ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Energy Engineering (33rd Cycle)

RGB-D methodologies for Face Expression Recognition

Luca Ulrich

* * * * * *

Supervisors
Prof. Maria Grazia Violante, Supervisor
Prof. Sandro Moos, Co-Supervisor

Doctoral Examination Committee:
Prof. Speranza Domenico, Università degli studi di Cassino e del Lazio meridionale
Prof. Motyl Barbara, Università degli Studi di Udine
Prof. Tistarelli Massimo, Università di Sassari
Prof. Ramieri Guglielmo, Facoltà di Medicina e Chirurgia dell’Università di Torino
Prof. Borgianni Yuri, Libera Università di Bolzano

Politecnico di Torino

I hereby declare that the contents and organisation of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Luca Ulrich
……………………………………

Luca Ulrich
Turin, December 15, 2020

Summary

In last years, application fields for face analysis have considerably increased, phenomenon that has been fed by a significant improvement of depth acquisition technologies. In this work, the focus is pointed on face expression recognition, which aims to recognize user's feelings through the analysis of facial data. Data considered in this research are those acquired from RGB-D cameras, a family of devices that allows to obtain both color and depth information.

Some examples of fields using face expression recognition are security, for instance face analysis can reveal criminal's truthfulness during an interrogation; automotive, to properly adapt the environment modifying music, lights and alerts of the vehicle monitoring the driver's mood or to adapt the driving style of an autonomous vehicle to passenger's emotional state; culture, to monitor and to react to changes in audience's mood, but also videogames and, as the reader will be able to read later on, emotional design.

This study aims to investigate the context of facial expression recognition and suitable sensors to acquire the human face, in order to develop an automatic procedure able to perform real-time face expression recognition using RGB-D cameras. In order to achieve this goal, work has been split in steps that have been briefly presented in the next few lines and described in different chapters.

First, a literature review has been conducted to understand differences about the main facial applications that have been identified in face detection, face authentication, face identification and face expression recognition. RGB-D cameras have been studied and compared to identify the most suitable technology and, subsequently, the most suitable camera for face expression recognition. RGB-D cameras have been chosen for our purposes to include the third dimension, increasing reliability and robustness. The research has led us to identify structured light as the most suitable depth acquisition technology for the purpose of this work,

so Intel RealSense SR300 has been selected to be used during the experiments. This part of the research has been described in Chapter 1.

Chapter 2 introduces a Support Vector Machine methodology aiming to identify the activation level of a subject's emotion. SVM relies on an automatic landmarking procedure involving geometrical descriptors and on geometrical descriptors themselves used as features for the classification. The method has been applied to a case study designed to make use of depth maps provided by the camera.

Chapter 3 introduces the usage of deep learning to obtain face expression recognition. This project has been conducted jointly with a team of Politecnico di Milano and aims to identify spontaneous emotions of people during an experiment and to build an ecological dataset. The experiment consists in showing some pictures belonging to public databases and validated to arouse specific emotions to the viewer. Meanwhile, an RGB-D camera records people reactions and data are stored for a later analysis through Convolutional Neural Network (CNN).

Chapter 4 describes the real-time procedure set up to obtain face expression recognition. The procedure consists of acquisition, data processing and recognition through CNN. In this chapter, the focus is on data processing, since all the operations (RGB and Depth alignment, face detection on depth map, cropping and resizing) have been automatized and optimized to obtain real-time.

Chapter 5 shows the work produced in collaboration with EURECOM, France. The study of women facial proportions has benefitted from the know-how of two different research groups regarding human face geometry and the concept of *standard face* and has resulted to be core to consolidate the background for the comprehension of feature extraction techniques. A ranking of the most significant measures (Euclidean distances, angular measures, and ratio between distances) has been drawn up.

Acknowledgment

I would like to acknowledge prof. Jean-Luc Dugelay, my external supervisor during the visiting period spent in EURECOM, Sophia Antipolis, France.

Contents

List of Tables

List of Figures

Chapter 1

Analysis of RGB-D camera technologies for supporting different facial usage scenarios

To change some characteristics on the face to appear more beautiful, older, or even frightful; to unlock a personal device pointing the camera on the users; to fulfill money transfers in a secure way; to create new, personal emojis to express itself at its best; to detect faces on pictures for tagging friends; to suggest a suitable playlist that fits the user's mood; these are all example of facial applications, which everyday increase and become more creative, thanks to their versatility and appreciation by the users.

Facial applications usually employ artificial intelligence techniques to identify and to extract patterns on faces from a huge quantity of data, images, and videos acquired with different types of cameras. In the last decade, RGB-D camera have become widespread, giving the opportunity to obtain not only the traditional color representation of scene, subject and environment, but also its depth representation, locating every single point in the space and opening up a world of possibilities through the usage of 3D data.

Since the presence on the market of a considerable number of RGB-D cameras has been observed, an overview on depth acquisition technologies and main facial applications has been considered necessary in this study to clarify the possible usage scenarios. This chapters aims to be the starting point of the thesis work, laying the foundations of the know-how regarding the acquisition process of RGB-D videos and images.

This part of the work has been published in Ulrich et al. [1].

1.1 Introduction

In recent years a considerable number of applications have benefited from the usage of the third dimension [2]; there are several research fields in which 3D is currently successfully used: safety, such as for autonomous driving [3]; orthopedics, for both diagnosis and treatment planning [4]; surgery, as 3D models reconstruction gives the possibility of organizing medical equipment [5], attending the surgeon during the intervention and supporting the post-operative evaluation of the results [6]; 3D printing applications [7], including facial prosthesis [8], dental implants [9] and pelvis prosthesis [10].

The ambition of accelerating the evolution process of cities into interconnected communities brings out other application areas as candidates for heavy 3D usage: land surveying [11], architecture [12], archaeology [13] for research and tourism purposes and also security. Smart cities, urban area equipped with interconnected sensors able to collect data to be used to manage products and services [14], aim to benefit of the spreading of face recognition technology and deep learning techniques to solve problems such as quickly finding missing children and identifying criminals [15] or monitoring public places such as airports [16]. Geometry of the surfaces acquired with sensors capable of capturing depth information can be used for a more accurate face reconstruction [17], to build 3D aging models [18], face manipulation [19] and landmarking [20]. As it will be better explained in the next section, facial applications are shiny examples of this consideration, since the face acquisition can be performed in different conditions depending on various usage scenarios.

3D techniques require a higher computational cost than 2D methods [21], especially if the 3D face model has to be reconstructed from multi-view images [22] or through 3D morphable models obtained from 2D images and 3D scans or even without 3D data [23]. Nonetheless, the robustness given by the opportunity to operate in critical lighting conditions [24], in presence of occlusions [25] [26] and regardless of the orientation of the subject [27] make a 3D approach preferable.

Literature about 3D is varied and fragmented due to lack of a shared methodology for analyzing the field and developing new applications in the face of a growing number of RGB-D cameras on the market. This scientific survey has been conducted to converge on a unique standard and to provide a baseline for the design of the following 3D facial applications in real-time: face detection, face authentication, face identification and face expression recognition.

Total time required by a facial application to be performed is the sum of the acquisition time and the processing time. The first one is the time required to obtain the RGB-D information and depends on 3D cameras; the second one involves the processing of the acquired depth information that is necessary to obtain a result. Since the latter does not depend on 3D cameras, but on other elements of the framework which constitutes an application, for instance the face analysis algorithms, it has not been analyzed in the present work, which focuses on acquisition technology.

This work aims to be a guide for the right choice of an RGB-D camera depending on the facial application that has to be implemented. The focus is on camera technologies able to provide RGB images and depth maps, namely images on which each pixel has a value representing the distance from the camera; 3D scanners have not been taken into consideration, because they do not work in real-time, since they require a too high minimum technical time to complete the scan.

The study is structured as follows. Section 2 focuses on facial applications and on 3D sensor technologies; an explanation of the methodologies used for the investigation is provided. Survey results are presented in Section 3, while in Section 4 conclusions have been drawn.

1.2 Methodological analysis

This survey has been carried out through a two steps analysis. First, a desk research has been performed to qualitatively investigate two aspects of 3D: the available technologies for computing the depth and the facial applications able to benefit from 3D usage that have been developed up to now. A desk research is a complete review of the literature, including articles and datasheets, indispensable to deeply analyze the functioning and the potentialities of the 3D sensors [28]. Secondly, QFDs (Quality Function Deployments) [29] have been used to quantitatively examine the relationships between two different orthogonal dimensions, namely the qualitative requirements typical of each facial application and the technical specifications of 3D acquisition technologies. Both dimensions have been obtained from the results of the desk research.

1.2.1 Desk research on facial applications

The opportunity of understanding and extracting information from human face has interested many researchers in past decades, giving birth to a new discipline called *Face Perception* [30]. Human brain has the ability of figuring out characteristics such as identity, age, sex, mood and better understanding a conversation if the person is looking at his interlocutor. It has been proved that infants already possess this ability from birth [31], and develops it during growth [32]. The lack of this capability is considered an impairment, a cognitive disorder called prosopagnosia [33] [34], that can be congenital or be caused by serious brain damages.

Since the recent spread of Computer Vision outcomes have highlighted that the utilization of technologies able to emulate human behavior is desirable, the idea of automatizing the face perception process has come up. Nevertheless, human brain functioning is highly complex and nowadays the possibility of reconstructing a model able to perfectly replay its behavior is remote. That explains why, in literature, all the applications related to the automatic recognition of specific features on human face are studied individually.

In this part of the research, the main facial applications have been considered: face detection, face recognition, with the two declinations in face authentication and face identification, and face expression recognition (Figure 1.1). The purpose of this study is to highlight the main aspects of these applications to have more elements on which to base the search for the best depth acquisition technology in relation to single applications.

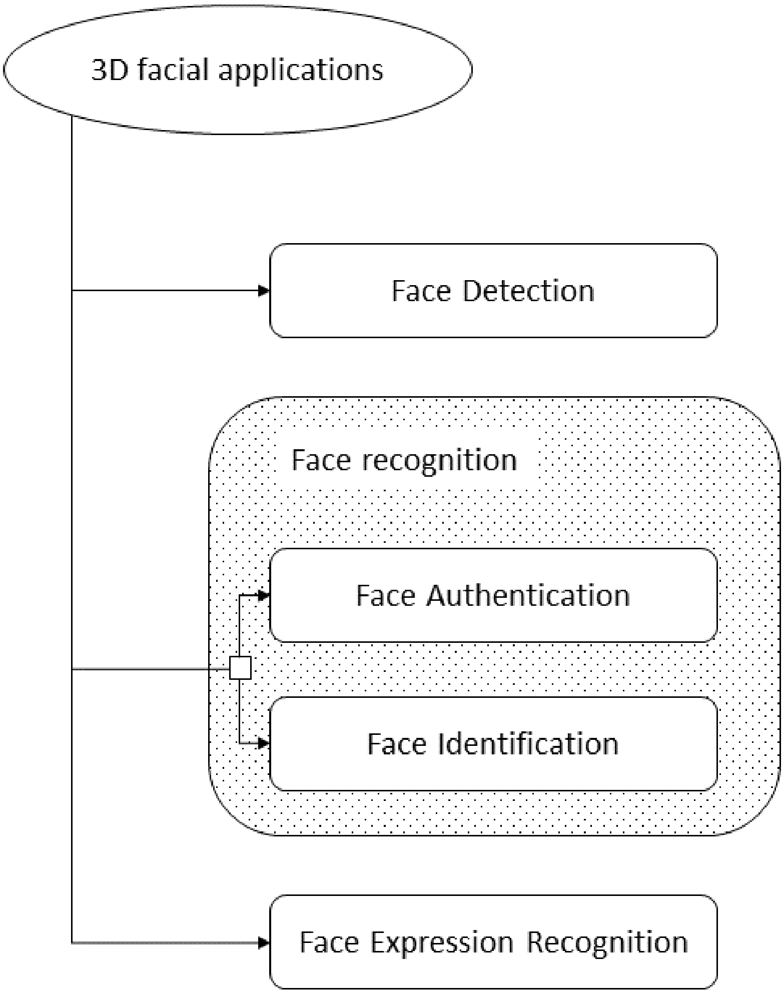


Figure 1.1 - 3D facial applications considered in the present work: Face Detection [35] [36] [37] [38], Face Authentication [39], Face Identification [40] [41] [42] and Face Expression Recognition [43] [44]

Face Detection

Face detection aims to detect a face shape inside an image or inside a frame in the case of a video stream [35]. This operation is often implemented to identify the region of interest into an image for further processing, typically another facial application, to discard the background and consequently to improve computational speed focusing on the area that carries the relevant information, eventually normalizing the image with rotation and scaling operations.

Nonetheless, face detection could also be used stand-alone in various applications: Facebook has implemented a face detection algorithm to insert tags on pictures [36]; in the context of video surveillance Erden et al. [37] have proposed a solution to count the number of people in a room using cascaded AdaBoost classifiers; Lamba et al. [38] have studied an approach to detect faces among crowd based on the color of skin and Histogram of Oriented Gradients (HOG); recently Loey et al. [39] have introduced a machine learning algorithm for face mask detection.

One of the main contribution to this field has been given by the Viola-Jones object detection framework definition [40], a robust and fast solution to the face detection problem, and in the next years other face detection techniques working on 2D images have been proposed [41], such as Liu [42], which has presented a Bayesian discriminating features method and Feraund et al. [43], that have used a constrained generative model (CGM) to detect side view faces.

Several studies have been carried on 3D techniques, eventually combining the information provided by both the color and the depth information. Colombo et al. [44] have presented a method to detect the eyes and the nose analyzing curvature of the surfaces of 3D faces acquired with a laser range scanner, Heisele et al. [45] have developed a framework to perform face detection and face identification using two different sets of facial elements to train the classifiers, Maes et al. [46] have adapted a scale-invariant feature transform for 3 surfaces, Neethu et al. [47] have combined Normalized Pixel Difference (NPD), haar classifier and haar classifier for profile face. More recently, deep learning has become the most widely used solution in the field of face detection: Zhang et al. [48] have introduced a deep cascaded framework using Convolutional Neural Network (CNN) to identify face and landmark location in a coarse-to-fine strategy, Jiang et al. [49] have proposed a Faster Region-based Convolutional Neural Network (Faster R-CNN) approach as well as Sun et al. [50].

The most evident advantage of using 3D techniques is an enhanced robustness result of lighting, pose and occlusion independence.

Face Authentication

A taxonomy clarification is mandatory to deepen the discussion about face recognition applications. Face recognition aims to recognize a face detected into an image or into a frame, comparing it with another face or with a set of faces contained into a database.

Face authentication belongs to biometric systems, that are solutions implemented to control the access to a private area using specific features of individuals [51].

Fingerprints [52] and iris scans [53] are two of the well-known biometric systems to authenticate an individual, but face authentication is becoming a more and more common solution in the case of identity certification for personal devices, especially for laptops and smartphones, due to their spread and their involvement in an increasing number of operations for the management of personal information, such as social accounts, email, bank accounts, and also to fulfil payment [54].

Some examples of methods explicitly stated as face authentication algorithms have been presented by Jonsonn at al. [55], that have used Support Vector Machine

(SVM), by Tao et al. [56], that have based their work on Viola-Jones detector and Samangouei et al. [57], which have used classifiers trained with specific facial attributes. Moreover, it is not uncommon that information obtained with different biometric systems are merged together to further improve robustness and to avoid issues due to data affected by noise, non-universality of biometric traits or simply to improve the error rates [58], for instance merging thermal and visible face images [59]; the score fusion methods are non-trivial tasks, since different users can experience different error rates, indeed new solutions to dynamically assign weights to the scores provided by different authentication systems are under study [60].

