method to generalize and its applicability to uncontrolled scenario.

Since Pointing'04 is a less reliable dataset, we tested on it only our best method, WSM, comparing its results to the state of the art, in Table 2.9. In this case we can observe that all the methods are not able to achieve considerable results. In this sense we have to underline that this dataset is collected using 15° as a step,

Table 2.8: Mean Absolute Error of Pitch, Yaw, and Roll Angles Across Different Methods over the AFLW2000 Dataset

Method	Yaw	Pitch	Roll	MAE
Multi-Loss ResNet50 [48]	6.470	6.559	5.436	6.155
Hyperface [54]	7.61	6.13	3.92	5.89
KEPLER [50]	6.45	5.85	8.75	7.01
3DDFA [69]	5.400	8.530	8.250	7.393
FAN [70]	6.358	12.277	8.714	9.116
QuatNet [68]	3.973	5.615	3.92	4.503
QT PYR [59]	7.6	7.6	7.17	7.45
WSM [58]	**3.11**	**4.82**	**2.25**	**3.39**
HP^{2}IFS	6.28	7.46	5.53	6.42

instead of 5°, it means that all the methods are under the level of sensitivity of the dataset and this makes them acceptable. In particular WSM performs better on pitch and hGLLim on yaw.

Table 2.9: Mean Absolute Error of Pitch and Yaw Angles Across Different Methods over the Pointing'04 Dataset

Method	Yaw	Pitch	MAE
Stiefelhagen [71]	9.7	9.5	9.6
Gourier et al. [72]	12.1	7.3	9.7
SVR [64]	12.82	11.25	12.035
hGLLiM [65]	**7.93**	8.47	**8.2**
Probabilistic HDR [73]	8.70	8.85	8.775
Kong et al. [74]	10.98	9.71	10.345
WSM [58]	10.63	**6.34**	8.4

From this analysis emerges that the only methods outperforming for some axis our algorithm are DNN based. Starting from this, we can conduct a comparison that evaluates the computational difference between training and training-free approach. Let us consider two specific methods, QuatNet [68] and our WSM [58]. We choose QuatNet because it achieves similar performance to WSM, revealing to be in more than one case the antagonist

of WSM, and, it is one of the few DNN methods that declare the computational time required to build the model and the devices on which the experiments were performed. Both the methods operated in 30 fps, in real-time, in testing an image. In Table 2.10, we can observe that, even if the MAE results very similar, the computational resources and the required time to obtain the model-perform the training, is very different among the two method. We can claim that WSM seems to be more suitable to a low resources device, and due to this characteristic may be applicable to mobile devices in the future.

Table 2.10: QuatNet (training) vs WSM (training-free)

Method	**GPU**	**Time**	**E_Pitch**	**E_Yaw**	**E_Roll**	**MAE**
QuatNet [49]	NVIDIA GTX 1080	4.5h	4.32	3.93	2.59	3.61
WSM [58]	Intel HD Graphics 515	0.16h	4.82	3.11	2.25	3.39

2.4 Conclusions

Starting from the kind of datasets and applications of the Head Pose Estimation, we examined in this Chapter our contribution to the state of the art. We focused on RGB images that can be captured by any kind of camera available and represent the most competitive fields in terms of errors. However, in the near future we will migrate our methods to depth images, because of the growing attention in the driver applications of HPE. In fact, other than the best frame selection and the face recognition, HPE has demonstrate its value also to evaluate the attention state of an individual only in the few years. This confirms the wide horizontal expansion that the HPE methods can reach in the future, justifying the growing number of methods recently born in this field. The high performances obtained by our methods are encouraging in their migration to depth images. In this sense we will also plan

to solve the landmark detection problem related to depth images, on which the first two method presented are based.

Chapter 3

Gait analysis as a soft biometric

In this Chapter we outline a human trait that can be used in different ways, in particular as a soft biometric: the gait. The study of the gait is, more generally, the analysis of the animal locomotion, an it is related to the muscles activities. The study of this distinctive biometric trait has its roots in *De Motu Animalium* [75] by Aristotle. The analysis of the human movements in particular were firstly introduced by Fischer and Braune in 1980 [76]. But, only after the advent of the photography and cinematography this trait has gained particular interest. The application field of the Gait Analysis were predominantly three in the last decades: medical analysis; biometric analysis; movements and sport analysis. In the following sections we will introduce the principal components of this trait and their differences among the applications. In particular we focus on biometric analysis and, even more in particular as a soft trait of human recognition.

3.1 Principal components of a human gait

In normal conditions, when no pathologies occurs, the movements related to walking operate at an unconscious level. On the other hand, when pathologies occurs, they can be detected by the modifications or the anomalies of the gait. It is clear that, before introducing any measure related to the gait, we have to contextualize the operation fields:

- Medical diagnostics: the anomalies on the gait movements can be at the same time a symptom or a cause of a medical disease. For this reason the gait analysis is mainly involved in the treatment of patients with Cerebral Palsy [77], or stroke patients [78].

- Sports: even if related to medical diagnostic, the analysis of the gait related to sports is less focused on serious pathology and more focused on the way to move of the athlete, for example the runners. An example of applications of this kind of analysis can be the choose of the proper shoes for a runner. In this case the study mainly involves the pronation, the way in which the humans foots roll inwards as they strikes the floor. This represent the way in which the body distributes the impact as a part of the gait cycle [79].

- Biometrics: the human gait demonstrates its value in distinguishing a subject from another. The combination of weight, limb length, footwear, posture and motion, leads to a unique feature object of an individual, as a print. As the identity recognition from gait improved, other soft biometric traits like cooperativeness, action recognition, and gender, demonstrate to be discriminated from the gait. This particular field is of interest in this research and will be properly detailed.

The gait can be divided in two main phases: stance and swing. When the foot touches the ground for the first time the stance

phase begins and ends when the same foot leaves the ground. When the foot leaves the ground the swing phase begin, and ends when the same foot touches the ground again. The stance and the swing phase cover approximately 60% and 40% of the cycle, respectively. An example of those phases can be seen in Figure 3.1.

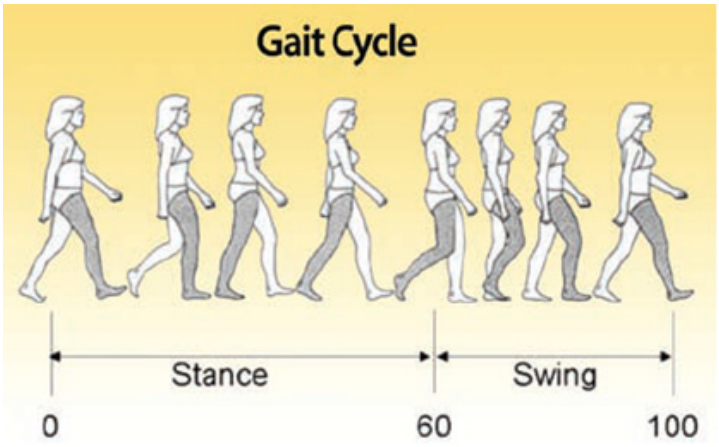

Figure 3.1: The gait phases.

The parameters related to the human gait can be divided in Spatial and Temporal.

Spatial parameters consider the spatial variation during the movements. Defined as the Line of Progression (LP) of the body, the subject direction during the data collection, the spatial parameters are:

- Step length: the length parallel to the LP, from the posterior contact of a foot to the posterior contact of the other foot.
- Stride length: the distance between two consecutive footprint of the same foot evaluated from the posterior heel.
- Step width: the distance between the left and the right foot from the LP.
- Step angle: the angle between the LP and the foot axes.

The temporal parameters are evaluated in time and include:

- Cadence: number of steps per unit time. Usually is evaluated as steps/minute.
- Speed: the distance covered per unit time. It represent the velocity.
- Motion: the body part movements during the gait, comprehensive of the analysis of tremors.
- Force and Pressure: the variation of those two characteristics during the gait.

As it can be deduced, not all of the mentioned parameters are used in the biometric applications of gait. For example, Motion is more helpful in Medical purposes. Force and Pressure are more used in sports application. Velocity is more used in medical and sport application and less used in Biometric applications. All the other parameters introduced are more or less used in all the applications. A more complete list of parameters can be found in Figure 3.2, where we underline the ones useful for Biometrics purposes.

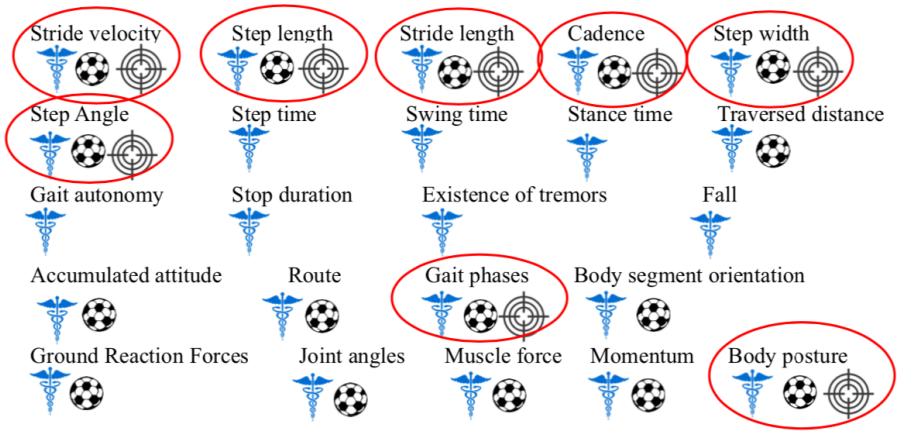

Figure 3.2: The gait parameters, useful for medical purposes, sport or biometrics.

3.1.0.1 Sensors

Depending on the parameters used to collect, there are several sensors involved. The first sensors born to capture gait parameters were floor sensors [80]. Floor sensors, as the name suggests, are floor platforms on which the subject should walk. They are also called "force platform" because they evaluated the gait by measuring the force and the pressure of the subjects that walk on them. From their ability to measure the force, we can split those kind of sensors in force platform and pressure measurement systems. The Force Platforms can measure the force vector applied in the foot. On the other hand, the pressure systems can quantify the pressure patterns of each foot but not the applied force in its horizontal component. Those kind of devices are often used in medical studies to observe the general problem of patients during the gait.

The wearable sensors are, on the contrary of the floor sensors, placed on the patient's body. In this case, the structure of the sensors and their aspect is strongly dependent on the kind of parameters to be measured. If we consider the sensors that measure the force, they need to be placed under the foot to return a current or voltage proportional to the pressure applied. Those instruments are similar to led sensor and are placed under particular shoes. There are also the inertial sensors that measure the velocity, the acceleration, the gravitational forces and the orientation, using accelerometers and gyroscopes. An insole is often used to equip those sensors. If we want to consider the contraction of the muscle during the gait, those kind of sensors do not fit the purpose. In this case, even if result less comfortable for the subject, it is preferable to use electrodes.

In Figure 3.3 we can see the Floor and wearable sensors we discussed.

All of the sensors previous introduced require the cooperation of the user. It is clear that in biometric methods, applied in security, surveillance and identity recognition in general, this constraint does not results efficient.

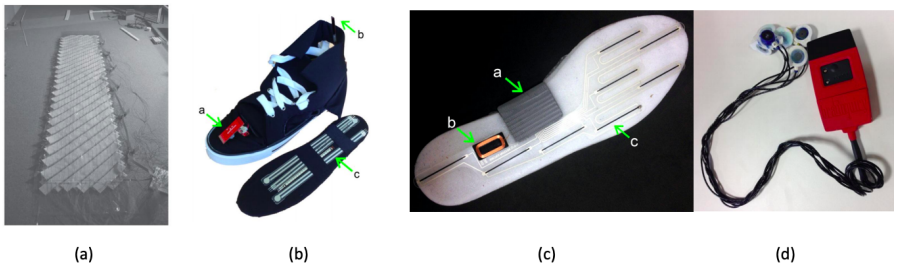

Figure 3.3: The gait sensors.(a) force floor, (b) force shoes, (c) Insole sensors, (d) electrodes.