On the customer-grade market face authentication has been introduced by Microsoft, with Microsoft Hello, Google, with Facenet, iProov, Megvii, with Face++ that is currently property of Alibaba and Apple with Face ID. Among the elements that have fostered the interest in face authentication by these companies, there are two reasons: one is methodological, one is technological. The first one is the publishing of a paper by Taigman et al. [61] in 2014, in which the possibility of using deep learning, and more specifically Convolutional Neural Network, has been shown. The neural network was trained with a huge number of labelled images and it is not surprising that some members of the research group were part of Facebook. Authors have claimed that the results reduce the state-of-the-art error by more than 27%. The latter is the usage of an ever-increasing number of sensors to retrieve more information from the scene and the improvement of the existing ones, such as the infrared illuminators by which it is possible to use the structured light technology and to get high quality depth information. The high degree of security requested to protect a personal device implicates the need of a great deal of skill in the recognizing process and consequently RGB-D cameras must provide the best images possible in terms of quality, so that the facial authentication algorithms can minimize the inevitable false positives and false negatives, having as much features as possible retrievable from images and depth maps provided as input data.

A curious anecdote is that the spread of personal mobile devices equipped with RGB-D cameras has been the cause of increased usage of face to perform user authentication to such an extent that a new taxonomy has been forged, the selfie biometrics [62], an expression that refers to the term *selfie* commonly used to define a self-portrait digital photography.

Face Identification

In this article, face identification refers to that variety of applications performing face recognition without authentication purposes. Many fields are concerned in the field of face identification applications: among them it is inevitable to mention security, recently Sajjad et al. [63] have introduced a framework to support law enforcement agencies in suspects identification and missing people finding in the context of smart cities; marketing, to target specific customers or at least some of their features such as age and gender [64], and healthcare; Hossain et al. [65] have presented a framework for health monitoring through a comparison between the current status of a patient and an image of the same patient in good health.

Some applications have benefitted of the technology development in terms of portability; Elrefaei et al. [66] have proposed a client-server architecture to acquire a video on an Android smartphone and to run a face recognition algorithm on the server side to transfer the most computational expensive task from the smartphone to remote computers, some years before Driver et al. [67] have presented a work to identify patients into a hospital through mobile face recognition and successively to retrieve patient's medical data, while Raghavendra et al. [68] have proposed an approach to reconstruct the depth information belonging to a face recording a video with the frontal camera of a smartphone. Therefore, as result of these examples, the identification of a proper depth acquisition technology plays a key role in the development of face identification applications.

Face recognition is probably the most studied facial application ever and the breadth of this research area has forced to select the main aspects to draw a realistic picture of the requirements for the choice of the proper depth acquisition technology, starting from a taxonomy clarification [69], since it has changed considerably over the years due to the spread of different kind of algorithms [70].

Among the countless studies about face identification, Turk et al. [71] have approached the problem using eigenfaces, namely eigenvectors referring to a set of faces, He et al. [72] have followed an appearance-based approach called Laplacianface to minimize lighting, facial expression and pose variation, Ahonen et al. [73] have proposed a face representation based on local binary pattern (LBP), Wright et al. [74] have studied the automatic recognition of human face in various conditions such as different expressions, illumination, occlusions and disguises via sparse representation, predicting how much occlusions the algorithm could handle. More recently, Parkhi et al. [75] have published their work on face recognition using deep learning, and many other studies involving neural networks have been presented, for instance Wen et al. [76], Liu et al. [77] and Deng et al. [78] highlighting that, nowadays, this is the best performing approach.

Face recognition algorithms relying only on 2D RGB images must be carefully used stand-alone due to their vulnerabilities to spoofing attacks. Indeed, some other methods as liveness detection [79] must be added to obtain a reliable face recognition technique [80]. The technological improvement has made depth data usage promising since depth maps provide different kind of information compared to RGB images; nonetheless, 3D information can be integrated in the existing methods making algorithms more robust to spoofing attacks [81].

Face Expression Recognition

Face expression recognition aims to understand human emotions by observing different parts of the face. Paul Ekman studies have made the way for this discipline, indeed he studied the nervous system activity response to emotions [82], he defined a set of six basic emotions, which are fear, anger, joy, sadness, disgust and surprise highlighting that his intent was not to deny the variety of affective phenomena, but to attempt to organize those phenomena [83], he studied the numerous cross-cultural agreements, but also the differences in the judgments of facial expressions [84] and he defined the Facial Action Coding System (FACS) to map facial muscles movements in Action Units (AUs) and trace back the emotions [85].

Bartelett et al. [86] have presented a system to automatically detect faces, to use Gabor representation and then to process data through a bank of SVM classifiers, Shan et al. [87] have used Local Binary Pattern (LBP) to recognize salient micro-patterns on the face and to develop a low-computational method, further developed using Boosted-LBP features and comparing different machine learning algorithms, among them SVM [88], Kotsia et al. [89] have proposed a methodology using geometrical deformation features and SVM, Guo et al. [90] have studied the problematic case of a small number of images to train classifiers, obtaining that the best results in these conditions are achieved by SVM and their own Feature Selection via Linear Programming (FSLP). The issue related to obtain large and reliable database for face expressions will be detailed in Chapter 3 of the present thesis.

First attempts to use deep learning in face expression recognition dates back to more than twenty years ago, one of the main works has been conducted by Matsugu et al. [91], with limited different emotions recognizable. More recently, with a few exceptions such as Jingxin et al. [92], that have presented a method based on automatic landmark detection and AUs, researches have definitely adopted deep learning approach, which literature review will be detailed in Chapter 4.

The increasing interest in face expression recognition is strongly connected to human-computer interaction (HCI) [93] and involves a variety of fields: Small et al. [94] have studied how much emotional photographs in charity advertisements can evoke sympathy to engender giving, a smart solution applied to marketing, Lee et at. [95] have conducted an experiment involving forty users to study the relationship between emotions and choice of contents on smart TVs, McDuff et al. [96] have presented a toolkit to design user interfaces able to adapt to multiple users' expressions, which could be interesting in interactive applications such as videogames, Calvo et al. [97] have studied the connection between facial expressions and expresser's internal feelings in the context of psychiatry, Olivetti et al. [98] have evaluated the user's engagement in a virtual environment, suggesting that the methodology can be applied to other virtual products, while Nonis et al. [99] have evaluated the engagement of users focused on learning 3D modeling and printing using a 3D simulator. Another relevant application field is robotics, since the capability to automatically understand human's mood significantly improve human robot interaction in terms of communication efficiency and safety [100].

Face expression recognition is a critical task since some expressions, especially fear and sadness according to Alexandre et al. [101], are ambiguous and difficult to be recognized even by a human observer. For this reason, it is necessary to retrieve all the possible kind of information from the face on which the emotion has to be detected and, on a visual level, geometrical analysis can be considered the basis of face expression recognition, because the usage of the third dimension is essential to detect and describe facial movements in the most accurate way possible.

1.2.2 Desk research on RGB-D camera technologies

The interest in the applications mentioned above has received a further impulse since the advent of low-cost 3D sensors, i.e. devices able to detect the third dimension. The Microsoft Kinect release on the market in 2010 is one of the milestones related to the diffusion of these devices. This sensor has been designed and developed for the specific purpose of recognizing human body actions to perform an original type of human-machine interaction aimed at controlling characters, vehicles, or whatever object movements inside a videogame.

Several types of 3D sensors have been released on the market during last years and technology is the most suitable characteristic for grouping up sensors according to the similarity of their main parameters (Figure 1.2).

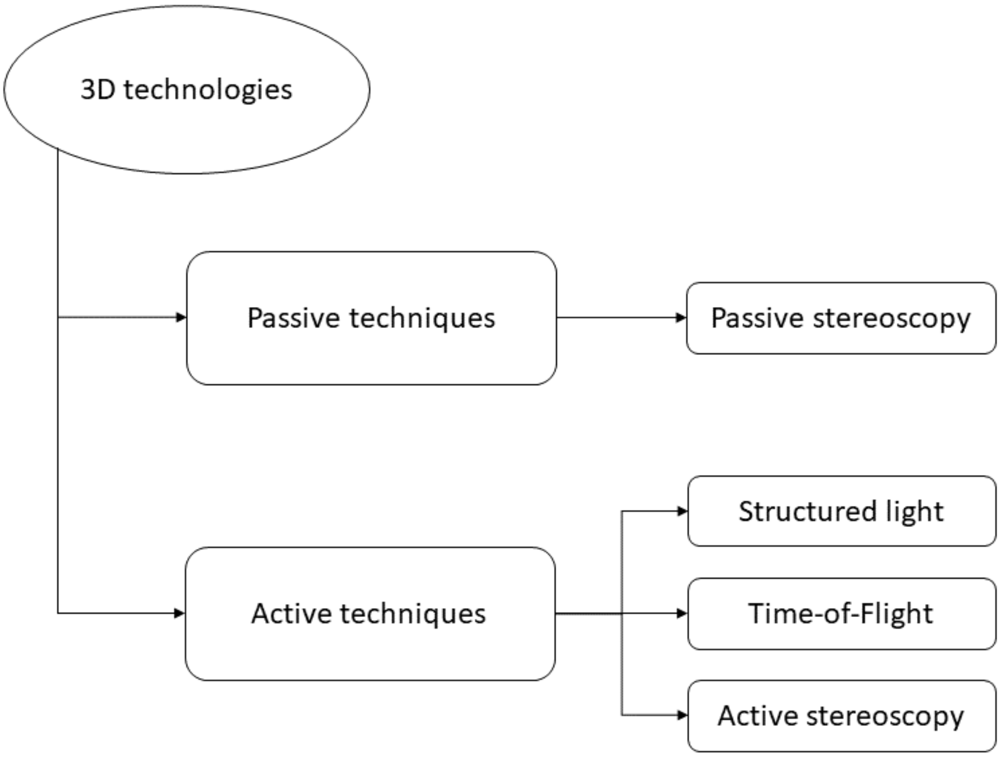


Figure 1.2 - 3D Technologies [70]

All the 3D sensors mentioned above are also known as RGB-D cameras, because they provide two types of data: RGB and D (depth). RGB refers to the color model thanks to which every color can be displayed using three primary color red, green and blue; in other words, it identifies the color images. Depth information is retrievable through depth maps, images on which each pixel has a value representing the distance from the camera. This type of data is an advance compared to 2D data in terms of reliability and is suitable for real-time applications, indeed it is possible to analyze the depth map without building a mesh; every 3D object is identified with x, y coordinates and the depth value instead of set of vertices, edges, and faces. The result is a more responsive acquisition system at the cost of accuracy. The present work focuses on 3D sensors because it is necessary to understand which technology can preserve high quality depth data working in real-time. It is mandatory to critically analyze which data will be adopted in the near future, when the accuracy of the third dimension will be exploited for several purposes and analyzing data real-time will be core for most of the acquisition systems [102].

Some of the applications mentioned above can have a considerable computational cost; nonetheless, 3D cameras and the devices that potentially can integrate them must be able to acquire information in real-time but the processing can be performed by systems located remotely. This solution can be planned at designed time before implementing a facial application, allowing not to be constrained by device capabilities in terms of processing, although they still must guarantee to maintain the 3D camera frame rate and to be connected with the remote system.

The way each depth acquisition technology provides the depth map is described in the following paragraphs.

Passive stereoscopy

Passive stereoscopy technology requires the presence of at least two cameras for acquiring different images of the same object or environment from different points of view.

To understand the distance of each point from the camera using this technology, the triangulation (or computational stereopsis) process must be performed, solving the so-called correspondence problem. Given the camera parameters calibration, the conjugate points [103], i.e. the two pixels representing the same point on the scene that are positioned on the two different acquired frames, must be found (Figure 1.3).

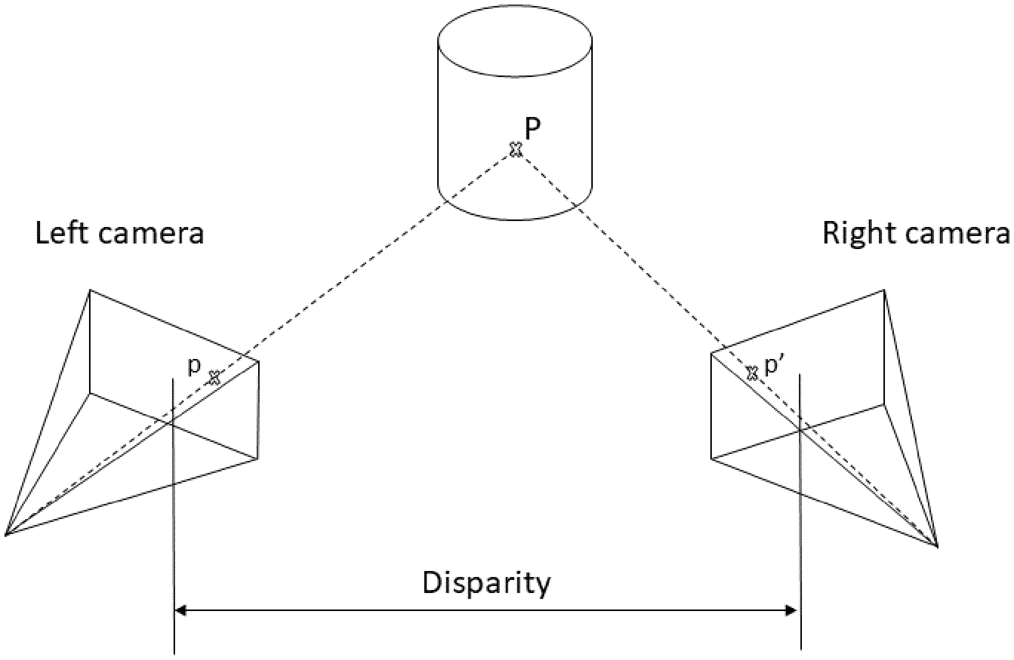


Figure 1.3 - Passive stereoscopy: conjugate points

The main drawback of stereo cameras is the need of a scene lacking occlusions, therefore the shape of the object can be detected from both the cameras, and this is not trivial, since the object geometry can be complex enough that some parts are visible from a camera and hidden to the other one, such as alae, namely the two points that lie on the right and on the left of the nose and are commonly considered the landmarks for computing nose width [104]. In addition, the scene must not be featureless since the correspondence problem can be solved only if the same features can be found by both the cameras.

Some cameras using passive stereoscopy are: Stereolabs ZED [105], Carnegie Robotics MultiSense S7 [106], E-Con Tara – Stereo Vision Camera [107], Nerian SPI [108], Roboception rc_visard [109], DUO 3D Sensor [110].

Price of these cameras can vary from 150 $ to 700 $ depending both on the features and the release on the market time.

Structured light

Structured light depth cameras have been designed to overcome the issue of reliability of correspondences that have to be identified in frames acquired by different cameras.

The technology consists in projecting a pattern on the object using a transmitter, typically an IR projector; successively, the deformation of the pattern on the object is detected by a receiver and sent to an application-specific integrated circuit

(ASIC). Data are processed and the result is the frame with depth information (Figure 1.4).

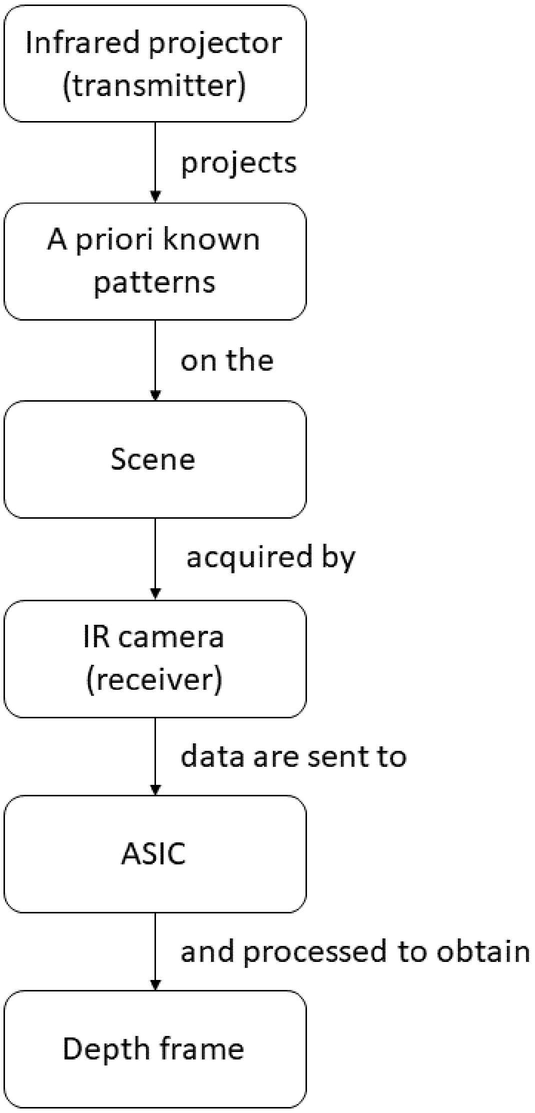


Figure 1.4 - Main steps involved of coded light technology

The projected pattern can assume different configurations to perform the correspondences estimation according to design concepts. Adopted strategies are wavelength multiplexing, range multiplexing, temporal multiplexing, and spatial multiplexing [111].

Wavelength multiplexing uses light emitted at different wavelengths but is has been recently discarded due to inter-channel crosstalk which makes this option useful only for scenes with limited illumination and albedo variation.

Range multiplexing uses binary (black and white) or grayscale light, but suffers the noise more than the other solutions, indeed different pixels of the projected texture have a different illumination power and these differences can be confused with the noise. Moreover, darker pixels are more difficult to be recognized by the

IR camera, making very complicated the computation of the furthest points on the scene.

Temporal multiplexing has been originally developed for still scenes. Patterns are black and white stripes built based on gray codes: firstly, the code word length N must be defined, then 2^N codework are generated. Each code word differs from the previous one only for one bit (one stripe). Finally, the code words are projected on the scene one after the other. The drawback of this solution is the presence of artifacts in dynamic scene, it is common to have a shadowed image of the object acquired due to the longer exposure time required.

Spatial multiplexing is the most widely used option to deal with patterns in coded light cameras. Patterns can be projected several times per frame, increasing accuracy of the final depth estimation. Patterns are very complex and aim to minimize the noise disturbance.

The different multiplexing solutions described above are not mutually exclusive. For instance, range multiplexing is often used together with the spatial multiplexing.

Structured light cameras allow to put transmitter and receiver close each other, since the distance is computed without the need of the disparity and consequently the occlusions issue is minimized, even if it is an issue not to underestimate in case color and depth frames must be aligned.

Some examples of cameras using structured light are Intel RealSense F200 [112], Intel RealSense SR300 [113], Microsoft Kinect v1 [114], Asus Xtion Pro Live [115], Ensenso N35-606-16-BL [116], Orbecc Astra Mini [117], Photoneo PhoXi [118], Structure Sensor [119].

This type of camera can be considered quite cheap compared to the other technologies: price is usually not higher than 200 $ with a few exceptions.

Time-of-flight

ToF cameras have been considered only at professional-grade level until Microsoft released the second version of the Kinect, commonly mentioned as Kinect v2 or Kinect One, since it has been developed for being used with the Microsoft X-Box One console, contrary to the Kinect v1 developed for X-Box 360.

This technology [120] relies on the knowledge of the light speed in the air ($c \approx 3 \times 10^8$ m/s). Distances can be evaluated projecting an electromagnetic wave on the scene and computing the time in which it has been received from the receiver (Figure 1.5).

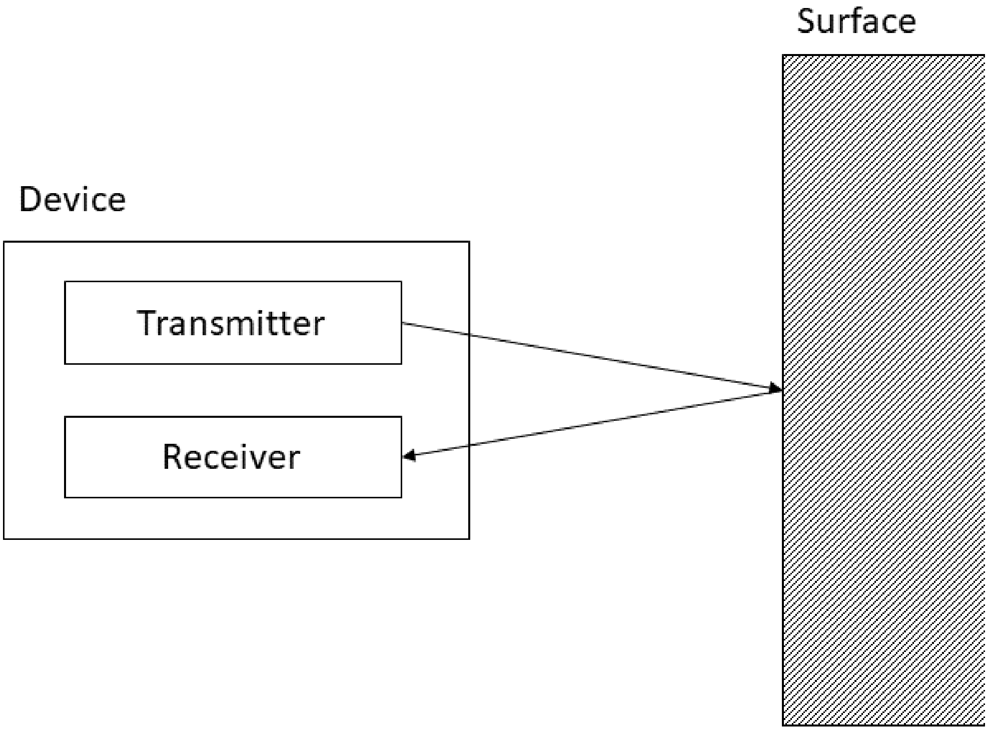


Figure 1.5 - ToF principle

The light emitted by the emitter is usually a square wave. Reading the shift between the transmitted wave and the received one it is possible to understand, integrating the two readings, the distance of the single points of the received image.

A remarkable advantage of this technology is the opportunity to put transmitter and receiver closer than the transmitter and the receiver needed for structured light depth cameras, making measurements occlusion-free. In this way, there are not artifacts on the acquired depth image; nonetheless, pixel matrix is usually smaller, and the resolution can be penalized.

Some examples of time-of-flight cameras are Microsoft Kinect v2 [121], IFM O3D303 [122], SICK Visionary-T [123], Basler ToF camera [124], PMD CamCube 3.0 [125], MESA SR 4000 [126], MESA SR 4500 [127], Soft Kinect DS325 [128].

On average, ToF cameras are the most expensive on the market since they were born for industrial applications. Nonetheless prices cover a very wide range: from 80 $ to thousands of dollars.

Active stereoscopy

Active stereoscopy is a depth acquisition technique which combines traditional passive stereoscopy with the emission of an IR pattern on the scene. The projected pattern increases the number of features on the scene (as if markers would be added), making the depth acquisition more accurate and more reliable; this solution is particularly useful to acquire flat surfaces of objects, subjects, and environments and to extend the operating range [129], in particular towards the shorter range, because at far distances passive stereoscopy suffers less from the occlusion problem.

An RGB-D camera built with this technology is equipped with two outdistanced cameras and a projector between them, usually working in IR spectrum.

Some active stereoscopy sensors are Intel RealSense D415 [130], Intel RealSense R200 [131], Intel RealSense D435 (Figure 1.6) [132] and Intel RealSense Euclid [133].

Active stereoscopy cameras are peculiar of Intel which initially proposed them at a cost between 130 $ and 400 $. Most recent devices cost 150 $ - 200 $.

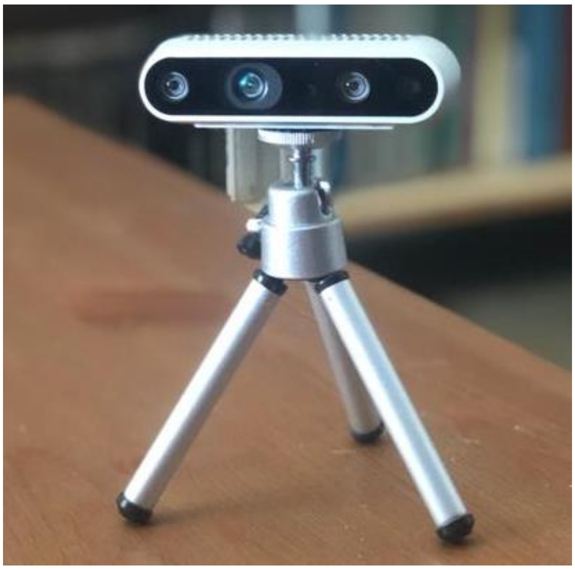


Figure 1.6 - Intel RealSense D435

1.2.3 Benchmarking

A benchmarking among 3D sensor technologies has been done evaluating the parameters available both in literature and in datasheets. Parameters taken into consideration are:

- Resolution: horizontal and vertical number of pixels
- Frame rate: number of images captured in one second (FPS, Frames Per Second)
- Minimum distance: this parameter establishes the lowest gap for camera functioning
- Maximum distance: this parameter establishes the greatest gap for camera functioning
- Range: difference between minimum distance and maximum distance
- Field of view (FOV): this parameter indicates the part of the scene visible by the camera
- Size: camera dimensions

Twenty-six sensors belonging to the four categories described above have been analyzed to identify strengths and weaknesses of each 3D detection technology. A recap is shown in Table 1.1.

Table 1.1 - RGB-D cameras considered in this work

Technology	Cameras
Passive stereoscopy	Stereolabs ZED Carnegie Robotics MultiSense S7 E-Con Tara – Stereo Vision Camera Nerian SPI Roboception rc_visard DUO 3D Sensor
Structured light	Intel RealSense F200 Intel RealSense SR300 Microsoft Kinect v1 Asus Xtion PRO LIVE Ensenso N35-606-16-BL Orbecc Astra Mini

	Photoneo PhoXi 3D Scanner L Structure Sensor
Time-of-Flight	Microsoft Kinect v2 IFM O3D303 SICK Visionary-T Basler ToF camera PMD CamCube 3.0 MESA SR 4000 MESA SR 4500 Soft Kinect DS325.
Active stereoscopy	Intel RealSense R200 Intel RealSense D415 Intel RealSense D435 (D435i) Intel RealSense Euclid.

Passive stereoscopy

Six passive stereo sensors have been considered (Table 1.2). Stereo cameras have quite good ranges of functioning, thanks to good maximum distance values that make most of them suitable for acquisition over 3 meters of distance, but a bad minimum distance of functioning. Values regarding minimum distance of functioning reported in this work, directly taken from sensor datasheets, are often misleading. That value means that it is possible to acquire the depth map, but its quality is very poor, especially in the case of facial application. This is a technological problem: passive stereoscopy uses disparity between two cameras to retrieve the depth information. If the camera is too close to the subject, a lot of points will be present in only one of the images due to occlusions, making the correspondence problem impossible to be solved. Resulting depth images contain too big holes, which make data impossible to use. In particular, a second minimum value is often shown in datasheets and it points out the optimal minimum distance that is usually greater than 50 cm.

On the contrary, resolution is excellent, while frame rate has quite different nominal values, from 3 to 60 FPS.

Table 1.2 - Passive stereoscopy sensors specs

Passive stereoscopy		
Parameter	**Best**	**Worst**
Frame rate	60 FPS	3 FPS
Maximum distance	15 m	1 m
Minimum distance	0.23 m	0.5 m
Range	14 m	0.3 m
Resolution	2208x1242	640x480
Size	57x30.5x14.7 mm	230x75x84 mm

Structured light

Eight structured light sensors have been analyzed (Table 1.3). Minimum distance is undoubtedly the strength for this technology, in fact several sensors minimum operating distance is between 20 cm and 40 cm. Frame rate is remarkable too, almost all sensors work at 30 or 60 FPS. Maximum distance and range operating functioning are the weaknesses of this technology, since most of sensors work with an upper limit that is suggested from 1.5 m to 2.5 m. Resolution is remarkable for short range, since only one sensor is 320x240 and the others are 640x480 or above.

Table 1.3 – Structured light sensor specs

	Structured light	
Parameter	**Best**	**Worst**
Frame rate	60 FPS	30 FPS
Maximum distance	5 m	0.49 m
Minimum distance	0.2 m	0.87 m
Range	4.4 m	0.23 m
Resolution	1280x1024	320x240
Size	80x20x20 mm	279.4x71.12x66.04 mm

Time-of-Flight

Among the 8 ToF sensors analyzed, just one of them can be considered suitable for facial applications. Other sensors belonging to this category have a magnificent maximum distance (at least greater than 4 meters), a decent frame rate (20-30 FPs), but poor minimum distance (0.5 m) and resolution (640x480 is the only remarkable value, all the others are below).

Some values in Table 1.4 are strongly influenced by a single sensor build with the specific purpose of working at close distance, the SoftKinetic DS325. This sensor's operating range is between 0.15 and 1 m, unexpected parameter compared to the other time-of-flight cameras, which usually have minimum distance between 0.3-0.5 m and maximum distance of functioning of several meters.

Table 1.4 – Time-of-Flight sensor specs

	Time-of-Flight	
Parameter	**Best**	**Worst**
Frame rate	30 FPS	20 FPS
Maximum distance	60 m	1 m
Minimum distance	0.15 m	0.8 m
Range	59.5 m	0.85 m
Resolution	640x480	176x144
Size	65x65x68 mm	250x70x45 mm

Active stereoscopy

The four active stereoscopy sensors considered (Table 1.5) are the most recent on the market, launched in 2015 or later.

They can be considered the best trade-off between all the parameters, with good minimum distance (around 30 cm), maximum distance (up to 10 meters), 30 FPS frame rate and good depth resolution (two of them reach 1280x800).

A special mention is deserved by the best minimum distance found during the desk research (0.11 m, belonging to Intel RealSense D435); nonetheless, minimum distance of functioning exceeds 0.3 m for the other cameras.

Table 1.5 - Active stereoscopy sensor specs

Active stereoscopy		
Parameter	**Best**	**Worst**
Frame rate	30 FPS	30 FPS
Maximum distance	10 m	2.8 m
Minimum distance	0.11 m	0.4 m
Range	10 m	2.25 m
Resolution	1280x800	640x480
Size	90x25x25 mm	126x254x345 mm

Sensors datasheets report the size including the chassis and the support dimensions. Customer-grade sensors can be integrated in personal devices such as smartphones, tablets and laptops without chassis and support, therefore it is desirable to understand the physical space that each technology requires. Passive stereo and active stereo need a larger space due to the presence of two different cameras for detecting the third dimensions through the disparity, while for what concerns structured light and ToF technologies size can be limited by the possibility of putting transmitter and receiver as close as possible.

A brief recap of main advantages and disadvantages for each technology can be found in Table 1.6.