For this reason, the preferable way to capture the gait for biometric purposes is using cameras. Depending on the kind of camera ivolved we can have RGB videos, Depth Videos or Infrared Videos. Depending on the visual of the subject, the acquisition can be sagittal or lateral and frontal or back [81]. In surveillance purposes, one single camera is the ideal to collect and test the gait because of the real context applications. However, use a single camera furnishes only a view of the subject. To compensate this characteristic, more than one camera can be used [82], or in alternative, a depth camera. The use of a single depth camera, as a Kinect, can be useful to extract relevant features from the gait [83] that can lead to the identity recognition. The depth information provided by the Kinect, can be also integrated by the RGB component to also evaluate pose estimation and abnormal gait [84]. The use of Kinect, with RGB information, is at the moment the most used input data to evaluate identification algorithms operating in surveillance scenario [85] [86].

3.2 The state-of-the-art on gait recognition

In the previous Section, we examined the traits and the instruments related to the gait analysis. Here, we want to provide a general overview of the recent methods, in the last five years, that

contribute to the state-of-the-art in this field. If we refer to wearable sensors for gait recognition, recent work can be split into two main categories depending on the kind of method involved [87]. In the first kind of methods we can find the use of signal matching algorithms. The signals were preliminary segmented by using the gait cycle or step phases above described and then different distances can be evaluated to obtain the recognition rates. An example of recent techniques of step segmentation can be found in [88] were the approach to the gait is based on the accelerometer. Techniques based on cycle segmentation, on the other hand, can be based on the extreme values of the cycle [89] or on the cycle length estimation [90]. After the segmentation step, the methods can perform recognition by comparing Dynamic Time Warping (DTW) [91], or by examinit both DTW than the frequency domain [92] [89]. However, as can be noticed, more recent techniques to perform recognition from gait using wearable sensors involve machine learning techniques. Those technique uses Support Vector Machine (SVM) as in [93], K-Nearest Neighbor (KNN) as in [94] and [95]. But the most popular methods involves Neural Networks and in particular Convolutional NN (CNN) [96] and [97].

As can be seen, the recent techniques mainly involve the use of smartphones and the trait collected for the gait recognition is furnished by the accelerometer. Despite this, non wearable sensors, like cameras, are still preferred in some scenario in which the user cooperation can not be required. In surveillance and security purposes, as already discussed when we introduced the sensors, videos are the most popular source of information to extract the preferred traits. For this reason it is vitally important to know how to extract gait information from videos. Differently from other biometric traits, the features of gait are strongly different depending on the source. For camera data, we can extract two kind of features: Binary Silhouettes or Human Poses [98]. In Figure 3.4 we show an example of those two features we are going to examine.

The binary silhouettes can be extracted using segmentation algorithms and then used to built the Gait Energy Image (GEI). GEI

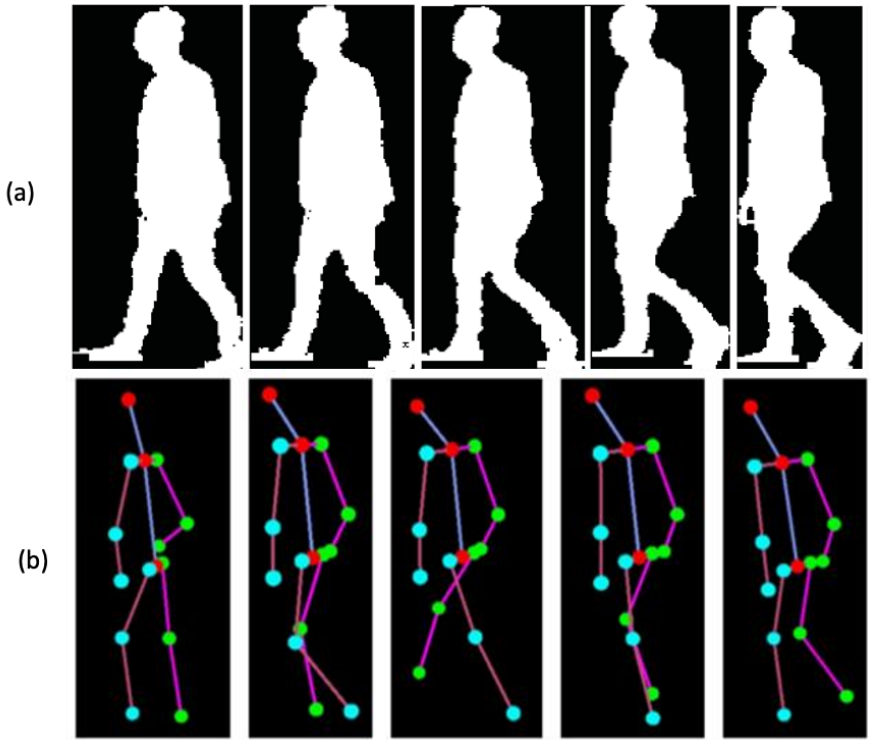

Figure 3.4: An examples of vision-based gait features. (a) the silhouettes and (b) the skeleton, extracted using OpenPose [99].

is the mean of a gait cycle and represents the frequency of a moving person being in a particular position. If, instead to compute the mean silhouette, we compute the difference between consecutive frame images of silhouettes, we will obtain the FDEI. The GEI is, however, the most used techniques in silhouette-based methods. Recently, GEI began the input of NNs, both small as [100] than more complex, involving Deep Neural Networks [101]. As mentioned, another features that can be extracted refer to the Human Pose. The human pose is extracted searching specific points of the human skeleton. For this reason those kinds of algorithms are classified as skeleton-based and can proceed without the use of training techniques, as in [102] focused on level-matching, or in-

volving different machine learning techniques as CNN [103], RNN [104], or a combination of the latter with Long short-term memory techniques, as in [105] combining CNN and LSTM, or as in [106], combining RNN and LSTM. Very new research is also particularly focused on the generalization of those techniques, deriving from the creation of multi-view large datasets for gait [107]. As it can be seen in the Vision-based gait recognition survey in [108], there are many datasets used in literature to solve this task using camera information. In Table 3.1 we reported only the most recent (since 2014). However it is clear, from the dataset creation year, that the attention on gait has been moved in different directions in last years. We will see how the gait is more analysed, instead of recognition tasks, in recent years, in the following section.

Table 3.1: Vision-Based Gait Datasets.

Dataset	**Year**	**Subjects**	**Resource**
KY4D Database-B	2014	42	16 cam
KY4D Shadow Database	2014	54	2 infrared cam, 1 cam
OUISIR Speed Transition	2014	179	1 cam
Human Motion	2014	15	10 osprey cam, 47 markers
KIST	2013	113	8 cam, 15 markers
OUSIR Large Populatin	2013	4007	2 cam
AVA-Multiview	2013	20	6 cam

3.3 Soft biometrics from gait

Gait, as other biometric traits, is born to solve recognition tasks. However, differently from other traits, it is less strong to identify peoples. The cause can be researched in all the components that impact the human gait. First of all, medical issues make changes to the way of walking or running. Then, also the emotion of

the individual changes its way to walk, in fact in Figure 3.4 (b) we introduced in the previous section, the subject is the same but different emotions are simulated during the walking phase. Also the action performed changes the ability of a method, mainly trained on walking sequences, to recognize a user. All of those factors that impact gait recognition, lead during the past years to a simple question: "can gait be used as a soft biometric trait instead of a recognition biometric?" Searching in literature, the amount of papers born to detect soft traits from gait certainly leads to a positive answer. We can distinguish two mainly traits to search studying the human gait, physical or behavioural. Physical traits, in this case, could help to detect some characteristic of an individual but are not sufficient to confirm his identity. Physical soft traits can be:

- the gender: can be useful in the medical gait analysis where some issues show differences among genders [109], to know who is using a smartphone [110], or more in general to be applicable as a preliminary step of recognition for security purposes [111], [112], [113]
- the age: classify the age of an individual from the way they walk finds its application especially in surveillance [114] [115] . In this context, the gait is also used in combination with other biometric traits [116], or alone to estimate both gender and age [117].
- cloth classification: even if not directly related to gait movements, it is possible from the gait silhouettes, to detect the cloth of a subject. This is used, in general, to help an identification step [118].

On the other hand, there are behavioural traits that are recently detected by the gait:

- action recognition: is one of the branch as in which the gait is involved the most [119] [120] [121]. It is also used to predict the progress of the action in a video [122] or for images from drones in surveillance scenarios [123] [124].

- emotion classification: emotions impact the way to walk, for this reason a literature is born to classify the emotions of a subject from his gait [125] [126]. This kind of analysis is also used for medical purposes [127].

- subject behaviour classification: this very new field, differently from detecting a specific action or emotion, is focused on detecting the intention of a subject, often classified in malicious or not. Examples of these techniques can be found in [128], or in [129] and it will be the focus of Section 3.3.3.

From this analysis, it is clear that the datasets previously introduced, labeled only for the user identity, are not suitable to be applied in these new contexts. For this reason we present in Table 3.2 other datasets in literature useful in these fields.

Table 3.2: Soft biometric from gait datasets.

Dataset	**Label**	**Classes**	**Samples**
Kinects [130]	Action	400	306245
NTU RGB+Dv [131]	Action	60	56880
UCF101 [132]	Action	101	13320
CMU [133]	Action Behaviour Gender	23	2605
EWalk [134]	Emotion	4	1336
OULP-Age dataset [135]	Age	88	63846
UFS [114]	Age	40	1870
ClothingAttributes [136]	Cloths	42	1856
GOTCHA-I [31]	Behaviour Gender Context	2 2 5	682

In particular we want to highlight the GOTCHA-I Dataset that has been built during these 3 years of research.

3.3.1 GOTCHA-I Dataset

The GOTCHA-I Dataset [31] is one of a kind. This because it is multi-view and multi-labeled. The 63 subjects in this dataset have been split in different folders, and this makes possible to perform identity recognition from gait. In addition, they are labeled with the gender, that also can be used to perform gender recognition from gait. The main innovation of this dataset is the presence of behaviour label. In particular, GOTCHA-I has data split into cooperative or non-cooperative mode. Differently from a classical non-cooperative mode, in which the subjects move freely, in this non-cooperative mode the subject tries to avoid the camera. This makes this dataset particularly useful in surveillance context were malicious subjects can be identified because they try to avoid the camera. Another thing that differentiates this dataset from the others in literature is the camera position. In fact the camera is not fixed, but moving, to simulate the wearable cameras in use by the police officers in several countries. Combining the cooperative and anti-cooperative mode with the environment in which the videos are captured we obtain the following classes:

- (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;

In Figure 3.5 we can appreciate the differences between the environments involved and the real scenario represented.

Figure 3.5: Different environments of Gotcha dataset. In the top: the stairs and path follow outdoor. In the bottom the walk modalities: outdoor, indoor, indoor with flashlight. An example of non-cooperative mode is in the bottom-middle image, representing the indoor environment.

In GOTCHA-I there are 62 subjects in 11 modalities, for a total of 682 videos. Experiments on GOTCHA-I are presented in the following Sections.

3.3.2 Gender from Gait

As we introduced, gender is one of the traits that can be deduced by the human gait. In this Section we will explore two methods we designed in the last three years in which the gender is detected

from the skeleton information of the gait of the subjects. In the preliminary work in [112], we proposed a geometric technique. The major step of this method are: data extraction; features creation and classifier selection. First of all we detected the human skeleton using OpenPose. OpenPose generates a set of keypoints in the body of the subject that can be further analysed. To extract relevant features from these keypoints we compute their distances. From this point we conducted two kinds of experiments:

- Total body keypoints: In a first attempt we tried to use all the keypoints and, as a consequence, all the available distances. Starting from the GOTCHA-I Dataset Videos, we extracted the distances that represent our features. The result is 5870 array of features, each of them representing a frame. In particular 3970 in cooperative mode and 1830 in non-cooperative mode. The limited amount of frames compared to the total number of videos, is a problem related to the video itself. In fact, not all the frames contain all of the keypoints, because not all the body of the subject is visible in each frame. Then, those features array were split in 70% for train and 30% for test. The gender has been classified using random forests of different depths. The results are shown in Table 3.3.