Table 1.6 - Advantages and disadvantages of analyzed technologies

Technology	Advantages	Disadvantages
Passive stereoscopy	Good operative range Good resolution	Difficulties in discriminating between minimum distance and optimal minimum distance of functioning Scene not featureless Problems with occlusions

Structured light	Excellent minimum distance of functioning Good frame rate	Low maximum distance of functioning Low range of functioning
Time-of-Flight	Excellent maximum distance of functioning	Poor minimum distance of functioning Poor resolution
Active stereoscopy	Good minimum distance of functioning Good maximum distance of functioning	Small problems with occlusions (limited by the IR light projection)

1.2.4 Quality Function Deployment (QFD)

Once the desk research has been completed, the QFD has been used to integrate two orthogonal dimensions, namely sensors' technical specifications and facial applications requirements. The aim of this stage is to identify their interconnections evaluating how much each technical specification is important in relation to a certain application requirement.

QFD [29] is a method applied to transform qualitative user demands into quantitative parameters and the basic design to implement it is the house of quality. On the vertical axis there are the user desires (What's), on the horizontal axis there are technical requirements (How's) that may be useful to satisfy the user desires. A weight between 1 and 5 is given to each user's desire according to the final application that has to be designed. In the other cells of the table a score of 1, 3 or 9 [134] is given according to the contribution that each technical requirement gives to each user desire, namely respectively *weak*, *moderate* and *strong*. 0 value has been given if there is no relationship. Scores to be attributed to the relationships can vary according to different ways of building a QFD [135]. In this case, 0, 1, 3, and 9 have been considered suitable, because they reflect at best the perception that people have with regards to the correlation process and strong correlation is awarded.

Four QFDs have been drawn up, one for each facial application previously explained, and they are structured as follows: qualitative application requirements, namely the main characteristics that an application should have, are listed on the first column and the importance of each qualitative requirement is listed on the second column. On the first row there are the technical specifications (How's), and contrary to the qualitative requirements, that are slightly different between the applications, the technical requirement list is the same for each of the four QFDs.

The considered technical specifications are the depth sensors parameters extracted by the desk research. Specifically, technical requirements are the frame rate, the minimum and the maximum distance to which the sensors work, the range, the FOV and the dimensions.

In the final row the relative total score of each technical specification is specified. Relative total score is a percentage of how much a technical requirement is important compared to the others. Its values are computed as follows:

- For each technical requirement, a total score is computed as a sum of products between the application requirement weights and the corresponding evaluation scores given to the technical requirements.
- For each total score obtained at point 1 the percentage is computed considering the sum of all the total scores as 100%.

1.3 Results and discussion

Generical raw data have been translated into values to be put in QFDs (9-3-1-0 score) after a discussion held by a focus group. The focus group has proved to be essential to accurately evaluate technical requirements thanks to the involvement of researchers from several areas and is composed by eleven people, five women and six men:

- four of them are computer science engineers, and their research field involve computer vision and RGB-D cameras
- three are management engineers
- two are biomedical engineers, experts in face analysis
- one is an electronic engineer
- one is a mathematical engineer, whose competences involve facial feature extraction

The focus group also assigned weights and scores to each of the requirements as a result of a discussion among all participants, so that everyone has intervened in the debate giving a contribution linked to the specific area of expertise, and the final value has been unanimously assigned.

Results are presented in the following section.

1.3.1 Face Detection

Even if accuracy is something to be taken care of in all contexts, this constraint can be considered not so strict for face detection stand-alone applications as for other facial applications. Once that the face is detectable, details on facial surface are not required. This does not mean accuracy is not relevant at all: a trade-off between accuracy and resources (computational and storage resources) is always necessary; nonetheless, in face detection applications the limit can be set closer to the resources than face authentication, face identification and face expression recognition applications. Moreover, flexibility should be a strength point for this application, so that it can work in all range, light, pose and occlusions situations (Table 1.7). Qualitative requirements are:

- Real-time: faces should be detected when an individual enters in the camera FOV [136].

- Wide operating range: faces should be detectable both if an individual is getting closer to the camera and moving away.

- Accurate at close distance: faces should be detectable if an individual is close to the camera.

- Accurate at far distance: faces should be detectable if an individual is far from the camera.

- Able to discriminate faces among other elements in the environment: the core of the application, if a face is present in the scene, then it should be detected.

- Integrable into a smartphone: sensors should allow to be put into a smartphone, a tablet, or a laptop to perform face detection.

- Portable: this requirement suggests having a sensor small enough to be easily carried by the user.

- Small output data: the detected face should be reported without spending too many resources in terms of memory, for reasons of storage and computational speed. Nonetheless, to preserve a level of accuracy that allows to detect faces is mandatory.

- Robust to light: faces should be detected whatever light conditions are (i.e. in the dark, in a sunny day…).

- Head pose invariant: faces should be detected whatever the individual relative orientation with respect to the camera is.

- Robust to occlusions: faces should be detected in presence of occlusions (i.e. glasses, scarves, …).

Table 1.7 - Face Detection

Application requirements	Application requirements weights	Technical requirements						
		Resolution	**FPS**	**Range**	**Min distance**	**Max distance**	**FOV**	**Dimensions**
Real-time	5	9	9	0	0	0	3	0
Wide operating range	5	0	0	9	1	9	0	0
Accurate at close distance	3	9	3	1	9	0	0	0
Accurate at far distance	3	9	1	3	0	9	0	0
Able to discriminate faces among other elements in the environment	5	9	0	3	3	9	1	0
Integrable into a smartphone	4	1	0	3	3	3	3	9
Portable	3	0	0	0	0	0	0	9
Small output data	5	9	9	0	0	0	3	0
Robust to light	4	3	1	3	3	3	3	0
Head pose invariant	5	3	1	3	3	3	0	0
Robust to occlusions	5	3	3	1	1	1	0	0

Relative total score	27%	15%	14%	11%	19%	7%	7%

Sensors parameters relative importance is shown in Figure 1.7.

Radar shows that the resolution is the most important parameter, followed by the maximum distance of functioning, since face detection applications must detect subjects that do not necessarily position themselves in front of the camera.

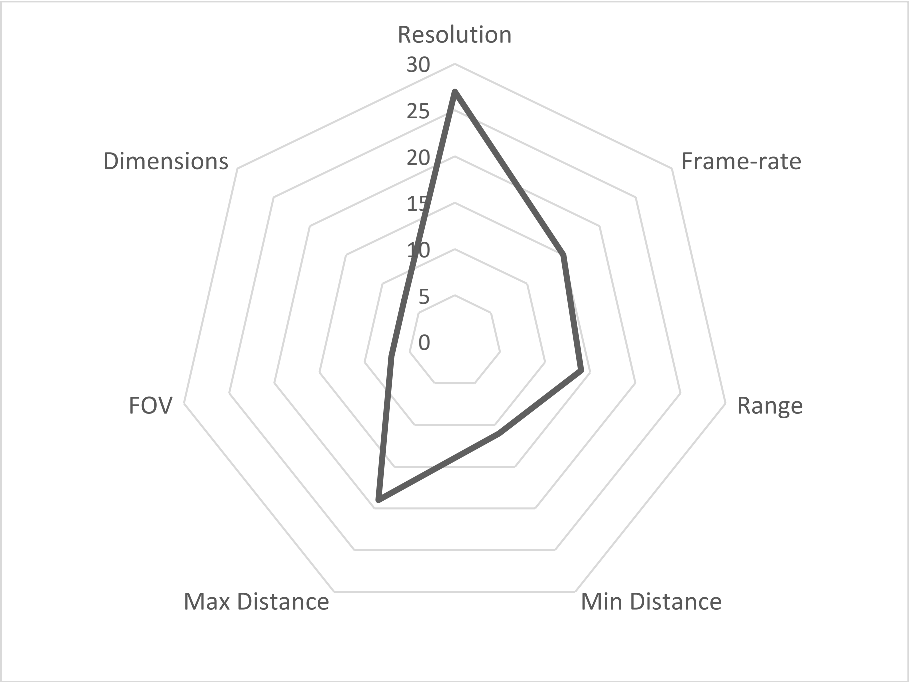


Figure 1.7 - Specs relative importance for Face Detection

1.3.2 Face Authentication

The minimum error rate in face authentication is required. User is aware of the sensitivity of this application so that real-time is not strictly required, but speed should be high enough to compete with other type of authentication (for instance, the insertion of a PIN code); nevertheless, speed must not sacrifice accuracy in any way, since for face authentication this is the main requirement on which to focus on. (Table 1.8). Qualitative requirements are:

- Fast enough to unblock a device: this application does not require real-time, unblocking speed should not be annoying for the user.

- Accurate at close distance: face should be recognized from a distance as close as a smartphone, a tablet or a laptop typical user is.

- Able to detect facial features: facial landmark for face analysis must be detected.

- Integrable into a smartphone: sensors should allow to be put into a smartphone, a tablet, or a laptop to perform face authentication.

- Robust to light: faces should be recognized whatever light conditions are (i.e. in the dark, in a sunny day…).

- Robust to occlusions: faces should be detected in presence of small occlusions (i.e. glasses, scarves, …).

Table 1.8 - Face Authentication

Application requirements	Application requirements weights	Technical requirements						
		Resolution	**FPS**	**Range**	**Min distance**	**Max distance**	**FOV**	**Dimensions**
Fast enough to unblock a device	5	9	9	1	3	3	3	0
Accurate at close distance	5	9	1	0	9	0	0	0
Able to detect facial features	5	9	0	3	9	3	0	0
Integrable into a smartphone	4	1	0	1	9	1	3	9
Robust to light	4	3	1	3	3	3	3	0
Robust to occlusions	3	3	3	1	1	1	0	0
Relative total score		28%	11%	6%	28%	9%	7%	11%

Sensors parameters relative importance is shown in Figure 1.8.

Radar shows that resolution and minimum distance of functioning are the most important technical requirements to satisfy, coherently with the most-common usage scenarios: a user that must unlock his personal device. Subsequently, frame rate and dimensions can be considered influential, since a user must not wait too much time to be authenticated, otherwise another authentication method would be preferable, and the system should have the possibility of being integrated in personal devices such as smartphones, tablets and laptops.

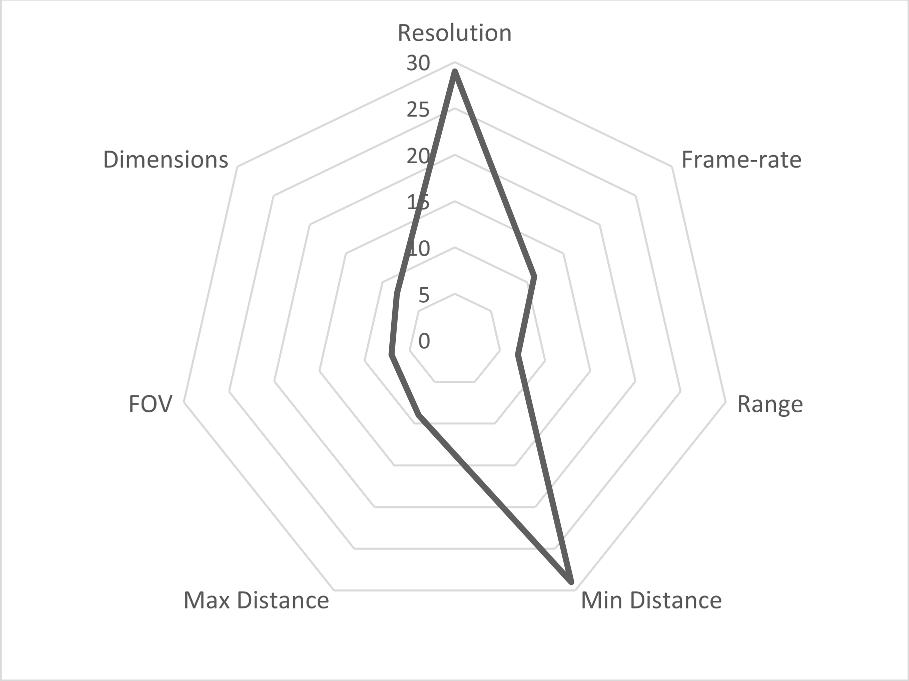


Figure 1.8 - Specs relative importance for Face Authentication

1.3.3 Face Identification

This application requires to council the accuracy for face analysis and the robustness to work in different range, light, pose and occlusions situation (Table 1.9). Close distance is not considered so relevant since face identification is different from face authentication as it has been previously explained.
Qualitative requirements about face identification are:

- Real-time: a subject should be identified before he leaves the field-of-view of the camera [137].

- Wide operating range: faces should be identified both if an individual is getting closer to the camera and moving away.

- Accurate at close distance: faces should be identified if an individual is close to the camera.

- Accurate at far distance: faces should be identified if an individual is far from the camera

- Able to detect facial features: facial landmarks for face analysis must be identified.

- Integrable into a smartphone: sensors should allow to be put into a smartphone, a tablet, or a laptop to perform face identification.

- Portable: this requirement suggests having a sensor small enough to be easily carried by the user.

- Robust to light: faces should be identified whatever light conditions are (i.e. in the dark, in a sunny day…).

- Head pose invariant: faces should be identified whatever the individual relative orientation with respect to the camera is.

- Robust to occlusions: faces should be identified in presence of occlusions (i.e. glasses, scarves…).

- Robust to different face expressions: faces should be identified whatever the individual mood is.

Table 1.9 - Face Identification

Application requirements	**Application requirements weights**	**Technical requirements**						
		Resolution	**FPS**	**Range**	**Min distance**	**Max distance**	**FOV**	**Dimensions**
Real-time	4	9	9	0	0	0	3	0
Wide operating range	5	0	0	9	1	9	0	0
Accurate at close distance	3	9	3	1	9	0	0	0
Accurate at far distance	4	9	1	3	0	9	0	0
Able to detect facial features	5	9	0	9	9	9	1	0
Integrable into a smartphone	3	1	0	3	3	3	3	9
Portable	2	0	0	0	0	0	0	9
Robust to light	5	3	1	3	3	3	3	0
Head pose invariant	5	3	1	3	3	3	0	0
Robust to occlusions	5	3	3	1	1	1	0	0
Robust to different face expressions	5	9	3	0	1	1	0	0

Relative total score	28%	10%	17%	15%	20%	5%	5%

Sensors parameters relative importance is shown in Figure 1.9.

Radar shows that the resolution confirms to be the most important technical requirement, indeed, to recognize features is mandatory to apply facial algorithms. All the technical requirements linked to the distance of functioning appears right after resolution in the ranking, since the sensor should be able to recognize subjects that could be more or less close to the camera.

This result is significantly different from face authentication and confirms the choice of splitting face recognition applications in face authentication and face identification.