- Upper body keypoints: In a second attempt we used only the keypoints, and the relative distances, of the upper body. This choice was made because of two reasons. First of all the total body keypoints are not visible in all frames, and this leads to lose a lot of considerable frames. The choice to prefer the upper body keypoints is also justified by the fact that, by anthropometric differences between men and women assumptions, women generally have hips that are wider than the shoulders while is generally the opposite for males. Using only upper body keypoints we have 17000 total arrays of features. Splitting the dataset in the same way of the previous experiment, and building the Random Forest (RF) classifier, we obtain the results in Table 3.3. In particular

we choose the best RF depth results.

Table 3.3: Result of the first method based on Random Forests.

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%

As it can be seen from the table, the possibility to use all the frames and to decrease the computational time to compute the distances, are paid with a lower accuracy using only some upper body feature.

From these preliminary results we made further experiments on this field, published in [137]. In particular, here, we examined also other classifiers. We started from 18 keypoints of 200 frames for each videos. Then, from those 18 keypoints, all over the body, we computed 153 distances. For this reason, for each video we will have 30600 features. We used 30 subjects in total to train the classifiers, equally distributed in men and women. Since there are 62 subjects in the GOTCHA-I Dataset, but only 15 women, we choose the 15 men randomly. The classifiers involved are:

- Random Forest(RF): we already used this classifier in the previous work, however, here we sensitively increase the number of trees in the forest. In particular we choose 100 trees, entropy as the function to compute the quality of the split at each step, and bootstrap samples during the building tree phases.

- K-Nearest Neighbor (KNN): KNN algorithm is based on the similarity between samples. This kind of algorithm requires to fix a number, K, of values to take under account as the neighbour of a sample to classify the latter. The experiments were conducted used 5 neighbors and euclidean distance computed between points.

- Support Vector Classifier (SVC): SVC solves the classification problem by using support vector. The aim of this algorithm is to create an hyperplane that separates the two classes involved.

- AdaBoost: this particular method solves a classification problem converting the latter into a set of simpler problems.

All the methods were training using 70% of the data as training and validation, and 30% as testing. In particular we perform the experiments using cooperative and non-cooperative mode separately, to underline their differences. Results are shown in Table 3.4.

As it can be seen, the non-cooperative modality makes the predictions more difficult. In opposition, the better environment is the indoor-with-flash, in which all classifiers perform better. The best classifier is, in general, the random forest. This confirms our initial choice to operate with RF in the previous work. These are the first results obtained on the GOTCHA-I Dataset, and they demonstrate how much this dataset is competitive and realistic, even if we perform a binary classification.

3.3.3 Cooperativeness detection human-gait based

Motivated by the competitiveness of the GOTCHA-I Dataset, we have increased the complexity of the algorithms involved to solve an even more difficult problem: cooperativeness detection. Detect if a user is moving freely or they are trying to avoid the camera, is a matter strictly related to surveillance scenarios. We mainly discussed this topic in [129]. The path we follow to solve the cooperativeness detection problem involves attentive recurrent neural networks. The pipeline of the method can be seen in Figure 3.6.

The steps of the method can be summarized as follow:

- Skeleton representation: As in the gender classification method, we start from the skeleton keypoints. In this case, as in the first analysis of gender, we use only the keypoints of the upper body. The keypoints are detected by the human body

Table 3.4: 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%	

model by OpenPose. The total number of keypoints considered are three and each of this point has two spatial coordinates and a coordinate that represents the confidence in the detection. To obtain the data for cooperativeness purposes, here we have to take into account the temporal variable. In our case we will consider the variation in distance about the spatial coordinates during time. We define the distance

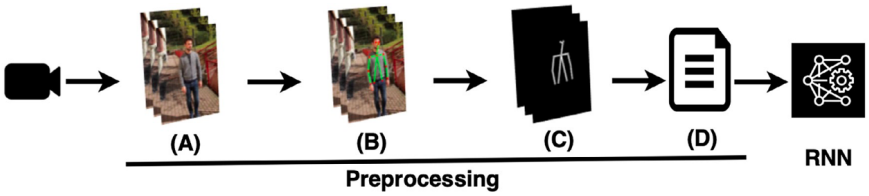

Figure 3.6: The pipeline of the cooperativeness detection method.

between two points in one frame

$$d_{p,q}(x,y) = ||v_p(x,y) - v_q(x,y)|| \tag{3.1}$$

where p and q represent two joints and, as a consequence, the associated confidence score, the third coordinates, becomes:

$$d_{p,q} = \frac{v_p(z) * v_q(z)}{(v_p(z) + v_q(z))/2} \tag{3.2}$$

where z is the original score. Using also the confidence scores during the following steps, we can detect the intention to avoid the camera using a part of the body, e.g. the hand.

- Bucketing algorithm: this preprocessing step is required to normalize the data provided to the network. This step is also very useful to handle input sequences of significant length with the aim to increase the training speed of the network. Summarizing this step, bucketing provides a tuple of the raw data and the differential data obtained from the differences between the successive frames. Those data are then normalized to feed the attentive neural network.

- Attentive recurrent network: This modeling start from a Recurrent Neural Network (RNN). The RNN characteristic is the presence of a self-connected hidden layer. An issue related to RNN is their problem to learn long-term contextual dependencies. Since we are working with videos, we need to solve this issue, and the technique chosen in this case is the multi-layer long short-term memory (LSTM). LSTM has

a separate memory cell inside that updates and exposes its content only when needed. The layers of those cells are called gates that allow the information to be kept or to be forgotten. In the method proposed, a particular LSTM is used, called BiLSTM and firstly proposed by [138]. This network is capable to move both forward and backwards (Bidirectional). To this architecture, RNN+LSTM, we also add a movement attention mechanism. This, because not all the movements impact in the same way in representing cooperativeness. This attention mechanism, starting from a vector of movements in both the modalities, compute the proper weights associated with the variation in distances, e.g. the movements. If we define as T the number of time steps in the sequence, a_t the weights computed at each time step t and h_t the hidden state vector, the attention can be defined as

$$S = \sum_{i=1}^{T} a_t * h_t \tag{3.3}$$

For this reason, the attention can be seen as 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 literature obtained in real scenario and labeled with cooperative and non-cooperative mode. In GOTCHA-I who we define as a cooperative subject is just ignoring the camera, and who we defined as non-cooperative subject is trying to avoid the camera intentionally. Train and test in the experiment were chosen randomly, and then the results averaged using a 5-fold cross validation. To avoid to create bias in our model, that can learn the gesture of a specific subject, we also do not use the same subject in train and in test. To train the models, the metric chosen is the Matthews Correlation Coefficient (MCC), taking into account the true positives (TP), true negatives (TN), false positive (FP) and false negatives (FN).

MCC can be computed as follow:

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TN + FN) * (TN + FP) * (TP + FN)}} \tag{3.4}$$

Differently from the accuracy, in the case of MCC, if we have a binary classification problem and it is balanced, its 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. An high number can reduce the noise, on the other hand, a lower number can provide a higher precision. In addition, we also considered different distance configurations, taking into account a different number and kind of keypoints. The best configuration, however, is reached when the distances used are head from neck, neck from right shoulder, neck from left shoulder.

Our architecture is then compared with a basic LSTM, a CNN-LSTM (CNN improved with LSTM), ConvLSTM (in which the convolutional step is inside the LSTM), BiLSTM and VGG16. The best results of our method were obtained for 100 buckets and are presented in Table 3.5. Also with 50 and 150 buckets our method outperform the other involved.

Table 3.5: 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 combining Att-RNN with other approaches mentioned in the previous table. The motivation is to study dependencies between the learners. In all the approaches we used the number of buckets that provide the better results in

the correspondent model. Then we weighted their responses in proportion to their trust or estimated performance. We obtained the results in Table 3.6.

Table 3.6: 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%

From this kind of analysis we can deduce that the correlation between stronger models is lower. In fact, in the first row we can notice that the accuracy obtained is lower than the use of our model in a single way. On the other hand, when we fuse our model with BiLSTM who reached lower performances, we obtain a higher accuracy.

The conclusions we can drawn with this approach are: the upper torso part provides the most important information about cooperativeness; Att-RNN can operate in strongly different illumination condition; some methods using similar techniques can be combined with Att-RNN to obtain a slightly better performance (a few lower than +1% in accuracy).

3.4 Conclusions

In this chapter we analysed the human gait and, in particular how it is involved in biometric recognition, from strong to soft traits. We focused on two soft traits belonging to different classes: gender recognition as a soft physical trait; cooperativeness detection

as soft behavioural trait. We introduced the dataset GOTCHA-I, developed by our lab, and particularly suitable to reproduce surveillance and security scenarios. Detect the gender of a user from their gait, reveal to be a competitive task on GOTCHA-I, that of sure need to be analysed with more sophisticated techniques (DNNs) because it demonstrates that is not possible to solve with an accuracy higher than 78% using Machine Learning Techniques that not involve Deep Layers. We focused the most on sophisticated techniques when we tried to solve the cooperativeness detection task. This task represents the main innovative step of the use of GOTCHA-I, since it is the only dataset with labeled data about subjects that, in a free-to-move scenario, tries intentionally to avoid the camera. The results obtained are very encouraging, since the Att-RNN proposed reached around 98% of accuracy recognizing cooperative and non-cooperative users. We think that this could have a high impact on the study of how to approach surveillance videos. We can conclude that soft biometrics trait recognition from gaits requires Deep architectures to be solved with high accuracy, for this reason we are expecting to see a growing use of them in future, since those technologies are now available in almost all academia contexts.

Chapter 4

Soft biometrics to robotics: looking towards the future

In this section we will examine a real use case of soft biometrics. In particular, in recent times, social humanoid robots, thanks to the recent advances in biometrics, and human understanding in general, shown their daily application. In this context, we propose two works we developed on the humanoid robot Pepper. The first one uses a combination of soft and strong biometric traits to built a trust function that allows the user to access at different smart objects. In the second work, we investigated the trust from a user point of view. In particular we develop an interactive technique to extrapolate information from the user, monitoring their level of trust in the robot by the soft biometrics traits.

4.1 The Robots evolution

The use of biometrics in robotics find its application firstly in Service Robot. The classical robots start, in fact, to be equipped with biometrics algorithm to better understand the human commands and behaviour to serve them [139]. We found, in this field, some work related to the ability to follow humans, that require to recognize the human skeleton using depth information [140] or RGB cameras [141]. To better understand the biometric involvement in

robotics and their evolution, we will present a brief history of this field with a particular focus on Humanoid Social Robots.