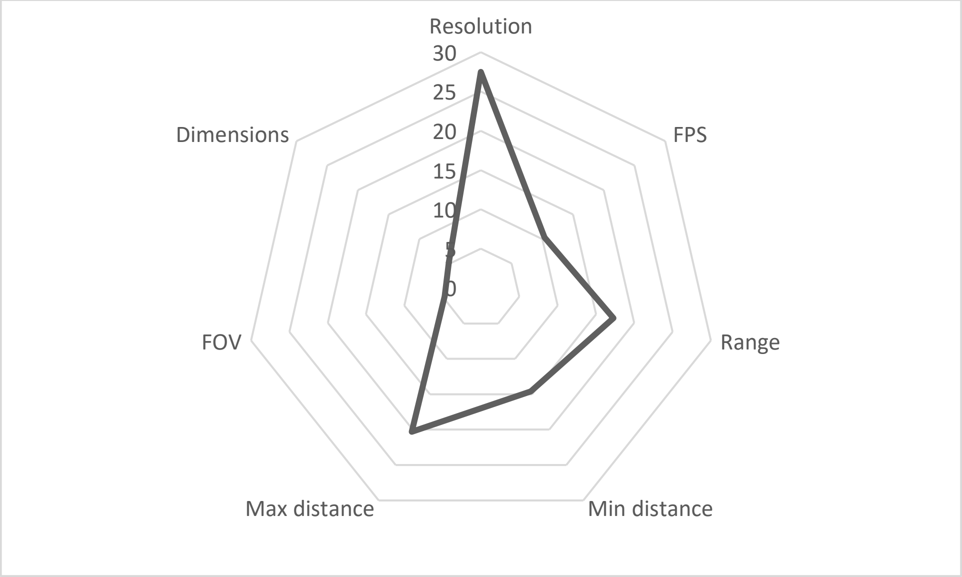


Figure 1.9 - Specs relative importance for Face Identification

1.3.4 Face Expression Recognition

Qualitative requirements about face expression recognition are very similar to the face identification ones since the operating conditions are almost the same (Table 1.10):

- Real-time: individual expressions should be recognized whenever an event associated to what they are assisting is triggered [138].
- Wide operating range: individuals' expressions should be recognized both if an individual is getting closer to the camera and moving away.
- Accurate at close distance: individuals' expressions should be recognized if an individual is close to the camera.
- Accurate at far distance: individuals' expressions should be recognized if an individual is far from the camera.
- Able to detect facial features: facial landmarks for face analysis must be recognized.
- Integrable into a smartphone: sensors should allow to be put into a smartphone, a tablet, or a laptop to perform face expression recognition.
- Portable: this requirement suggests having a sensor small enough to be easily carried by the user.
- Robust to light: individuals' expressions should be recognized whatever light conditions are (i.e. in the dark, in a sunny day…).
- Head pose invariant: individuals' expressions should be recognized whatever the individual relative orientation with respect to the camera is.
- Robust to occlusions: individuals' expressions should be recognized in presence of occlusions (i.e. glasses, scarves…).

Table 1.10 - Face Expression Recognition

Application requirements	Application requirements weights	Technical requirements						
		Resolution	**FPS**	**Range**	**Min distance**	**Max distance**	**FOV**	**Dimensions**
Real-time	5	9	9	0	0	0	3	0
Wide operating range	4	0	0	9	1	9	0	0
Accurate at close distance	4	9	3	1	9	0	0	0
Accurate at far distance	4	9	1	3	0	9	0	0
Able to detect facial features	5	9	0	3	9	9	1	0
Integrable into a smartphone	4	1	0	3	3	3	3	9
Portable	3	0	0	0	0	0	0	9
Robust to light	4	3	1	3	3	3	3	0
Head pose invariant	4	3	1	3	3	3	0	0
Robust to occlusions	4	3	3	1	1	1	0	0
Relative total score		26%	10%	14%	16%	20%	6%	8%

Sensors parameters relative importance is shown in Figure 1.10.

The radar appears to be very similar to the face identification one, but this result should not be surprising. In both cases resolution must be excellent in order to discriminate between different features on resulting images. Data should be retrievable both if the subjects is close or far from the camera, and, regarding the frame rate, data should be available several times per second (this requirement is satisfied by the vast majority of analyzed sensors). Finally, dimensions and field of view are not so much considered, because sensors should be not necessarily portable and can be placed in strategic locations in order to avoid FOV issues.

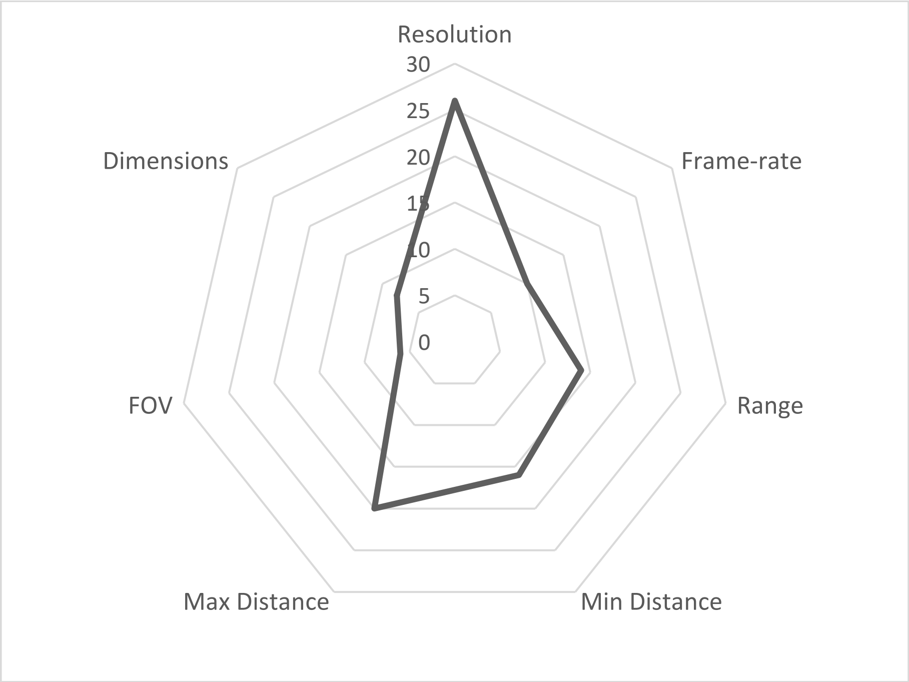


Figure 1.10 - Specs relative importance for Face Expression Recognition

A comparison between facial application specs is reported in Figure 1.11. Supplementing the comments already reported, resolution can be universally recognized as the most important parameter, followed by technical requirements linked to the distance of functioning, minimum, maximum and range, depending on the facial application. Frame rate varies from 10% to 15% and this result can be explained as follows: nowadays real-time is a mandatory requirements for facial applications; nonetheless, the bottleneck is not the choice of the sensor, but the computationally demanding techniques, thus the focus to solve issues linked to time performances must be moved on the choice of the suitable facial algorithm.

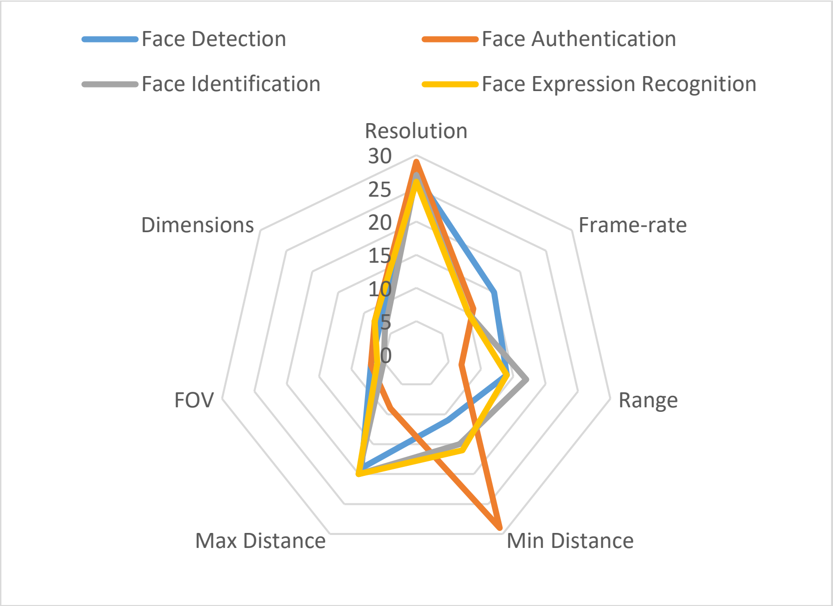


Figure 1.11 - Specs comparison for the different facial applications

Afterwards technical specifications and facial applications have been analyzed, the most suitable 3D detection technology can now be identified (Table 1.11).

ToF cameras are the best in terms of long range operating functioning [139] [140], but this strength is not feasible for facial applications, and they are weaker than other technologies in terms of resolution, this is the reason why it is the worst choice for the considered facial applications. In Figure 1.12 an example of 3D shell obtained from a depth map acquired using Microsoft Kinect One (ToF depth acquisition technology) is shown. Facial shape is recognizable, but facial surface is compromised, due to artifacts that strongly influence the facial surface.

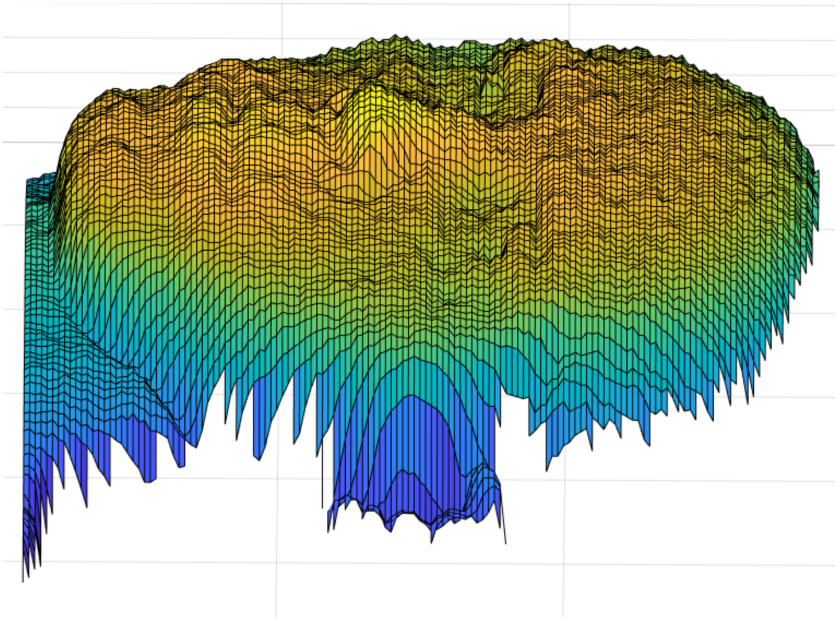


Figure 1.12 - 3D shell obtained with Kinect v2

Passive stereo technology has resulted to be the most suitable choice for face detection applications, due to the trade-off between high resolution and remarkable maximum operating functioning distance [141] [142], followed by the active stereo technology and, in third position, by structured light cameras, because of their too poor maximum operating functioning distance.

Face detection has been taken into account during all the evaluation process not only as stand-alone application, but also as preliminary step of face authentication, face identification and face expression recognition.

Scores of these facial applications have been given from a global point of view. In particular, when the focus group gathered for the evaluation, the main facial application steps were taken into account and this means that they discussed about the face detection step as well as the subsequent steps such as feature extraction or neural networks constraints in terms of input data.

Face detection requirements in stand-alone applications are different from face detection requirements as preliminary step. The requirements of face authentication, face identification and face expression recognition definitely consider face detection as a part of their algorithms, but some of the requirements may change based on the application within which they are incorporated. Going into the detail, if included in a face authentication application, face detection can accept a higher response time than face detection as a stand-alone application such as counting people in a room; furthermore, the operational range need not be wide, because face authentication use cases are limited to short distances.

Considering face identification and face expression recognition, the shape of the radars related to these applications are very similar each other and the face detection one is not too different. This testify that face detection requirements played a role in the evaluation of these facial application requirements. Indeed, they have not been twisted if they are considered as stand-alone or integrated, nonetheless there are some differences. In terms of relative importance, frame rate has a greater value in stand-alone face detection applications, while minimum distance acquires importance in face identification and face expression recognition.

The situation is inverted in face authentication. Since minimum distance is the most important parameter, together with the resolution, the excellent minimum operating functioning distance of structured light technology has resulted to be the best for this application [143] [144]. It is mandatory to observe that an active stereoscopy sensor seems to be the best at close range, but this is false to a broader set of sensors. Since active stereoscopy is the most recent technology, it is wise to bear in mind this result, but the time is not yet ripe to claim that it is the best one for close-range applications and, consequently, face authentication.

Face identification and face expression recognition have resulted to be similar in terms of qualitative requirements, in fact the shapes of their technical specifications relative importance are very close to each other. Active stereoscopy is the most suitable technology for these applications [145], because of the presence of good resolution both at close distance and long distance operating functioning at the same time. Passive stereoscopy is the second-best choice, thanks to its very high resolution and operating functioning at high distance, that is more relevant with respect to close distance. This is the reason why structured light is in third position, in fact the poor maximum operating function distance has been penalizing for this sensor category.

Key technical specifications used to analyze 3D sensor technologies are strongly linked to accuracy more than the acquisition time. From the datasheet analysis, it has been found that all the considered 3D sensors can provide several FPS when acquiring single shot acquisitions; if all of them can satisfy the real-time requirement, it has been unavoidable to focus on other technical specifications to discriminate between RGB-D cameras and to evaluate 3D acquisition technologies.

Table 1.11 - Technology ranking for facial application

	Face Detection	**Face Authentication**	**Face Identification**	**Face Expression Recognition**
1st	Passive stereoscopy	Structured light	Active stereoscopy	Active stereoscopy
2nd	Active stereoscopy	Active stereoscopy	Passive stereoscopy	Passive stereoscopy
3rd	Structured light	Passive stereoscopy	Structured light	Structured light
4th	ToF	ToF	ToF	ToF

1.4 Conclusions

In this chapter a survey to understand which depth acquisition technology can fit better different facial applications has been conducted.

Qualitative requirements for the most common face applications and RGB-D camera specifications considered in the present survey are the result of a literature review about 3D facial applications and a desk research among various depth acquisition technologies adopted by 3D sensors.

A focus group has filled-in four QFDs to identify the main features involved in each facial application and to understand which the most suitable technology for depth acquisition is.

Results show that passive stereoscopy is the best technology choice for face detection applications due to the great resolution and long operational range; structured light is the most suitable sensor technology for face authentication and, in general, for short-range applications since it does not need different points of view to acquire the depth information, making it easier to locate the sensor close to the subject/object of interest; active stereoscopy is the most interesting technology for face recognition and face expression recognition. This technology is the most recent and care must be taken to its development, nonetheless, active stereoscopy can already be considered the most versatile depth acquisition technology available up to now.

This part of the research has been the starting point to orient within the world of RGB-D cameras and facial application methodologies. Then, the choice of the proper 3D sensor has been evaluated analyzing the individual technical specifications of the considered models and the specific needs of the facial application to be developed for this thesis.

The choice has been fallen on Intel RealSense SR300.

Chapter 2

Face Expression Recognition on depth maps: a case study

Third dimension is an extremely powerful source of information, since it can face issues such as lighting variations, different pose orientations, disguises, and occlusions.

In the context of facial applications these advantages have become even more evident: a camera sensitive to the light could impede its user to access to his personal device when he is outdoor, in presence of very bright lamps or inside too dark rooms; a face recognition system could fail in front of a subject with makeup, as well as with a scarf put on, or even with a mask covering mouth and nose; a face expression recognition system could not work if the subject is not exactly aligned with the camera lens.

The usage of depth information can help to solve these criticalities and the key element to analyze the third dimension using an RGB-D camera is the depth map. The depth map is the counterpart of the color frame, indeed it is an image on which each pixel represents the distance of a point in the real world from the camera, instead of the color information, and its characteristics differ on the basis of the depth acquisition technology adopted.

In this case-study, a structured light camera has been chosen to properly acquire the depth information; indeed, the application has required to record the facial expressions of a subject during an interview. This means that the main parameter to choose the most suitable RGB-D camera technology for this purpose has been the minimum operational distance, since the RGB-D camera had to provide the best performances at short-range distances.

A face expression recognition methodology has been applied; the focus has been moved on the level of activation of the emotions rather that the identification of the emotions themselves, to monitor the engagement of the interviewee in every question.

2.1 Introduction

Face expression recognition is involved in an increasing number of application fields such as safety, entertainment, security. Among these, *emotional engineering* is a recent discipline that has introduced methods for predicting and likely being able to control users' emotional responses with respect to product attributes, in order to be able to design and engineer them [146].

The essential part of the design process is the capability of understanding the user's feelings and emotions [147]. Within this context, a case-study has been set up to touch the considerations made in the previous chapter and to start to integrate the technology potentialities of an RGB-D camera with a face expression recognition method.

In emotional design, a subject is usually submitted to an interview, which is an ideal situation to acquire depth data, given the statics of the scene; indeed, the interviewee is seated in the same position, except from little adjustments that can be faced in the post-processing phase.

Support Vector Machine is one of the most used classifiers for face expression recognition; furthermore, it is particularly suitable for circumstances in which the dataset has a low cardinality, but data are characterized by high dimensionality, so it has been chosen to pursue the goal of this chapter.

2.2 Method

The reason why to involve emotions through facial expression monitoring and evaluation has been to measure the emotional engagement during the interview.

A model has been searched in the literature for matching quantitative values to specific emotions, especially in the perspective of user's involvement. Russell proposed an emotional model, called *circumplex model* and shown in Figure 2.1, that can be discretized as it is placed on a Cartesian plane [148] [149]. According to this model, the x-axis quantifies the positivity/negativity of the emotion, while the y-axis quantifies the emotional activation and involvement.

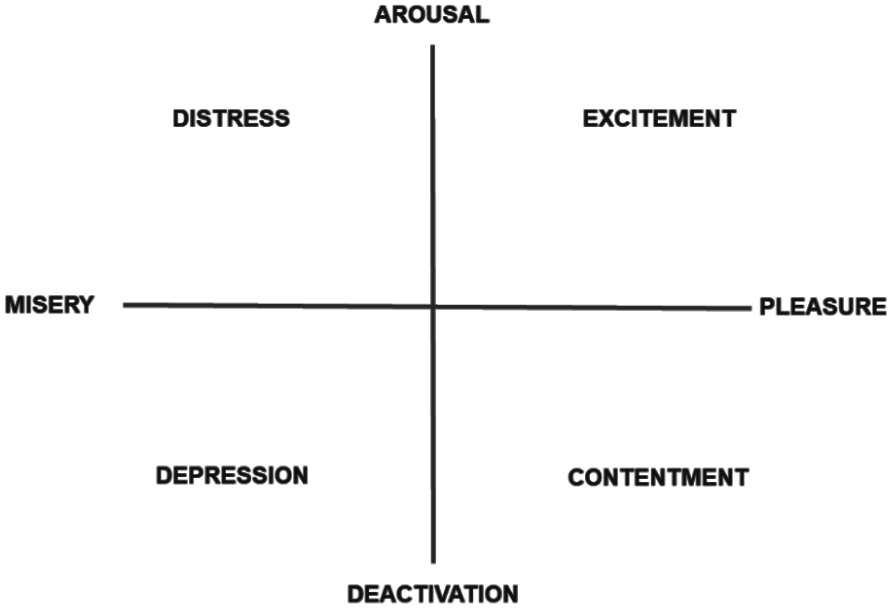


Figure 2.1 - The eight affect concept elaborated by Russel [148]

Considering the canonical emotional tone of a professional interview regarding a product to be conceptualized, we have considered for our model only the quadrants identifying positive emotions, i.e. the first and the fourth. Weights 1, 2, 3, 4, 5 have respectively been assigned to emotions *deactivation*, *contentment*, *pleasure*, *excitement*, and *arousal*, according to the circumplex model.

We also took into consideration the simplified model with weights 1, 3, 5 assigned respectively to *deactivation*, *pleasure*, and *arousal*. Due to the experimental nature of this study, the simplified model has been preferred, in order to point out differences among the classes more clearly.

For the sake of completeness, the two models are shown in Figure 2.2.

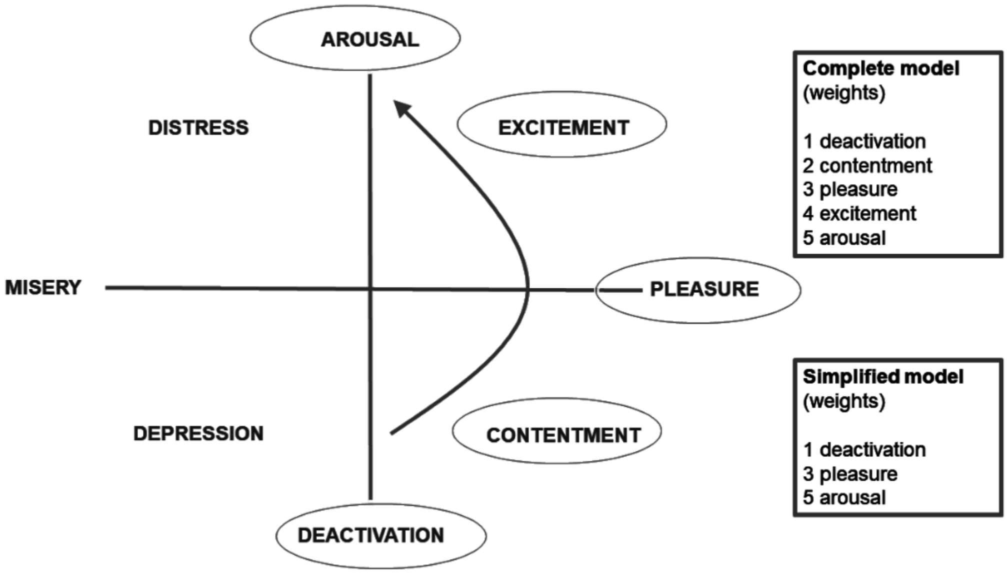


Figure 2.2 - Chosen emotions from Russell's model

During the interview, a depth camera should be placed in front of him/her to acquire the face frame-by-frame. The camera is started when the interview starts, together with a vocal recording. Then, the vocal recording is analyzed, and notes should be taken about the exact range of seconds in which the interviewer's question is about to finish and the interviewee starts answering. This is supposed to be the very moment in which the expression about the inner emotion (the *ground truth feeling*) is displayed by the face.

Facial data undergo a selection of the significant frames, which are then post-processed so that a depth map remains framing the face alone. Then, an algorithm is run on the facial depth map to automatically localize 17 landmarks (Nasion, Pronasale, Subnasale, left and right Endocanthion, left and right Exocanthion, left and right Inner Eyebrow, left and right Outer Eyebrow, right and left Alare, left and right Chelion, Labiale Superius, Labiale Inferius), shown in

Figure 2.3, with a thresholding methodology [150] based on Differential Geometry descriptors [151].

An insight about geometrical descriptors is provided in Appendix A.

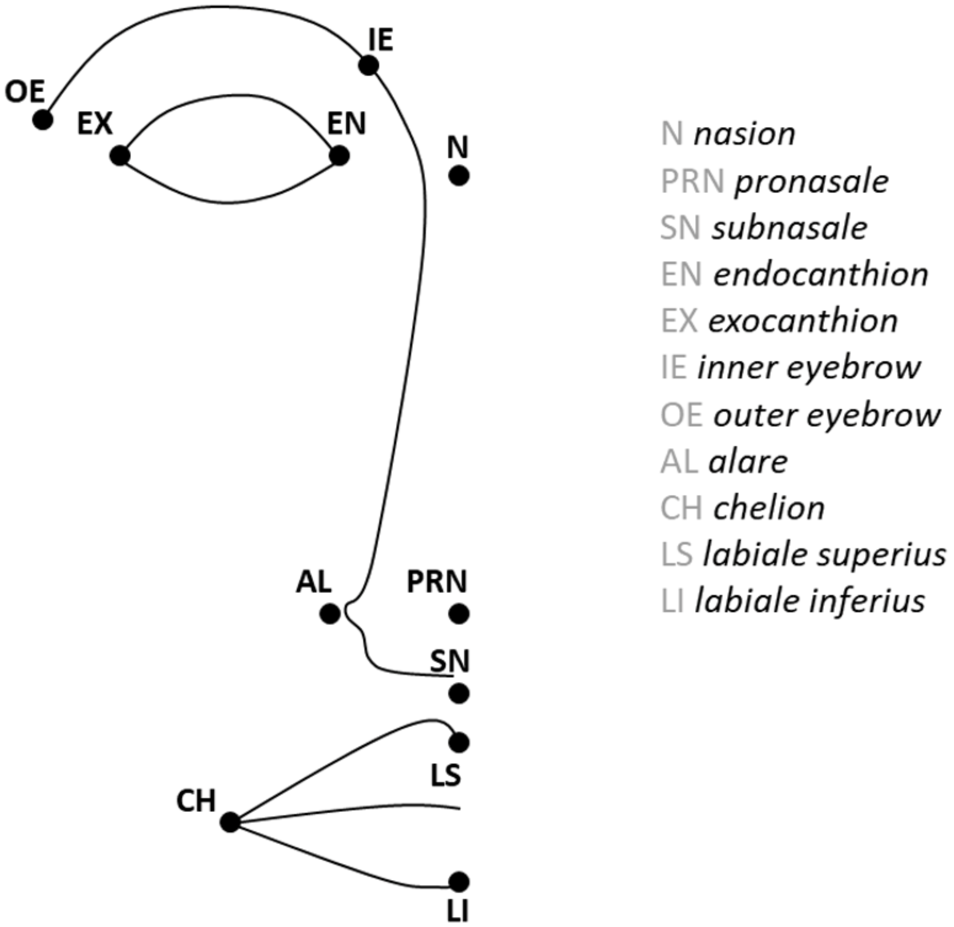


Figure 2.3 - Set of anatomical landmarks adopted in this study. The formal definitions of every landmark are provided by Swennen et al. [152], with the exception of the eyebrow points.

After the landmarking localization process, Euclidean distances are calculated between landmarks by adopting the *Pronasale* as anchor point, as shown in Figure 2.3. Thus, distances between every landmark and the *Pronasale* have been computed (Figure 2.4).

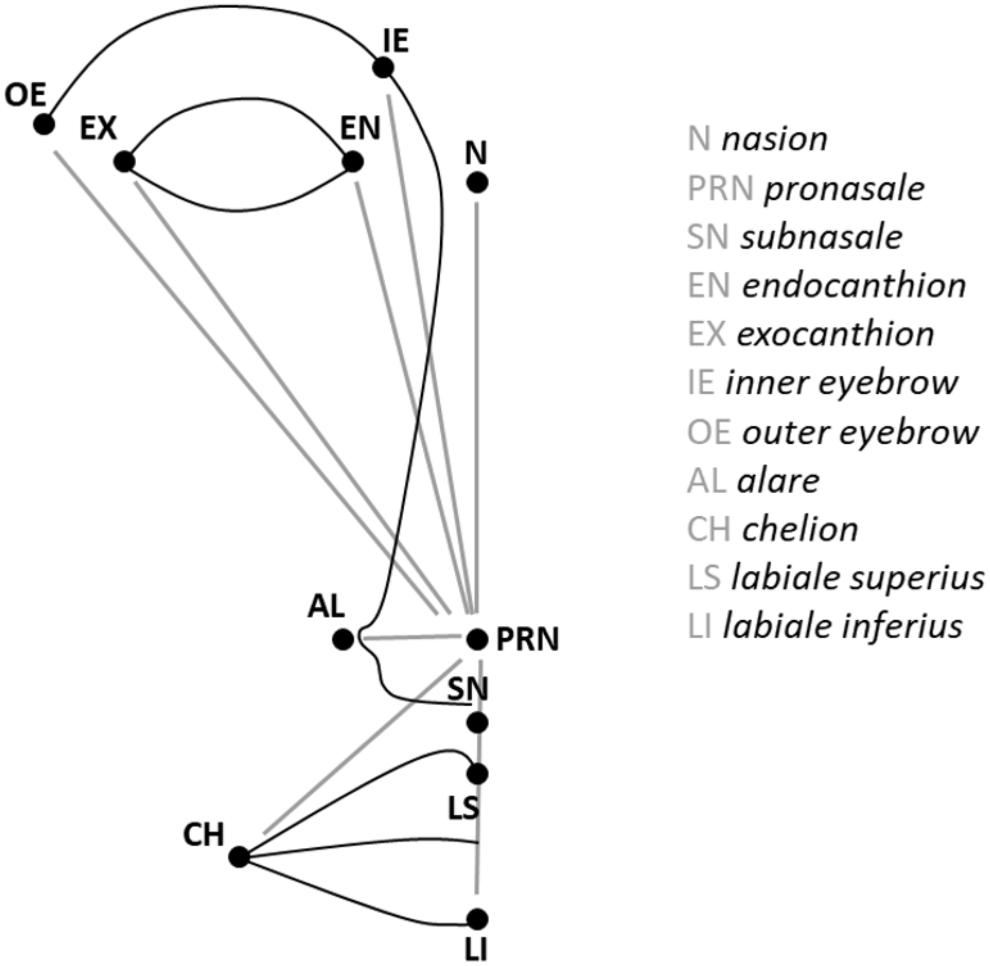


Figure 2.4 - Computed Euclidean distances using PRN as anchor point

Support Vector Machine (SVM) has been chosen as face expression recognition method, due to the high dimensionality of the data and the low cardinality of the dataset.

The classification is obtained by providing in advance a number of clusters in which to funnel the data. These clusters represent the activation levels of emotions; thus, in this case there are the 3 classes defined by the simplified model, which map the deactivation, the pleasure and the arousal levels.

SVM is a classification technique, as such the goal is to find a separator between classes, a problem that possibly can have infinite solutions (Figure 2.5).

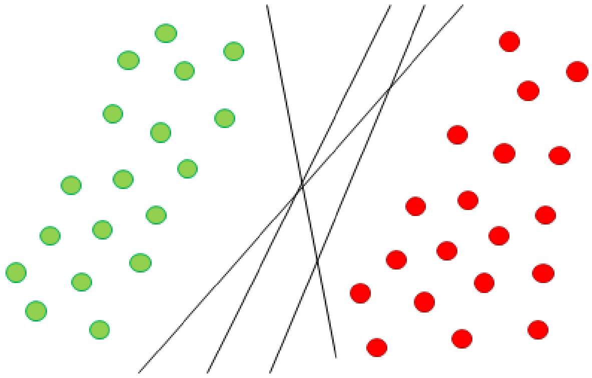


Figure 2.5 - Possible classes separators

SVM aims to find the so-called hyperplane, represented by the central continuous black line, that separates two classes with the widest margin possible, identified by the two dashed lines. This is a constrained optimization problem since no support vectors, namely no members of the classes, can be within the margin (constrain) and the margin must be as wide as possible (optimization).

If each support vector is located on the same side of the separator, the approach is called large-margin (Figure 2.6).

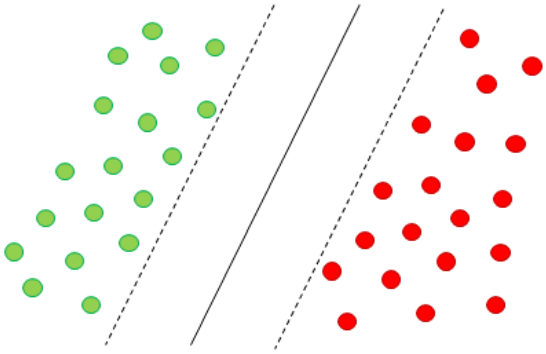


Figure 2.6 - SVM large-margin

If some exceptions are allowed, thus some members of a class can be on the wrong side of the separator (Figure 2.7), the approach is called soft-margin.