4.1.1 A brief history of robots and social robots

In 320 B.C., Aristotle was already thinking about robotics with the famous quote:

> *If every instrument could accomplish its own work, obeying or anticipating the will of others, like the statues of Daedalus, or the tripods of Hephaestus, which, says the poet, "of their own accord entered the assembly of the Gods;"if, in like manner, the shuttle would weave and the plectrum touch the lyre without a hand to guide them, chief workmen would not want servants, nor masters slaves.*

The idea of artificial intelligence is deeply rooted in mythology and in popular traditions. As an example, the God Efesto, of the Greek mythology, built the first automata to be helped in his forge, the automata Talos. From the ancient Greek traditions, we move to Alexandria, in 60 B.C., where Erone built the first three-wheeler able to follow a fixed path. The idea of the self propelled wagon will appear again in 1478 by Leonardo Da Vinci, using a spring mechanism. Leonardo himself developed the first humanoid automata documented project: the knight in armor. The innovations in the mechanical and electricity fields, lead to an increasing interest and applications to robotics. It was only in 1920 that the term "Robotics" was coined by Karel Capek, Czechoslovak writer that utilises the latter in the drama R.U.R.. But only in 1940, the term Robot was used for the first time in its modern acceptation, by Isaac Asimov in the field "I, Robot", also introducing three robotic laws. In parallel to the literature, that always tries to anticipate the future applications, we found in 1960 the first robot industrially produced: UNIMATE. UNIMATE was a robot arm, created by George Devol in 1961, for an assembly line. UNIMATE was able to carry melted components and merge them

to the machines. Two years after UNIMATE, in California was developed the first robot arm for people with disabilities. The idea to follow a path, also avoiding obstacles was resumed in 1966 and 1972 by the Stanford Research Institute, but it was only in 1973 that we see for the first time a modern humanoid robot. Wabot-1, was the first robot with arms and legs, able to talk or carry objects. On the other hand, one of the first robot that shows the necessity to be equipped with camera was the rover Viking 1 of NASA, in 1976. After this event, the robotics has had an increasing use in several fields. We can highlight in particular:

- Domestic Robots: the robots to help in domestic actions, in particular concerning the cleaning. As an example the automatic vacuum cleaner, the modern mixers, etc.

- Industrial robot: they are employed in industrial processes, and often look like robotic arms. This permits to speed up the production process, increase profits and avoid workplace accidents.

- Agricultural Robots: they are involved in agriculture, and help different step of the production process. Some robots in this field are used to seed, to harvest, or recently to pollinate.

- Welfare Robots: they are involved in case of need of support for hospitalized patient and elders. As an example, we can find robots that help people to walk or that distribute medicines.

- Medical Robots: They are used to help the doctors during their operations. We can find robots able to guide the doctors during surgery or to help patients during recovery.

- Entertainer Robots: they are used to entertain humans, as playful robots for children, animal robots as Disney Animatronics, etc.

- Educational Robots: they are used to help in the stages of learning. They are mainly programmable robots to help develop this ability.
- Military Robots: they are the automated version of military staff, as an example the automatic tank.
- Domotic Robots: differently from the domestic robots, those kind of machines, help also in other tasks. We can find systems to open doors, change luminosity, activate or deactivate the ventilation system, etc.
- Ecological Robots: this new category of robots are involved in ecological tasks, as clean the seabed, dismantle radioactive waste, etc.
- Space robots: they are involved in space missions to collect information, photos, raise materials or perform mapping.
- Humanoid or Animaloid Robots: the robots we involved in our research, they have an human or animal aspect and are mainly involved in socialization tasks.

In Figure 4.1 we can appreciate the different kind of robots actually involved in real use case.

4.1.2 Humanoid Social Robots characteristics

Humanoid robots are a class of robot specifically built for interaction. They have articulations as legs and arms, and for this reason they are able to perform difficult movement tasks. The Softbank Robotics [142], is actually one of the company selling both humanoid robots and solutions for humanoid robots. The robot we choose for our experiments is developed by this company. The first robot they developed is NAO. Nao is only 58 centimeters tall and weights 5 kg. For this reason, NAO is more involved in the educational field or with kids. In its latest version, NAO has 25 degrees of freedom, it is able to understand if it is sitting or standing. It

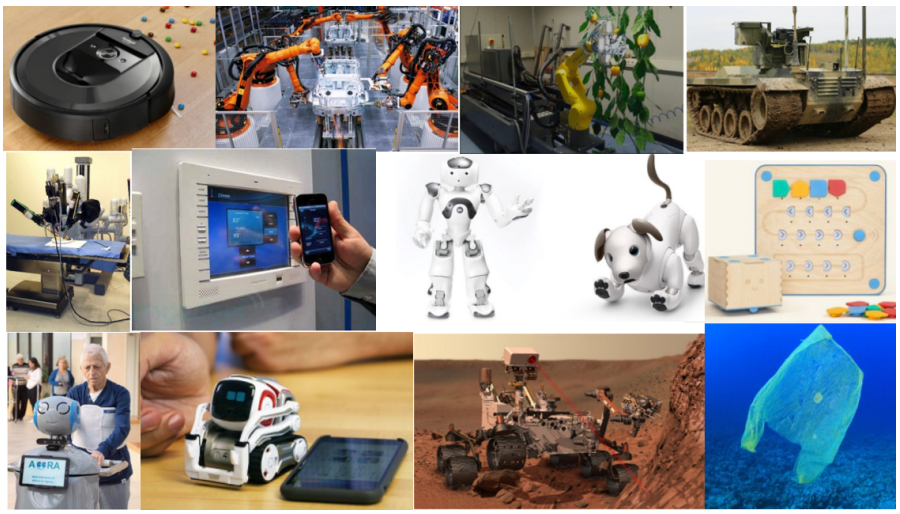

Figure 4.1: From left to right: in the first row, domestic robot, industrial robot, military robot; in the second row, medical robot, domotic robot, humanoid robot, animaloid robot, educational robot; in the third row, welfare robot, entertainment robot, space robot, ecological robot.

is able to walk or run. Equipped with 4 microphones and able to talk in 20 languages. It has also two 2D cameras. The second robot of Softbank is Romeo, born in 2012. It is taller than NAO, 146 centimeters, with 36 kg of weight. It is equipped with the same sensors of NAO and presents 37 degrees of freedom. Both Romeo and NAO can be programmed using the NAOqi library in python, C++ or NET, NAO using 2 processors, and Romeo 4.

The robot used in our research is Pepper, the newest, born in 2014. Differently from Romeo, Pepper has two arms but only one "leg", it is in fact defined semi-humanoid. In fact, Pepper has wheels and a tablet on its torso. As a consequence, it has only 20 degrees of freedom, but, on the other hands, it is more able to interact with humans. It is built to have a friendly aspect and equipped to be a companion robot. The battery autonomy is of 12 hour and its weight is 28 kg for 120 centimeters tall. In Figure 4.2 it is possible to appreciate the proportions and the appearance

of this three robots.

Figure 4.2: From left to right: NAO, Romeo and Pepper.

In Table 4.1 it is possible to check the Pepper specification that we used in the experiments of the methods in Section 4.3.

4.2 Social robots in the state-of-the-art

The state of the art on social robots in recent time mainly involves the human response to robots. In particular, an interesting field of application is represented by children. In some cases, the social robot is not even humanoid, but is capable to help improving social skills in children with ASD using a long-term, in-home interaction [143]. The involvement of the humanoids can be more effective when more interaction is required, as in the case of storytelling frameworks [144], or in the learning of a new language, as in [145]

Table 4.1: PEPPER TECHNICAL SPECIFICATIONS

Size (H x D x W)	1210 x 425 x 480 [mm]
Weight	28kg
Battery	Lithium-ion Typ. Capacity 30.0Ah Energy 795 Wh
Sensors (head) Sensors (trunk) Sensors (hand) Sensors (leg)	Mic x4, RGB camera x2 3D sensor x1, touch sensor x3 Gyroscope sensor x1 Touch sensor x2 Ultrasonic sensor x2, Laser sensor x6, Bumper sensor x3, Gyroscope sensor x1
Degrees of freedom	20
Display	10.1" inches touchable display
OS	NAOqi OS
Network	Wireless / wired interfaces
Velocity	Max. 3 km/h

where NAO is used. In some cases, it was analysed the ability of the social robots to express emotions using a combination of colors, sounds and vibrations [146]. Those characteristics are by default on Pepper, that simulated those traits by its eyes. The abilities of social robots also appear to be involved in motivational tasks [147] or to reduce children anxiety during vaccination [148]. In addition, in a similar application that we will see in Section 4.3.1, social robots can be also useful to help users to interact with the objects of a smart home [149].

If we focus on applications specifically built for Pepper, the fields are mainly dictated by the robot aspect and abilities [150]. Pepper has been tested on assistance on elders, to stimulate them with creative and cognitive activities [151]. On this field, are also

under study the metrics that best represent the elder responses to Pepper stimuli [152]. To be applied to a wider population, the social abilities of Pepper were also tested in Shopping Malls, as in the case of [153] were the expectations of the user about the robot as a service are analysed, or the study in [154], again in a Shopping Mall but with Pepper in the role of entertainer. Because of its friendly aspect is interesting during those experiments, analyse the level of trust of the users that interact with Pepper. Some research in this sense is focused to put in relation the people awareness in robot's capability with their trust in assign the robot different tasks [155]. Or, as in [156], to evaluate the acceptance of Pepper in a program of higher education, in this case a course of academic writing. All these applications have in common the necessity to collect data from the users, both in the case of assistance than to better understand the users reaction to Pepper. For this reason we can find work as [157] focused on health data acquisition, or [158], focused on the way to collect data from the Human-Robot interaction.

4.3 Experiencing soft biometrics on Pepper

In this section we will present two frameworks that use soft biometrics during human-robot interaction. As previously introduced, we used the humanoid social robot Pepper to conduct our experiments. In the first, presented in Section 4.3.1 we used Pepper as an additional check to understand user behaviour and intention and to permit or deny to the user the possibility to interact with the objects in a smart-home. In the second experiment, presented in 4.3.2 we will use Pepper to extract information of the users by some adaptive questions that take into account the emotion of the subject during the interaction. Both the first and the second experiments are based on the concept of trust, however, in the first the trust is evaluated by Pepper, in the second the trust is evaluated by the user.

4.3.1 An eco-system for security in IoT

In the work here proposed, published in [159], we will use Pepper as an interface between the user and the smart objects in a home environment. In a previous framework we used Pepper as a voicemail system secretary to access in a secure way to private messages [160]. However, we used strong biometrics to solve this task, in particular face recognition. The aim of the work here presented is to also integrate soft biometrics on Pepper and extend the trust to all the possible levels required by the smart objects of a smart home. In particular, due to the increasing number of home and infrastructure connected to the Internet of Things (IoT) devices, we want to present a work that considers the involvement of the humanoid robot as a reliable ally. The applications involving IoT must offer a high reliability, not only to protect the private information stored, but also to manage the large amount of information that is acquired to produce the most adequate and suitable answer. For this reason, here we present an IoT eco-system based on the cooperation of different devices with Pepper. Their cooperation has the aim to build a trust model that allows or denies the user to interact with smart objects. In particular, we can split our focus on three main aspects: empowered cameras, empowered Pepper and trust model.

4.3.1.1 Empowered cameras

Fixed cameras represent undoubtedly a valid support to Pepper in an eco-system. Their cost is negligible and this make possible to add them in various room/locations of a building. Their ideal use is in association with methods working on images, at a distance. This is the case of soft biometrics, that, evaluating soft traits, can work even if the resolution is low. In this sense, the first method that we can add to our cameras is the cooperativeness detection by gait, described in Chapter 2, Section 3.3.3. By this system, we can detect if a subject is intentionally trying to avoid the camera. This can be interpreted as a first signal of malicious intent of the subject. Once the non-cooperativeness is detected,

Pepper can reach the subject to better understand their behaviour and intentions. Another method that works at distance, on gait, is the gender detector proposed in Chapter 2, Section 3.3.2. Since both the methods use gait as a skeleton, they can use the same preprocessed data, in order to make the cooperativeness detection and gender classification even faster. In addition, in this ecosystem on cameras we also proposed a facial attributes detection able to work on very different faces in low resolution, proposed by Abate et al in [161]. Those methods reached in literature: 80.7% of accuracy to distinguish gender from gait, 97.58% of accuracy to detect cooperativeness, and 91% of accuracy to classify 40 facial attributes of a subject.