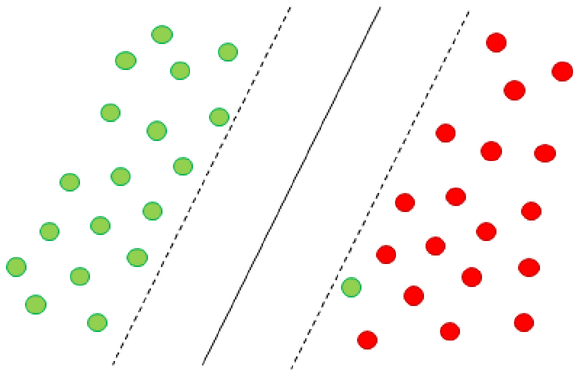


Figure 2.7 - SVM soft-margin

Large-margin and soft-margin are not a binary choice, but a tradeoff to find using the C parameter: the higher C is, the more accurate the hyperplane will split the model; the lower C is, the more importance will be given to the creation of a wide margin. In the first case the priority is making a few mistakes during the training phase, but this could be dangerous in case of introduction of a new support vector: indeed, it could influence the location of the hyperplane, making old support vectors misclassified due to the few tolerance given by the narrow margin. In the second case the priority is the simplicity, the large margin provides more flexibility, but less accuracy.

Furthermore, there are some cases, such as Figure 2.8, for which it is impossible to separate two classes using a straight line (linear kernel) and it is necessary to use a kernel trick, i.e. a non-linear kernel.

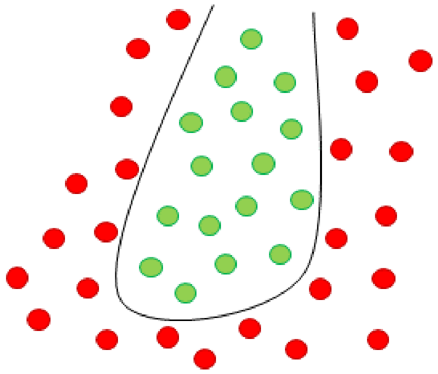


Figure 2.8 - Data distribution that requires non-linear kernel

One of the most used non-linear kernels is the RFB (Radial Basis Function) that has two parameters to set:

- *C*: it is a parameter that influences the choice between the large-margin and the soft-margin approach.

- *gamma*: it is a parameter that influences the curvature of the line separating the data

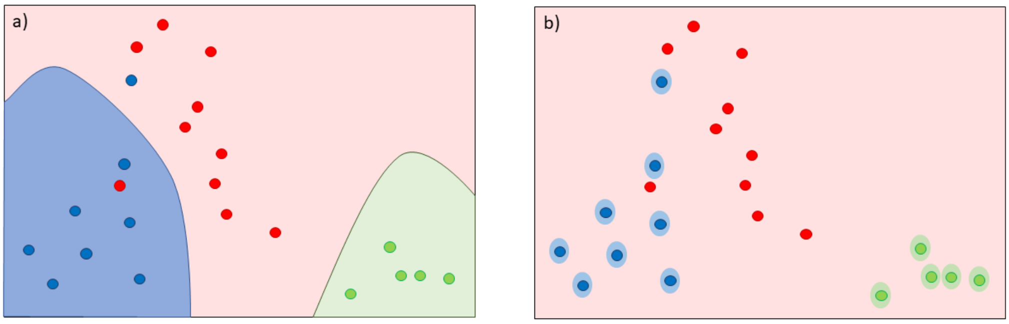


Figure 2.9 - SVM classification with RFB kernel. On the left a low gamma has been used, on the right the chosen gamma value is high

When classes to identify are more than two (multi-class), another choice to take is how to manage comparisons between classes and, consequently, the location of the separator. Having three classes A, B, C to deal with:

- One vs rest (OVR): the algorithm finds the first hyperplane that separates class A and the other classes (class B and class C), then the second hyperplane that separates class B and class C
- One vs one (OVO): the algorithm compares all the classes separately and then combine the results together

Comparing OVR and OVO, the first approach is less expensive in terms of computational cost, the latter is less sensitive to unbalance because hyperplanes are defined in the same conditions for each class

2.3 Case study

The distance sensing for the depth camera is mapped using coded light technology. As seen in Chapter 1, coded light depth acquisition technology consists in projecting an *a priori* known pattern on the surface of an object, a face or any other element of the environment and obtaining depth information of the surface through a receiver according to the way in which the pattern is deformed. This technology is designed specifically for applications at close range and this kind of sensors specifically works best within the range 0.2-1.5 m.

Intel RealSense SR300 has been chosen. The projected pattern is infrared (IR) light, therefore out of the visible spectrum. The depth video format, as well as the frame rate, is adjustable. Typically, the recommended resolution is the best available (640x480), while the frame rate is 30 FPS. The output of every frame is a depth map.

Figure 2.10 shows a picture taken during the interview.

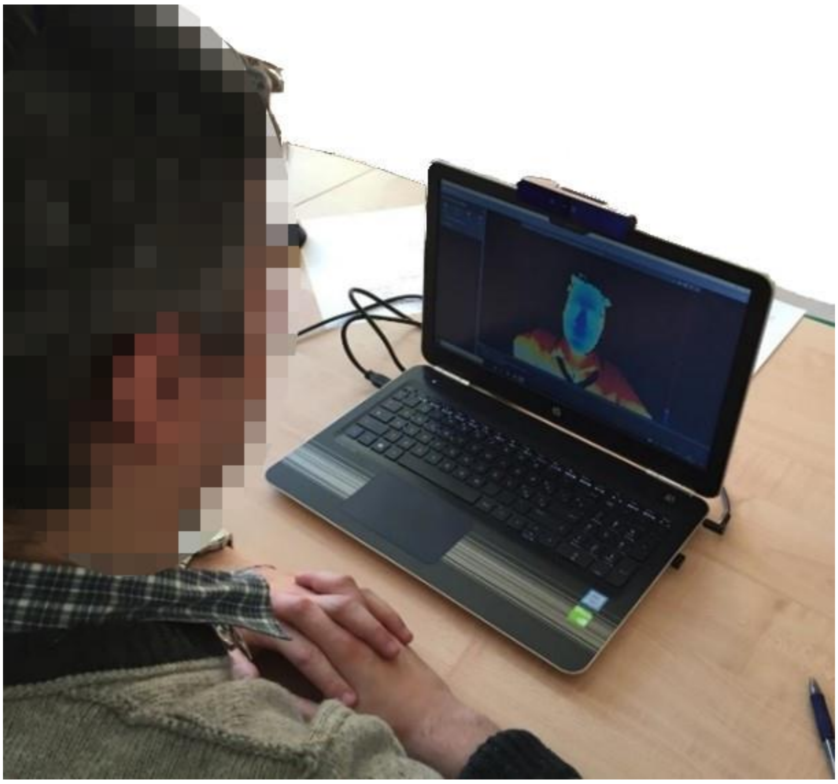


Figure 2.10 - Interview setup. On the laptop monitor, it is possible to see one of the depth frames acquired

234 frames and subsequent facial depth maps have been selected and post-processed, relying on the identification of the key moments of the interview between every question and every answer.

.

Parameters described in the Section 2.2 cannot be chosen a priori, but several trials must be done to set the proper values. In this case-study, classes to consider are three (deactivation, pleasure, arousal); OVR approach has been preferred to OVO; the kernel is RBF since data could not be split in a linear way; the C parameter has been set to 100, a high value compared to the default value that is 1, this testify the need of a great accuracy at the expense of less flexibility not necessary since the training set of frames has been fixed; the gamma parameter has been set to 10^{-8}, a very low parameter not to advantage one class over the others.

SVM belongs to machine learning supervised methodologies, thus needs a training phase to establish a relationship between input and output. After several trials, an amount of 234 depth frames has been extracted from the video acquired with Intel RealSense SR300 and has been subdivided into:

- Training set (30%): these are the frames labelled and used to train the SVM classifier.
- Validation set (10%): these frames have not been used to train the SVM nor to test it, but to tune C and gamma parameters before using the SVM for the frames of interest.
- Testing set (60%): these 138 frames have been used to test the SVM method. Furthermore, they were the frames of interest, namely the frames identifying the emotional level of activation during the interview.

Data dimensionality is considerable: 7 geometrical descriptors, i.e. Smean, Fden2, k1mean, sing, k2median, gmean and Hmean (variations of the geometrical descriptors previously described) and 16 Euclidean distances between landmarks have been computed for every frame. Moreover, geometrical descriptors have been computed on the whole face and on specific facial areas (eyes and mouth), and information have been stored considering 9 bins for every geometrical descriptor; for the sake of clarity, bins are the bars of the histogram of a facial descriptor.

In the next pages some images representing the steps previously described have been reported. Figure 2.11 represents the point cloud in 3D in Matlab Viewer, which allows to compute geometrical descriptors. Figure 2.12 represents the result of the landmarking step, which allows to subsequently compute the Euclidean distances.

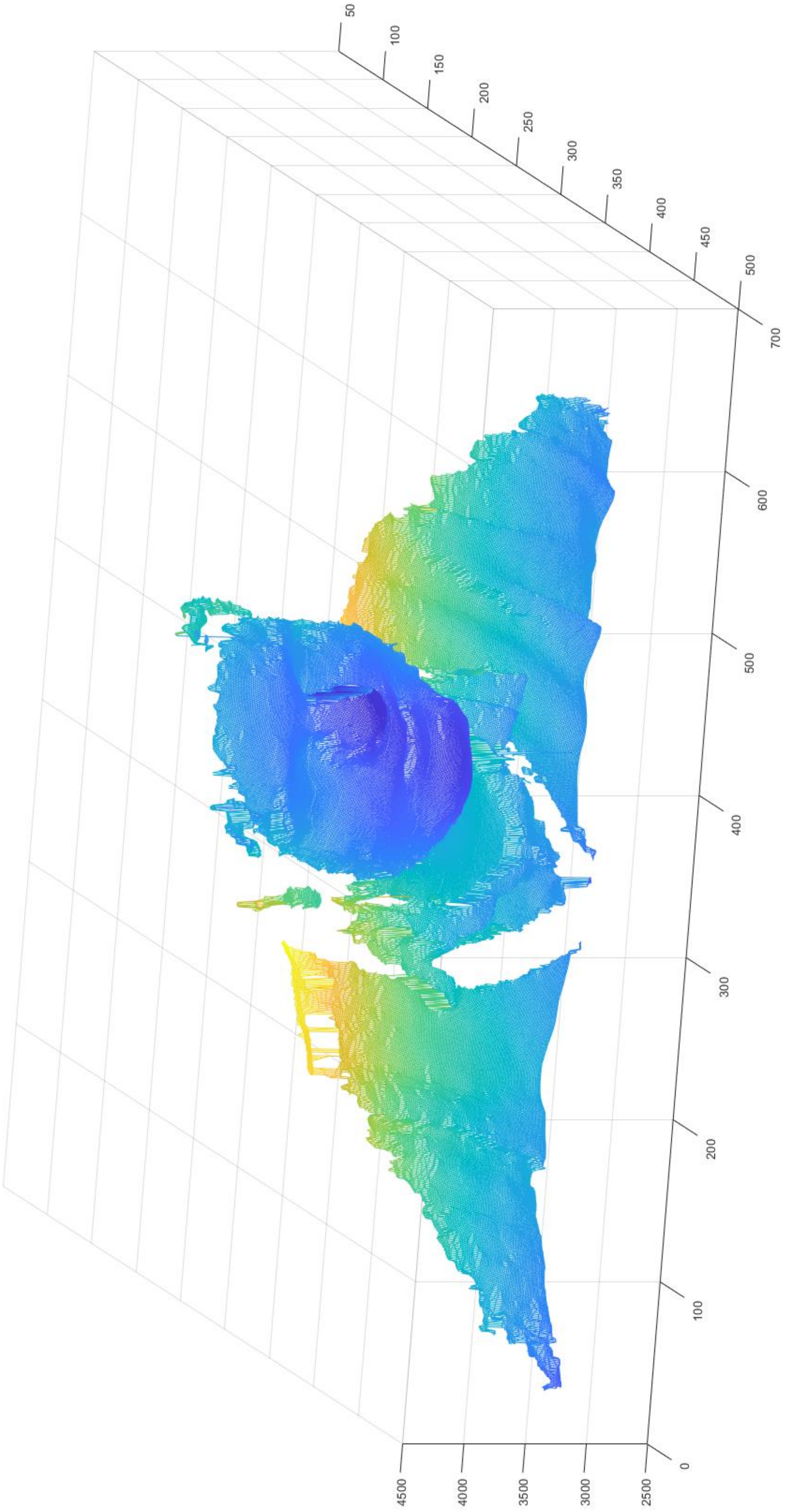


Figure 2.11 - Depth map represented in 3D in Matlab

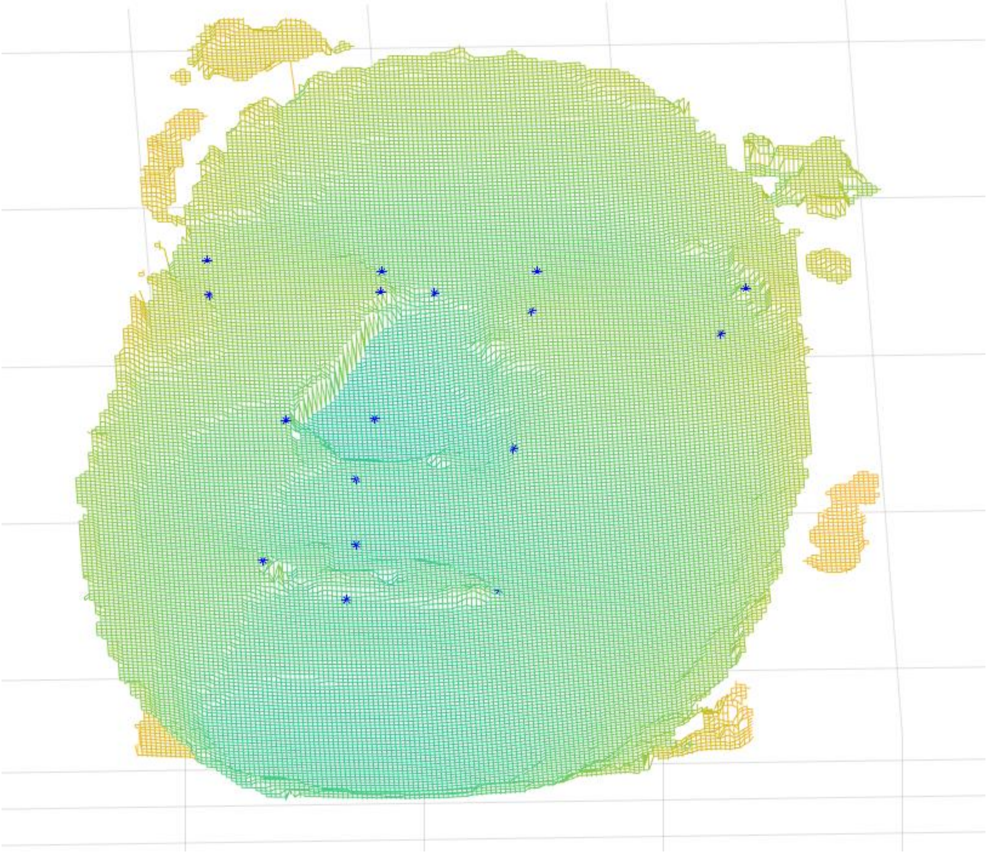


Figure 2.12 - Landmarks localization. Landmarks are displayed with blue asterisks

According to the type of data and the emotions displayed by the interviewee's face, the simplified model has been chosen in this case, to make easier, but more accurate, the classification. Thus, three clusters have been adopted: *deactivation*, *pleasure,* and *arousal*.