4.3.1.2 Empowered Pepper

It makes sense to use a facial recognition methods on Pepper, as the first step, because of its success in the previously mentioned voicemail application, presented in [160], that reached 91.4% of accuracy in this task. In addition, the soft biometrics we want to use to improve Pepper abilities to understand the user behaviours, involve sentiment analysis. Sentiment analysis is something that can be associated to both facial expression and voice. For this reason, we decided to fuse those two biometrics traits to obtain a general score for emotion recognition. In particular here, we will use the NAOqi system to capture the traits, then two different NNs are used to detect seven emotions: anger; disgust; fear; happiness; sadness; surprise; neutral. The face NN works on Region of Interest (ROI) and the voice NN works on the peaks and intensities of the voice signal. The two scores obtained are fused using a weighted average, that for each of them is represented as follows:

$$res = \frac{v_{em} + acc}{2}$$

where v_{em} is the percentage of the predicted value of the emotion while acc is the Accuracy of the model. The emotion trait, face or voice, associated with the highest weight will provide the final emotion. Starting from only 45% of accuracy from face (tested

on Pepper) and 73% on voice (tested on Pepper), the proposed fusion reaches on Pepper an accuracy of 74.38% outperforming both NNs.

Another method that we used on Pepper is represented by the Heart Rate detection. The heart rate gives us significative information about the user behaviour. We can imagine that an increment in user Heart Rate during questions about his/her behaviour or intentions can be associated to malicious intent. Pepper will record a video sequence, then the user face is extracted and the corresponding ROIs are detected. From this video sequence, it estimates the heart-rate (bpm) during the interaction. To evaluate the HeartRate in bpm, we are interested in the green value of those ROIs, in particular extracted using the mean of 100 consecutive frames. Those method is contactless and has an error margin of 5bpm.

4.3.1.3 Semantic model

To interconnect the smart objects using the information provided by Pepper and Cameras, we developed a trust model based on ontology. In particular, we associate a trust level to every sensor. More trust means the involvement of more accurate biometrics traits. Considering the ontology, the Truster is the entity (i.e., the smart device or Pepper) that estimates the trust value, and the Trustee is the user that performs the action. Applying a semantic approach, we can require, for each command or command chain that the user want to execute, the trust required to perform the execution. From this level of trust, one or more devices and techniques are involved to obtain the trust of this specific user basing on both soft and strong biometrics traits. In Figure 4.4 we can assess the computation required to execute a command chain with the correspondent level of trust associated to the use of each object.

Once the model is defined, we need to define the function that will return the trust collected from the user. In particular, since some biometric traits are stronger than others, as discussed in the

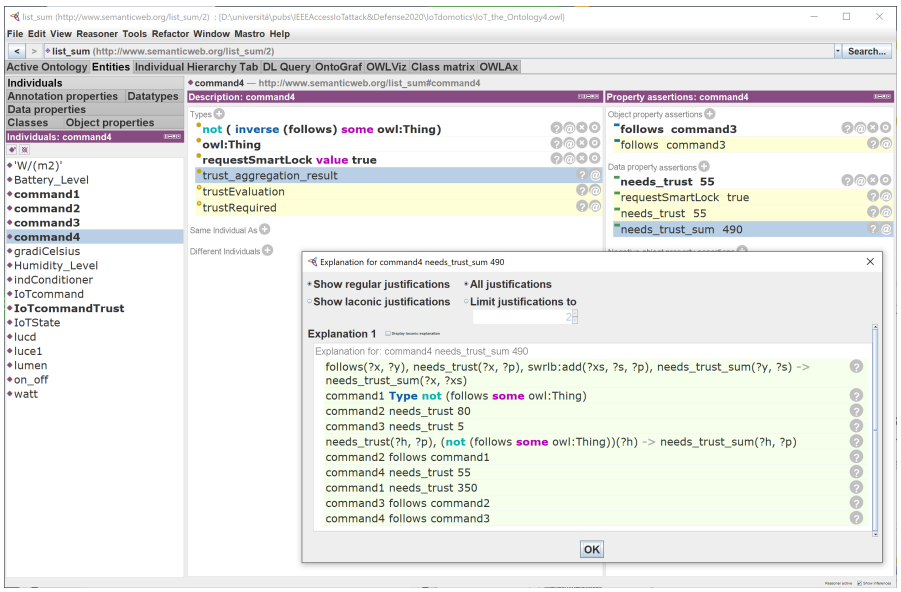

Figure 4.3: The computation of required trust for a command chain.

Introduction Chapter, we will define as k_i the strength coefficient of each biometric trait i. Those coefficient follow the reliability of each involved method from 0 to 1, in ascendant order.

The methods on Pepper and on Cameras are represented, together with their accuracy and k_i associated, in Table 4.2.

Table 4.2: Methods involved, devices, accuracy.

Method	Device	Accuracy	k
Face+Voice Emotion	Pepper	74.38%	0.7
HeartRate	Pepper	5 bpm	0.6
Identity from Face	Pepper	94.1%	1
Gender from Gait	Camera	80.7%	0.3
Cooperativity from Gait	Camera	97.58%	0.2
40 Facial Attributes	Camera	91%	0.4

The function involved in the trust evaluation is defined as:

$$t = \pm k_1 * a_1 \pm k_2 * a_2 \pm ... \pm k_n * a_n \tag{4.1}$$

where, the positive or negative biometric response-characteristics are determined by the sign, the values k_i represent the biometric coefficients and a_i represent the accuracy of the biometric recognition algorithm involved.

4.3.1.4 The smart-home application

We focused our application field on a specific environment: a smart home. In particular, we can imagine a situation as represented in Figure 4.4. In particular, we identified 3 different smart devices, representing different level of required trust.

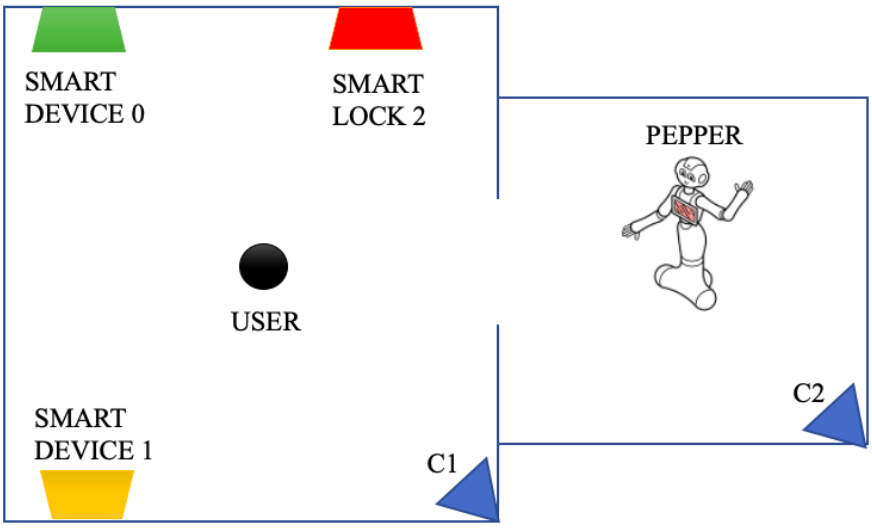

Figure 4.4: The computation of required trust for a command chain.

A smart light or a smart air conditioner can be classified as a smart Device of trust level 0. This kind of devices requires a low level of trust, that can be achieved using only username and

password in a dedicated application. If username and password of an user are stolen, the impostor can only access to very simple devices that do not cause a break in the home security.

A smart device of level 1 is a device precluded from some components of the family or strangers. In this case, the camera C1 of our configuration will be activated to detect soft biometrics from the user. If a user tries to attack this device, the cooperation detection and gait gender recognition can detect malicious intent even if the user tries to impersonate another person.

A smart lock, can be seen as a device that require level 2 of trust to be used, in particular the user must be correctly identified both by the cameras and Pepper with the methods described above. In this case the HeartRate represents the core of our security system. In fact the Robot can ask specific questions to the user, as "Why are you trying to avoid the camera?", "Why do you want to perform this action?", and, by the modification in HeartRate, it is able to detect a malicious intent.

We can conclude, after the presentation of the overall work, that he case study reveals that this kind of approach offers interesting characteristics of generality, extensibility and robustness with respect to possible device compromises and fraud attempts.

4.3.2 A Social Engineering approach

In this approach, as mentioned above, we want to evaluate the user trust in the robot [162]. In particular, here the trust will be essential to perform a human-robot interaction with the robot aiming to extract useful information about the user. The project proposed is, for its nature, in the field of Social Engineering, to convince the user that there are no evil intent during the interaction. Our aim is to make the robot more "empathic" during the interaction and, using this characteristic, try to discover passwords and/or personal data of the user. The framework proposed in this sense are two: SASD (Short Attempt to extract Sensitive Data) and LASD (Long Attempt to extract Sensitive Data), both of them emotionally adaptive.

To build the emotional model, we used what we called the "Emotion Module". This module, depicted in Figure 4.5, is composed by the Emotion from Voice NN and Emotion from Face NN, already used in the work presented in the previous section.

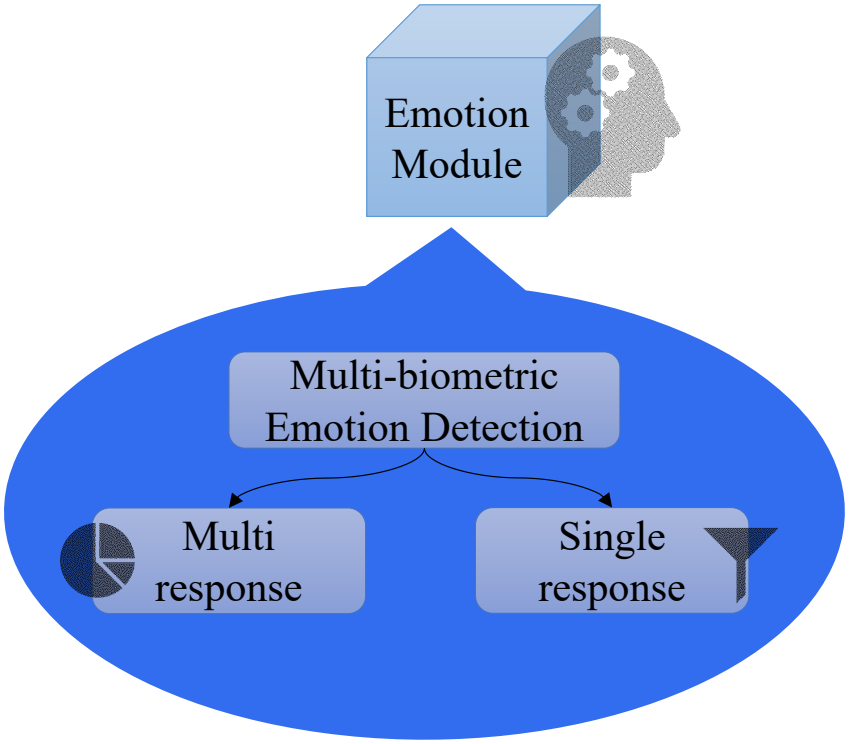

Figure 4.5: The methods inside the Emotion Module.

4.3.2.1 SASD

The SASD architecture, depicted in Figure 4.6 aims at obtaining in a short time-frame, as much information is possible. For this reason, the robot will directly ask the user some key questions to obtain the desired information. However, in order not to arouse or annoying the user, the questions must be asked in a proper way. For this reason, the set of questions proposed to the user in the

SASD modality, must be fixed and accorded with the support of psychology.

This approach can be used, for example, to discover a user password, this because usually the users have as recovery information to obtain a forgotten password, personal information about family, preferences and so on. The Robot, knowing the user identity, by face recognition we already discussed in previous sections, can focus on specific questions to discover these preferences. At the same time, it must be able to recognize some behaviour of the user and change the questions or try to ask the question in a different way once discomfort is detected. The Emotion Module, is here used in this way: if the user has a positive reaction as happiness, curiosity, calm, the robot will continue to interact on the same topic; if on the other hand, the subject seems to be stressed, angry, annoyed, then the robot will change topic.

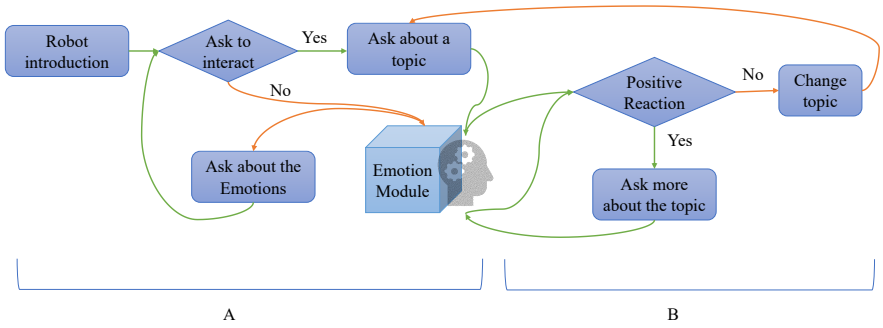

Figure 4.6: The SASD architecture.

A set of preliminary questions will be submitted to the user to start the conversation, as it can be seen from the SASD framework figure, and then, when the user will be ready to interact, the robot will start to ask about a topic. The dialogue on a topic can be made by using specific questions or properly "hidden" in innocuous ones. Pepper will continue to ask questions until all questions will be answered, or until all the *possible* questions will be answered. This because if the robot detects some discomfort in a topic it will not ask about this particular topic again. This characteristic

can be changed in the architecture, if and only if the information is essential. In this case a new set of questions about this topic, formulate in a complete different way, is necessary in the initial set of questions.

4.3.2.2 LASD

In this case, the aim of the conversation is not to obtain targeted information but a large amount of information in a way more comfortable to the user. To reach this aim a large amount of questions are required. For this reason, in this case, we plan to substitute the initial set of questions of the SASD framework with a set of topics and concepts. These topics and concepts will be combined to obtain a larger set of questions. To do this, it is necessary to generate the questions only when necessary, using semantic and grammar rules, combined with an algorithm of column generation.

The conversation, in this case has the first part in common with SASD module (part A) and the second depicted as in Figure 4.7.

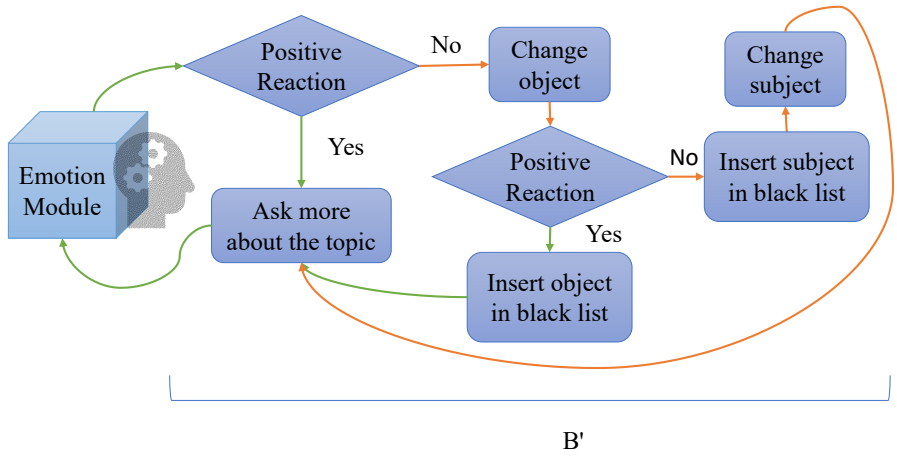

Figure 4.7: The module B of LASD, in red the negative reactions, in green the positive reactions.

Here, we can appreciate the different approach that involves

the negative or positive response given by the Emotion Module. If the reaction on a topic is positive, Pepper will ask more about the topic, examining also subtopics related to the topic (e.g. "family" as the topic, "parents" as a sub-topic of "family"). When a negative reaction is detected, the robot, prior to change the topic, will try to change the subject. As an example, a negative emotional response to the question "What is your father's job?" will lead to a question in which the subtopic of the topic "family" is the same (e.g. "father") but related to another object (not the "job"). If a positive reaction is detected after this change, the related topic (e.g. "job") will be inserted in a black list and not mentioned again. In case the negative reaction, even if the object is changed, the subtopic (the subject "father" in this case) will be inserted in a black list, and the object ("job") will remain available for future questions. In this framework, differently from the ability to steal targeted information in a fixed time, we will evaluate the success of the experiment in the amount of time spent by the user in a positive conversation. In other words, the ability of the robot to perform a pleasant conversation with the users, even if it is trying to store their personal information.

It is clear, from the two frameworks proposed, that we want to demonstrate how it is possible to use soft biometrics not to help the user or guarantee a level of security to a system. But, on the contrary, to steal personal data and information rigging the users, convincing them to be involved in a simple conversation with an humanoid.

4.4 Conclusions

In this Chapter we analysed the real application of soft biometrics on humanoid robots, also supported in some cases by strong biometric traits. We evaluated the involvement of the robot from two opposite point of view. If the use of a humanoid can be more comfortable to provide a security system in a smart home, also supported by the friendly aspect of the robot. On the other hand,

the same friendly aspect of Pepper can be used against the users to steal their personal information in a social engineering framework. These applications of soft biometrics, that have an *essential* role in the success of both methods, demonstrate that those particular traits can be used to guarantee the safety of the users but also as a weapon against them. From this point of view, soft biometrics applied to humanoid robots are undoubtedly a field that needs to be further explored in the future, since the malicious or benevolent intention of the developer can be both amplified by the advances of the state-of-the-art in both fields.

Chapter 5

Conclusions and Future Works

In this Thesis we exposed our proposal in terms of soft biometrics applications, developed in the last three years. From the experience acquired in this context we can claim that soft biometric traits are not intended as an alternative where it is not possible to proceed with classical biometrics, but rather an additional source of information. We explored both physical, as HPE than gender, than behavioural traits, as cooperativeness. From our analysis is clear that the differences in soft biometrics concerning physical and behavioural traits is quite labile. The same trait may be used in different ways to extract different information. Another keypoint we notice during our analysis is the extremely wide range of applications of soft biometrics. Differently from a recognition purpose that can be applied to different contexts, but without differ from the task that is in each case the identification of the subject, soft biometrics are more flexible to be applicable in context that can be also completely unrelated to the goal to recognize the user.

If we analyse those considerations from a technical point of view, related to the specific methods we presented, it is possible to say that:

- The Head Pose Estimation, usually applied to frontalization or best frame selection for recognition, already moved to

a new task: driver attention detection. In this case each rotation of the head is related to a risk factor for the driver that can be quantified by this method. For this reason it will be of great interest to apply this method also to different illumination condition, like thermal, infrared or, even more interesting, depth images.

- Gender recognition from gait can highlight social aspects and controversial related to the concept of gender. On one hand, its concept is related to skeleton, assumed as a physical trait. On the other hand, the temporal modification of skeleton points during gait is something related to the behaviour during walking, e.g. the perceived gender, that is not always related to the physical one. It is indubitable that this aspect should be further analysed, maybe applying those techniques in contexts where only one flow of information is available (skeleton from a single frame, temporal movements in games without skeleton values etc.)

- The cooperativeness detection, here applied to a single dataset built by us, the only one with cooperativeness annotation, shows very interesting results and should be applied also to wider contexts. In this sense, we plan to combine the cooperativeness concept with action recognition, to respond the question: is an anti-cooperative behaviour related to a previous or next malicious action?

- The aspect of social humanoid robots has an high impact on the use of biometrics and soft biometrics, as we demonstrated here. For this reason it is essential to protect the information stored by the service robots in house, that, after days, month or years, knows each habit of the users. In this sense, we can plan to built an efficient way to store and protect those information, maybe combining biometrics and encryption techniques as analysed in other contexts.

In conclusion, we can undoubtedly say that there is a lot of R&D to do in soft biometrics, and that their future involvement will be

of increasing interest for a wide range of applications.

Bibliography

[1] [Online]. Available: https://www.paymentscardsandmobile.com/biometrics-and-fraud-youre-one-in-7-billion/

[2] W. Dahea and H. Fadewar, "Multimodal biometric system: A review," *International Journal of Engineering and Technology*, vol. 4, pp. 25–31, 01 2018.

[3] N. Bhartiya, N. Jangid, and S. Jannu, "Biometric authentication systems: Security concerns and solutions," *Proceedings of the 2018 3rd International Conference for Convergence in Technology (I2CT)*, pp. 1–6, 2018.

[4] "On matching digital face images against scanned passport photos," in *Proceedings of the 2009 First IEEE International Conference on Biometrics, Identity and Security (BIdS)*, 2009, pp. 1–10.

[5] H. Yokozawa, T. Shinzaki, A. Yonenaga, and A. Wada, "Biometric authentication technologies of client terminals in pursuit of security and convenience," *Fujitsu scientific technical journal*, vol. 52, pp. 23–27, 07 2016.

[6] M. De Marsico, M. Nappi, D. Riccio, and H. Wechsler, "Mobile iris challenge evaluation (miche)-i, biometric iris dataset and protocols," *Pattern Recognition Letters*, vol. 57, pp. 17 – 23, 2015, mobile Iris CHallenge Evaluation part I (MICHE I). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865515000574

[7] P. Wei, H. Li, and P. Hu, "Inverse discriminative networks for handwritten signature verification," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[8] G. Iovane, C. Bisogni, L. De Maio, and M. Nappi, "An encryption approach using information fusion techniques involving prime numbers and face biometrics," *IEEE Transactions on Sustainable Computing*, vol. 5, no. 2, pp. 260–267, 2020.

[9] M. Awais, M. J. Iqbal, I. Ahmad, M. O. Alassafi, R. Alghamdi, M. Basheri, and M. Waqas, "Real-time surveillance through face recognition using hog and feedforward neural networks," *IEEE Access*, vol. 7, pp. 121 236–121 244, 2019.

[10] I. Joshi, A. Anand, M. Vatsa, R. Singh, S. D. Roy, and P. Kalra, "Latent fingerprint enhancement using generative adversarial networks," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 895–903.

[11] H. Wechsler and A. S. Toor, "Modern art challenges face detection," *Pattern Recognition Letters*, vol. 126, pp. 3 – 10, 2019, robustness, Security and Regulation Aspects in Current Biometric Systems. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865518300576

[12] S. Barra, C. Bisogni, M. Nappi, and S. Ricciardi, "F-fid: fast fuzzy-based iris de-noising for mobile security applications," *Multimedia Tools and Applications*, vol. 78, pp. 1–21, 05 2019.

[13] S. S. Mgaga, N. P. Khanyile, and J. Tapamo, "A review of wavelet transform based techniques for denoising latent fingerprint images," in *2019 Open Innovations (OI)*, 2019, pp. 57–62.

[14] S. Nand, "The role of speech technology in biometrics, forensics and man-machine interface," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, p. 281, 02 2019.

[15] C. Bisogni and M. Nappi, "Multibiometric score-level fusion through optimization and training," in *2019, Proceedings of the 3rd International Conference on Bio-engineering for Smart Technologies (BioSMART)*, 2019, pp. 1–5.

[16] A. Abate, C. Bisogni, A. Castiglione, R. Distasi, and A. Petrosino, *Optimization of Score-Level Biometric Data Fusion by Constraint Construction Training*, 11 2019, pp. 167–179.

[17] F. Becerra-Riera, A. Morales-González, and H. Vazquez, "A survey on facial soft biometrics for video surveillance and forensic applications," *Artificial Intelligence Review*, vol. 52, 06 2019.

[18] J. D. S. Ortega, P. Cardinal, and A. L. Koerich, "Emotion recognition using fusion of audio and video features," in *Proceedings of the 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 2019, pp. 3847–3852.

[19] D. Ghadiyaram, D. Tran, and D. Mahajan, "Large-scale weakly-supervised pre-training for video action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[20] R. Dupre and V. Argyriou, "A human and group behavior simulation evaluation framework utilizing composition and video analysis," *Computer Animation and Virtual Worlds*, vol. 30, no. 1, p. e1844, 2019, e1844 cav.1844.

[21] M. I. S. B. (MISB), "Misb standard 0601," *UAS Datalink Local Metadata*, 2014.

[22] J. Clark Weeden, C.-A. Trotman, and J. J. Faraway, "Three Dimensional Analysis of Facial Movement in Normal Adults: Influence of Sex and Facial Shape," *The Angle Orthodontist*, vol. 71, no. 2, pp. 132–140, 04 2001.

[23] T. Bakirman, M. U. Gumusay, H. C. Reis, M. O. Selbesoglu, S. Yosmaoglu, M. C. Yaras, D. Z. Seker, and B. Bayram, "Comparison of low cost 3d structured light scanners for face modeling," *Appl. Opt.*, vol. 56, no. 4, pp. 985–992, Feb 2017. [Online]. Available: http://ao.osa.org/abstract.cfm?URI=ao-56-4-985

[24] 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://www.kaggle.com/kmader/biwi-kinect-head-pose-database

[25] I. Lusi, S. Escalera, and G. Anbarjafari, "Sase: Rgb-depth database for human head pose estimation," *Computer Vision â ECCV 2016 Workshops*, pp. 325–336, 11 2016. [Online]. Available: https://icv.tuit.ut.ee/databases/

[26] T. Baltrusaitis, P. Robinson, and L. Morency, "3d constrained local model for rigid and non-rigid facial tracking," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2610–2617. [Online]. Available: https://projects.ict.usc.edu/3dhp/

[27] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization," in *Proceedings of the First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011. [Online]. Available: https://www.tugraz.at/institute/icg/research/team-bischof/lrs/downloads/aflw/

[28] 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: http://cvlab.cse.msu.edu/lfw-and-aflw2000-datasets.html

[29] N. Gourier, D. Hall, and J. L. Crowley, "Estimating face orientation from robust detection of salient facial structures," in *FG NET WORKSHOP ON VISUAL OBSERVATION OF DEICTIC GESTURES*, 2004. [Online]. Available: https://www-prima.inrialpes.fr/Pointing04/data-face.html

[30] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Li, "Face alignment across large poses: A 3d solution," 06 2016, pp. 146–155. [Online]. Available: http://www.cbsr.ia.ac.cn/users/xiangyuzhu/projects/3DDFA/main.htm

[31] P. Barra, C. Bisogni, M. Nappi, D. Freire-Obregón, and M. Castrillón-Santana, "Gotcha-i: A multiview human videos

dataset," in *Security in Computing and Communications*, S. M. Thampi, G. Martinez Perez, R. Ko, and D. B. Rawat, Eds. Singapore: Springer Singapore, 2020, pp. 213–224. [Online]. Available: https://gotchaproject.github.io/

[32] M. Ariz, J. J. Bengoechea, A. Villanueva, and R. Cabeza, "A novel 2d/3d database with automatic face annotation for head tracking and pose estimation," *Comput. Vis. Image Underst.*, vol. 148, no. C, p. 201â210, Jul. 2016. [Online]. Available: http://www.unavarra.es/gi4e/databases/hpdb

[33] S. Muralidhar, L. S. Nguyen, D. Frauendorfer, J.-M. Odobez, M. Schmid Mast, and D. Gatica-Perez, "Training on the job: Behavioral analysis of job interviews in hospitality," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ser. ICMI â16. New York, NY, USA: Association for Computing Machinery, 2016, p. 84â91. [Online]. Available: https://doi.org/10.1145/2993148.2993191

[34] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vision*, vol. 57, no. 2, p. 137â154, May 2004. [Online]. Available: https://doi.org/10.1023/B:VISI.0000013087.49260.fb

[35] Y. Pang, Y. Yuan, X. Li, and J. Pan, "Efficient hog human detection," *Signal Processing*, vol. 91, no. 4, pp. 773 – 781, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168410003476

[36] 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.

[37] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.

[38] F. M. Sukno, J. L. Waddington, and P. F. Whelan, "3-d facial landmark localization with asymmetry patterns and shape regression from incomplete local features," *IEEE Transactions on Cybernetics*, vol. 45, no. 9, pp. 1717–1730, 2015.

[39] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "Poseidon: Face-from-depth for driver pose estimation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 5494–5503. [Online]. Available: https://aimagelab.ing.unimore.it/pandora/

[40] G. Borghi, M. Fabbri, R. Vezzani, S. Calderara, and R. Cucchiara, "Face-from-depth for head pose estimation on depth images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 596–609, 2020.

[41] 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)*, 2015, pp. 4624–4628.

[42] J. Chen, J. Wu, K. Richter, J. Konrad, and P. Ishwar, "Estimating head pose orientation using extremely low resolution images," in *2016 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, 2016, pp. 65–68.

[43] K. Diaz-Chito, J. MartÃnez Del Rincon, A. Hernandez-Sabate, and D. Gil, "Continuous head pose estimation using manifold subspace embedding and multivariate regression," *IEEE Access*, vol. 6, pp. 18 325–18 334, 2018.

[44] Z. Zhao, Q. Zheng, Y. Zhang, and X. Shi, "A head pose estimation method based on multi-feature fusion," in *Proceedings of the 2019 IEEE 7th International Conference on Bioinformatics and Computational Biology (ICBCB)*, 2019, pp. 150–155.

[45] N. Alioua, A. Amine, A. Rogozan, A. Bensrhair, and M. Rziza, "Driver head pose estimation using efficient descriptor fusion," *EURASIP Journal on Image and Video Processing*, vol. 2016, pp. 1–14, 2016.

[46] S. Lathuiliere, R. Juge, P. Mesejo, R. Munoz-Salinas, and R. Horaud, "Deep mixture of linear inverse regressions applied to head-pose estimation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7149–7157.

[47] C. Gou, Y. Wu, F. Wang, and Q. Ji, "Coupled cascade regression for simultaneous facial landmark detection and head pose estimation," in *Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 2906–2910.

[48] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[49] H. Hsu, T. Wu, S. Wan, W. H. Wong, and C. Lee, "Quatnet: Quaternion-based head pose estimation with multiregression loss," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 1035–1046, 2019.

[50] A. Kumar, A. Alavi, and R. Chellappa, "Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors," in *Proceedings of the 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, May 2017, pp. 258–265.

[51] X. Xu and I. A. Kakadiaris, "Joint head pose estimation and face alignment framework using global and local cnn features," *Proceedings of the 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pp. 642–649, 2017.

[52] 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.

[53] J. Xia, L. Cao, G. Zhang, and J. Liao, "Head pose estimation in the wild assisted by facial landmarks based on convolutional neural networks," *IEEE Access*, vol. 7, pp. 48 470–48 483, 2019.

[54] 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 Trans-*

actions on Pattern Analysis and Machine Intelligence, vol. 41, no. 1, pp. 121–135, Jan 2019.

[55] S. Ackland, F. Chiclana, H. Istance, and S. Coupland, "Real-time 3d head pose tracking through 2.5d constrained local models with local neural fields," *Int. J. Comput. Vision*, vol. 127, no. 6â7, p. 579â598, Jun. 2019. [Online]. Available: https://doi.org/10.1007/s11263-019-01152-w

[56] J. Kim, G. Lee, J. Jung, and K. Choi, "Real-time head pose estimation framework for mobile devices," *Mobile Networks and Applications*, vol. 22, 12 2016.

[57] S. Li, K. N. Ngan, R. Paramesran, and L. Sheng, "Real-time head pose tracking with online face template reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1922–1928, 2016.

[58] 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.

[59] P. Barra, C. Bisogni, M. Nappi, and S. Ricciardi, "Fast quadtree-based pose estimation for security applications using face biometrics," in *Network and System Security - Proceedings of the 12th International Conference, NSS 2018, Hong Kong, China, August 27-29, 2018, Proceedings*, ser. Lecture Notes in Computer Science, vol. 11058. Springer, 2018, pp. 160–173. [Online]. Available: https://doi.org/10.1007/978-3-030-02744-5_12

[60] 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.

[61] C. Bisogni, M. Nappi, C. Pero, and S. Ricciardi, "Hp2ifs: Head pose estimation exploiting partitioned iterated function systems," *Proceedings of the 25th International Conference on Pattern Recognition (ICPR2020)*, 2020.

[62] 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

[63] H. Abdi, "Partial least square regression (pls regression)," *Encyclopedia for research methods for the social sciences*, vol. 6, no. 4, pp. 792–795, 2003.

[64] 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

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] Aristotle, *On the Gait of Animals*. Kessinger Publishing, 2004.

[76] O. Fisher and W. Braune, *Der Gang des Menschen: Versuche am unbelasteten und belasteten Menschen*, 1980.

[77] F. Miller and J. Henley, *Diagnostic Gait Analysis Use in the Treatment Protocol for Cerebral Palsy*, 01 2017.

[78] M. Wang, X. Wang, Z. Fan, F. Chen, S. Zhang, and C. Peng, "Research on feature extraction algorithm for plantar pressure image and gait analysis in stroke patients," *Journal of Visual Communication and Image Representation*, vol. 58, pp. 525 – 531, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047320318303419

[79] N. Farahpour, A. Jafarnezhad, M. Damavandi, A. Bakhtiari, and P. Allard, "Gait ground reaction force characteristics of low back pain patients with pronated foot and able-bodied individuals

with and without foot pronation," *Journal of Biomechanics*, vol. 49, no. 9, pp. 1705 – 1710, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0021929016304237

[80] A. Muro-de-la Herran, B. Garcia-Zapirain, and A. Mendez-Zorrilla, "Gait analysis methods: An overview of wearable and non-wearable systems, highlighting clinical applications," *Sensors*, vol. 14, no. 2, p. 3362â3394, Feb 2014. [Online]. Available: http://dx.doi.org/10.3390/s140203362

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] A. S. M. H. Bari and M. L. Gavrilova, "Artificial neural network based gait recognition using kinect sensor," *IEEE Access*, vol. 7, pp. 162 708–162 722, 2019.

[86] R. Sahak, N. K. Zakaria, N. M. Tahir, A. I. M. Yassin, and R. Jailani, "Review on current methods of gait analysis and recognition using kinect," in *Proceedings of the 2019 IEEE 15th International Colloquium on Signal Processing Its Applications (CSPA)*, 2019, pp. 229–234.

[87] M. D. Marsico and A. Mecca, "A survey on gait recognition via wearable sensors," *ACM Comput. Surv.*, vol. 52, no. 4, Aug. 2019. [Online]. Available: https://doi.org/10.1145/3340293

[88] M. De Marsico and A. Mecca, "Biometric walk recognizer: Gait recognition by a single smartphone accelerometer," *Multimedia Tools and Applications*, vol. 76, 06 2016.

[89] L. Rong, D. Zhiguo, Z. Jianzhong, and L. Ming, "Identification of individual walking patterns using gait acceleration," in *Proceedings of the 2007 1st International Conference on Bioinformatics and Biomedical Engineering*, 2007, pp. 543–546.

[90] M. O. Derawi, P. Bours, and K. Holien, "Improved cycle detection for accelerometer based gait authentication," in *Proceedings of the 2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2010, pp. 312–317.

[91] M. O. Derawi, C. Nickel, P. Bours, and C. Busch, "Unobtrusive user-authentication on mobile phones using biometric gait recognition," in *Proceedings of the 2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2010, pp. 306–311.

[92] L. Rong, Z. Jianzhong, L. Ming, and H. Xiangfeng, "A wearable acceleration sensor system for gait recognition," in *Proceedings of the 2007 2nd IEEE Conference on Industrial Electronics and Applications*, 2007, pp. 2654–2659.

[93] 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.

[94] 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.

[95] M. Nowlan, "Human recognition via gait identification using accelerometer gyro forces," 2009.

[96] M. Gadaleta and M. Rossi, "Idnet: Smartphone-based gait recognition with convolutional neural networks," *Pattern Recognition*, vol. 74, pp. 25 – 37, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320317303485

[97] 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.

[98] A. Sokolova and A. Konushin, "Methods of gait recognition in video," *Programming and Computer Software*, vol. 45, pp. 213–220, 07 2019.

[99] 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.

[100] 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.

[101] 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.

[102] S. Choi, J. Kim, W. Kim, and C. Kim, "Skeleton-based gait recognition via robust frame-level matching," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 10, pp. 2577–2592, 2019.

[103] L. Yao, W. Kusakunniran, Q. Wu, J. Zhang, and Z. Tang, "Robust cnn-based gait verification and identification using skeleton gait energy image," in *2018 Digital Image Computing: Techniques and Applications (DICTA)*, 2018, pp. 1–7.

[104] K. Jun, D. Lee, K. Lee, S. Lee, and M. S. Kim, "Feature extraction using an rnn autoencoder for skeleton-based abnormal gait recognition," *IEEE Access*, vol. 8, pp. 19 196–19 207, 2020.

[105] Y. Liu, X. Jiang, T. Sun, and K. Xu, "3d gait recognition based on a cnn-lstm network with the fusion of skegei and da features," in *Proceedings of the 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2019, pp. 1–8.

[106] D. Lee, K. Jun, S. Lee, J. Ko, and M. S. Kim, "Abnormal gait recognition using 3d joint information of multiple kinects system and rnn-lstm," in *Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 542–545.

[107] W. An, S. Yu, Y. Makihara, X. Wu, C. Xu, Y. Yu, R. Liao, and Y. Yagi, "Performance evaluation of model-based gait on multi-view very large population database with pose sequences," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 4, pp. 421–430, 2020.

[108] J. P. Singh, S. Jain, S. Arora, and U. P. Singh, "Vision-based gait recognition: A survey," *IEEE Access*, vol. 6, pp. 70 497–70 527, 2018.

[109] 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.

[110] A. Jain and V. Kanhangad, "Gender classification in smartphones using gait information," *Expert Systems with Applications*, vol. 93, pp. 257 – 266, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417417306917

[111] 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.

[112] 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)*, 2019, pp. 1–4.

[113] 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*, vol. 1123 CCIS, 2019.

[114] 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*, vol. 30, pp. 629–644, 2019.

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

[116] P. Punyani, R. Gupta, and A. Kumar, "Human age-estimation system based on double-level feature fusion of face and gait images," *International Journal of Image and Data Fusion*, vol. 9, no. 3, pp. 222–236, 2018.

[117] T. Islam, L. Awasthi, and U. Garg, *Gender and Age Estimation from Gait: A Review*, 01 2021, pp. 947–962.

[118] E. S. Jaha and M. S. Nixon, "From clothing to identity: Manual and automatic soft biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 10, pp. 2377–2390, 2016.

[119] 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*, vol. 19, no. 13, pp. 5195–5203, 2019.

[120] 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*, vol. 14, no. 5, pp. 818–829, 2020.

[121] 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

[122] F. Becattini, T. Uricchio, L. Ballan, L. Seidenari, and A. Del Bimbo, "Am i done? predicting action progress in videos," 05 2017.

[123] 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

[124] 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)*, March 2020.

[125] S. Srivastava, V. Rastogi, C. Prakash, and D. Sethi, *Robust Approach for Emotion Classification Using Gait*, 01 2021, pp. 885–894.

[126] S. Xu, J. Fang, X. Hu, E. Ngai, Y. Guo, V. C. M. Leung, J. Cheng, and B. Hu, "Emotion recognition from gait analyses: Current research and future directions," 2020.

[127] "Relationships between gait and emotion in parkinson's disease: A narrative review," *Gait Posture*, vol. 65, pp. 57 – 64, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0966636218309494

[128] Y. Yagi, I. Mitsugami, S. Shioiri, and H. Habe, *Behavior Understanding Based on Intention-Gait Model*, 04 2017, pp. 139–172.

[129] 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*,

vol. 197-198, p. 102991, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S107731422030062X

[130] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017.

[131] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," 06 2016.

[132] K. Soomro, A. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, 12 2012.

[133] "Cmu motion capture database." [Online]. Available: http://mocap.cs.cmu.edu/

[134] T. Randhavane, A. Bera, K. Kapsaskis, U. Bhattacharya, K. Gray, and D. Manocha, "Identifying emotions from walking using affective and deep features," 06 2019.

[135] C. Xu, Y. Makihara, G. Ogi, X. Li, Y. Yagi, and J. Lu, "The ou-isir gait database comprising the large population dataset with age and performance evaluation of age estimation," *IPSJ Transactions on Computer Vision and Applications*, vol. 9, 12 2017.

[136] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," 10 2012, pp. 609–623.

[137] P. Barra, C. Bisogni, M. Nappi, D. Freire-Obregón, and M. C. Santana, "Gait analysis for gender classification in forensics," in *DependSys*, 2019.

[138] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[139] J. Wirtz, P. Patterson, W. Kunz, T. Gruber, V. Lu, S. Paluch, and A. Martins, "Brave new world: service robots in the frontline," *Journal of Service Management*, vol. 29, pp. 907–931, 2018.

[140] S. Sun, N. An, X. Zhao, and M. Tan, "Human recognition for following robots with a kinect sensor," in *Proceedings of the 2016 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2016, pp. 1331–1336.

[141] W. Chi, J. Wang, and M. Q. . Meng, "A gait recognition method for human following in service robots," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 9, pp. 1429–1440, 2018.

[142] [Online]. Available: https://www.softbankrobotics.com/emea/en

[143] B. Scassellati, L. Boccanfuso, C.-M. Huang, M. Mademtzi, M. Qin, N. Salomons, P. Ventola, and F. Shic, "Improving social skills in children with asd using a long-term, in-home social robot," *Science Robotics*, vol. 3, no. 21, 2018.

[144] D. Conti, C. Cirasa, S. Di Nuovo, and A. Di Nuovo, ""robot, tell me a tale!": A social robot as tool for teachers in kindergarten," *Interaction Studies*, vol. 21, no. 2, pp. 220–242, 2020. [Online]. Available: https://www.jbe-platform.com/content/journals/10.1075/is.18024.con

[145] J. Kennedy, P. Baxter, E. Senft, and T. Belpaeme, "Social robot tutoring for child second language learning," in *Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016, pp. 231–238.

[146] S. Song and S. Yamada, "Expressing emotions through color, sound, and vibration with an appearance-constrained social robot," in *Proceedings of the 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, 2017, pp. 2–11.

[147] H. W. Park, R. Rosenberg-Kima, M. Rosenberg, G. Gordon, and C. Breazeal, "Growing growth mindset with a social robot peer," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 137â145. [Online]. Available: https://doi.org/10.1145/2909824.3020213

[148] S. Rossi, M. Larafa, and M. Ruocco, "Emotional and behavioural distraction by a social robot for children anxiety reduction during vaccination," *Proceedings of the International Journal of Social Robotics*, vol. 12, pp. 765–777, 2020.

[149] M. Luria, G. Hoffman, and O. Zuckerman, "Comparing social robot, screen and voice interfaces for smart-home control," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 580â628. [Online]. Available: https://doi.org/10.1145/3025453.3025786

[150] A. K. Pandey and R. Gelin, "A mass-produced sociable humanoid robot: Pepper: The first machine of its kind," *IEEE Robotics Automation Magazine*, vol. 25, no. 3, pp. 40–48, 2018.

[151] D. Unbehaun, K. Aal, and R. Wieching, "Creative and cognitive activities in social assistive robots and older adults: Results from an exploratory field study with pepper," 06 2019.

[152] L. Bechade, G. Dubuisson-Duplessis, G. Pittaro, M. Garcia, and L. Devillers, *Towards Metrics of Evaluation of Pepper Robot as a Social Companion for the Elderly*, 01 2019, pp. 89–101.

[153] M. Niemelä, P. Heikkilä, and H. Lammi, "A social service robot in a shopping mall: Expectations of the management, retailers and consumers," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 227â228. [Online]. Available: https://doi.org/10.1145/3029798.3038301

[154] I. Aaltonen, A. Arvola, P. Heikkilä, and H. Lammi, "Hello pepper, may i tickle you? children's and adults' responses to an entertainment robot at a shopping mall," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 53â54. [Online]. Available: https://doi.org/10.1145/3029798.3038362

[155] A. Rossi, P. Holthaus, K. Dautenhahn, K. L. Koay, and M. L. Walters, "Getting to know pepper: Effects of people's awareness of a robot's capabilities on their trust in the robot," in *Proceedings of the 6th International Conference on Human-Agent Interaction*, ser. HAI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 246â252. [Online]. Available: https://doi.org/10.1145/3284432.3284464

[156] J. Guggemos, S. Seufert, and S. Sonderegger, "Humanoid robots in higher education: Evaluating the acceptance of pepper in the context of an academic writing course using the utaut," *British Journal of Educational Technology*, vol. 51, no. 5, pp. 1864–1883, 2020.

[157] D. van der Putte, R. Boumans, M. Neerincx, M. O. Rikkert, and M. de Mul, "A social robot for autonomous health data acquisition among hospitalized patients: An exploratory field study," in *Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2019, pp. 658–659.

[158] G. Suddrey and N. Robinson, "A software system for human-robot interaction to collect research data: A html/javascript service on the pepper robot," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 459â461. [Online]. Available: https://doi.org/10.1145/3371382.3378287

[159] A. F. Abate, P. Barra, C. Bisogni, L. Casone, and I. Passero, "Contextual trust model with a humanoid robot defense for attacks to smart eco-systems," *IEEE Access*, 2020.

[160] P. Barra, C. Bisogni, A. Rapuano, A. F. Abate, and G. Iovane, "Himessage: An interactive voice mail system with the humanoid robot pepper," in *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, 2019, pp. 652–656.

[161] 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*, vol. 8, pp. 9037–9045, 2020.

[162] C. Bisogni, L. Cascone, A. Castiglione, G. Costabile, and I. Mercuri, "Social robot interactions for social engineering: Opportunities and open issues," *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing*, 2020.

Acknowledgements

Riferendosi al suo mentore, Rita Levi-Montalcini diceva *"La scelta di un giovane dipende dalla sua inclinazione, ma anche dalla fortuna di incontrare un grande maestro."* E nelle sue parole mi ritrovo completamente nel ringraziare il Prof. Michele Nappi per il grande lavoro che ha svolto nel fornirmi gli strumenti e le opportunitá per orientarmi nel mondo della Ricerca. Unitamente, rivolgo la mia gratitudine anche ai miei co-autori, in particolare ai Prof. Aniello Castiglione e Fabio Narducci, che piú volte mi sono stati di esempio e da guida in questi anni.

E siccome la famiglia é la patria delle virtú, non posso certo evitare di ringraziare la mia famiglia, di nascita ed acquisita, nell'avermi sempre incoraggiata e appoggiata nelle mie scelte, nonché di avermi fornito, con l'educazione e l'esempio, la determinazione che mi ha sempre contraddistinto.

"Occorre notevole ardimento per affrontare i nemici, ma molto di piú per affrontare gli amici." E per questo ringrazio Umberto, con la sua onestá, sempre in grado di tenermi testa permettendomi di migliorare come ricercatore e come individuo.

E, infine, con ancora una frase del premio Nobel Levi-Montalcini, *"Le donne che hanno cambiato il mondo non hanno mai avuto bisogno di mostrare nulla, se non la loro intelligenza."*, faccio un enorme in bocca al lupo alle mie colleghe, le Dottoresse Paola Barra, Lucia Cascone e Chiara Pero, che, unitamente a me, stanno costruendo il loro futuro su questo principio.