2.4 Results and discussion

The SVM methodology has been developed to recognize different emotional levels of activation, not different emotions.

The advantage of the present technique is that the interviewee involvement in the discussion can be analyzed for every question and answer or, from the technological point of view, frame-by-frame.

The counterpart is that the proposed technique requires a depth camera, which, however, has a relatively low cost and a certain knowledge about classification techniques. Nonetheless, several software and routines are already open and available for the community.

Obtained results must be analyzed under different perspectives.

Firstly, the proposed recognition method for emotional activation has an accuracy of 81%. This recognition rate could improve changing or adding new features to SVM classifier, such as different geometrical descriptors or even some elements provided by the color information acquired with the Intel RealSense SR300. Moreover, a calibration step could help if the usage of Euclidean distances is maintained, otherwise those specific features could be misleading among different users.

Secondly, this work must be seen as a case-study introduced and faced to acquire experience within the context of a thesis which has started with the RGB-D cameras analysis and aims to lay the foundation for a reliable and completely automatic face expression recognition application.

Chapter 3

Face Expression Recognition using Deep Learning and ecological valid dataset

Facial expressions are the result of muscles' movements on the face. These movements can be voluntary or involuntary and represent one of the main forms of nonverbal communication between humans. This aspect has been deepened by several researches and is also strongly considered in the popular culture; in the American crime drama *Lie to me*, the main character, Dr. Cal Lightman (Tim Roth), is an expert of body language, with a particular ability on identifying and interpreting the facial microexpressions. It is not surprising to discover that one of the major scientific consultants for the realization of this series has been Paul Ekman, one of the founding father of face expression recognition.

As Dr. Lightman had to find out the real emotions hidden by suspects', our research brought to light the difficulties that classifiers could encounter in learning how to recognize expressions from fake images. In this case, *fake* is referred to the fact that some databases of people showing facial expressions exists, but people are usually actors who pretend to feel an emotion, which is not spontaneous, and could not perfectly reflect the movements and the micromovements of subjects that truly feel an emotion.

Recent studies show that deep learning methodologies proved to be very promising and to easily provide state-of-art results in the context of face expression recognition, thus the works described in the next two chapters have been carried on employing a convolutional neural network, a class of deep neural network which resembles the organization of the animal visual cortex in the connectivity pattern between neurons.

The work described in this chapter is the result of a fruitful collaboration with a research group of Politecnico di Milano, which has given a great contribution especially regarding the psychological aspects of the research.

3.1 Introduction

The main problem in using a supervised classifier to perform face expression recognition is to find valid data to train machine learning algorithms. Data that have to be inputted in the classifier for the training phase must be labelled and to do such an operation a scientific approach to describe emotions is necessary.

The first study dealing with a quantification of emotions was initiated by Wundt [153] and continued by Scholsberg [154], that introduced a three-dimensional model which dimensions were pleasant-unpleasant, tension-relaxion, and excitation-calm. Ekman [155] recommended to merge the two last dimensions, because they resulted too similar each other, and Russell [148] developed the Circumplex Model of Affect, that has been used in Chapter 2 for the definition of the SVM classes.

In Figure 3.1 Circumplex Model of Affect is shown. On the left the eight affect concepts are arranged in a circular order, as well as the twenty-eight affect words that are displayed on the right.

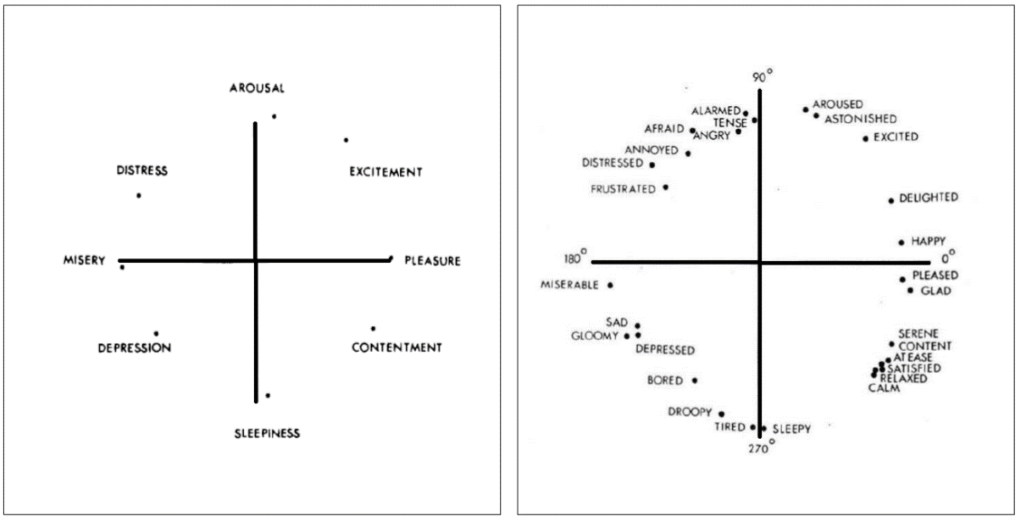


Figure 3.1 - Cicumplex Model of Affect

The expression that fits at best with the purpose that researchers must achieve to obtain useful data regarding facial expressions is *ecological validity*. Ecological validity refers to the condition according to which the facial expression of a subject must be due to a certain stimulus and not to boundary conditions. Within the context of an experiment, this means to find a trade-off between the experimental rigor, necessary to compare results obtained from different subjects, and the comfortability of the subjects themselves, who must express feelings only due to the stimulus received, not conditioned by constraints imposed by the experiment.

An experiment that aim to study facial expressions can be set up in different ways. First of all, participants can be asked to act or to express their spontaneous feelings; then, the format of results must be established [156] and, consequently, the necessary equipment must be procured (a standard video camera, an RGB-D camera, sensors to obtain physiological data, a database to store answers to a questionnaire, …); moreover, if the choice falls on spontaneous reactions, stimuli must be defined.

Among the variety of stimuli raising emotions, such as audio-visual [157], movie clips [158], music tracks and game scenarios [159], for the present work images stored in affective databases have been chosen; the next section will deepen the description of affective images. After that, the topic about affective measures and scales (above all, the Self-Assessment Manikin) will be faced. Later on, the focus will be moved to the participants to the experiment for both spontaneously raising and recognizing their facial expressions, where the recognition has been performed with a Convolutional Neural Network (CNN); the participants have undergone a double test about empathy and alexithymia to guarantee their reliability both for the expressions they have shown and the questionnaire they have compiled. The description of the experiment will precede the discussion of the results and the further improvements in the future work.

For the sake of clarity, in Figure 3.2 the description of the main step of the experiment has been reported.

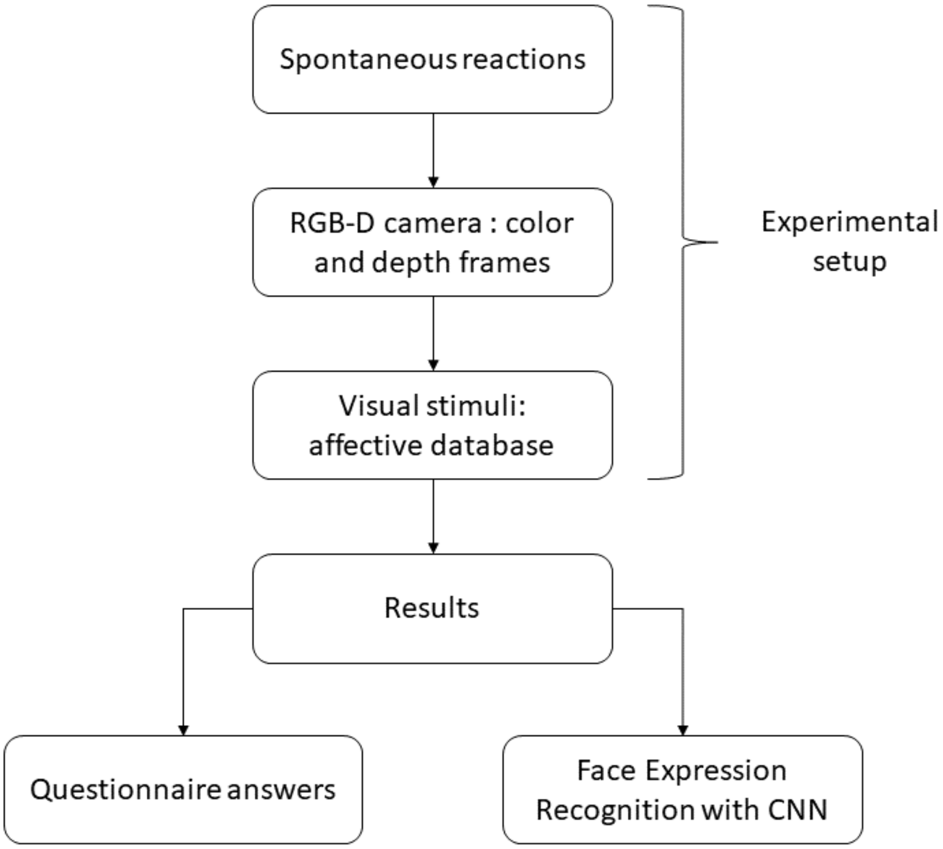

Spontaneous reactions
RGB-D camera : color and depth frames
Visual stimuli: affective database
Experimental setup
Results
Questionnaire answers
Face Expression Recognition with CNN

Figure 3.2 - Flow-chart of the planned experiment

The final aim of this ambitious project is to build an ecological valid dataset within which RGB-D images are stored. These images should represent spontaneous facial expressions, indispensable to train a deep learning algorithm, in particular a CNN, even if could be used to train other supervised machine learning algorithms as well.

The result that have been obtained up to now is twofold: on one side, a comparison between elicited emotions and expected emotions has been drawn up;

on the other side, a remarkable recognition rate of the CNN has been achieved. As it will be shown in Section 3.3, these satisfying and promising results have let to clearly identify the following steps to obtain a complete ecological dataset and to further improve the CNN recognition rate.

3.1.1 Affective databases

Once that visual static stimuli have been chosen, proper images had to be selected. After a review about affective databases, i.e. datasets containing images that arouse specific emotions, IAPS and GAPED have been chosen.

The International Affective Picture System (IAPS) is the most known affective databases, intended for research use only. The version of the database used for this experiment is composed by 1182 images (the database has been subjected to updates through the years) subdivided in semantic categories to arouse different emotions. IAPS has been largely used in psychiatric applications: Taskiran et al. [160] have studied the responses to emotional stimuli in patients affected by attention-deficit hyperactivity disorder, Migliore et al. [161] have conducted a similar study in Relapsing-Remitting Multiple Sclerosis (RRMS) patients, Moret-Tatay et al. [162] and Bekele et al. [163] have dealt with middle-aged adults and older men respectively affected by Schizophrenia, Peter et al. [164] have compared emotional responses in subjects with personality disorders, cluster-C personality disorders and non-patients. Furthermore, the ability of processing emotions after traumatic situations has been investigated, such as post-earthquake distress by Pistoia et al. [165] and violated women by Martinez Navarro [166], but also the dependence from alcohol can have implications in the way of processing emotions according to Dominguez-Centeno et al. [167]. The vast majority of the above-mentioned studies focused on evaluating subjects' emotions through the analysis of physiological signals: electroencephalography (EEG), electrocardiography (ECG) and magnetic resonance imaging (MRI).

The Geneva Affective Picture Database (GAPED) is composed by 730 images. The intent in building this new dataset was to overcome a problem that arouse in using IAPS extensively: the impact of those images seemed to drop in terms of efficacy both for positive and negative emotions. In particular, regarding the negative ones, GAPED designers subdivided images in four classes: two of them represents animals (snakes and spiders), one concerns the violation of the social norms (defined by legality), one concerns the violation of personal norms (determined by morality). According to Dan-Glauser and Scherer [168], estimation of low tolerability of the stimuli related to social norms becomes relevant in the elicitation of anger, but also in disgust, pity, guilt, shame, and contempt. There are predictable dissimilarities in valence marks among the positive, neutral, and negative categories, but also in arousal rates, indeed it is usually possible to find a correlation since valence scores are rarely independent from arousal levels.

3.1.2 Affective measures and scales

To quantify an emotion is a critical task that has been largely discussed over the years; what people refer to when using the term *feeling* is only the conscious experience of an emotion. Nonetheless, Mehrabian and Russell have developed a three-dimensional model (Valence-Arousal-Dominance, VAD) to map an emotion into three independent dimensions:

- *Valence*: it describes the positivity or negativity of an emotion. Formerly, adjectives used to label valence were in the range *happy-unhappy*; later, the concept of positivity has been associated not only to happiness, but also to serenity and relaxation, as well as the concept of negativity has been associated not only to unhappiness, but also to frustration, annoyance.
- *Arousal*: it describes the level of mental activity inducted by the received stimulus, where the lowest can be associated to a status of boredom and the highest to frenetic excitement. Adjectives used to label this dimension are *stimulated-relaxed*, *excited-calm*, *awake-sleepy*.
- *Dominance*: it describes how a subject feels with regards to the aroused emotion in terms of *submission-dominance*. It is the most critical dimension to define; for example, fear and anger are similar emotions in terms of valence and arousal, both of them are negative, but the first is considered dominant, the latter is considered submissive.

Values to be given to valence, arousal and dominance should be continuous, hence the need of establishing a proper tool to evaluate these dimensions shows up.

The affective slider [169] is an alternative that allows to provide high-resolution measurements, specifically designed for arousal and pleasure (valence). It is a digital self-reporting device (Figure 3.3), that does not need to provide explanations to the user, since it is sufficiently self-explanatory.

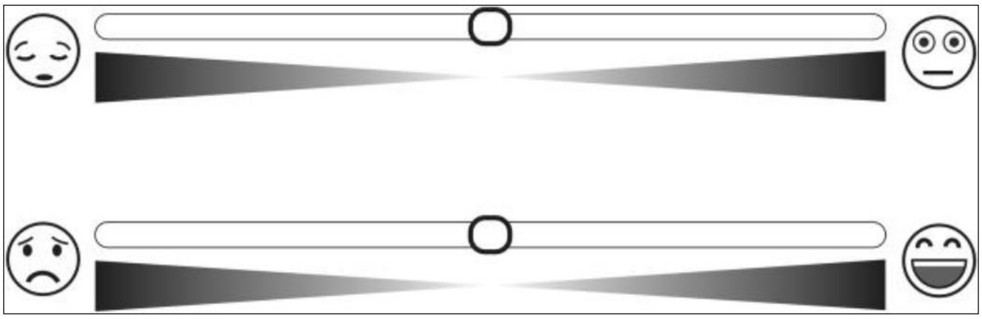


Figure 3.3 - The affective slider

The Self-Assessment Manikin (SAM) is a solution that maps the three dimensions into three different scales.

Valence ranges from pleasant to unpleasant; in the SAM implementation selected for this experiment, the lowest value is represented by a smiling figure, while the highest value is represented by a frowning figure.

Arousal ranges from calm to excited; in the SAM implementation selected for this experiment, the lowest value is represented by a sleepy figure, while the highest value is represented by a wide-eyed figure.

Dominance has not been used in this experiment not to move the focus of the participants forcing them to answer a too demanding questionnaire. Furthermore, in literature and in describing images in affective databases, it is the least used dimension; nonetheless, for the sake of completeness, SAM represents the lowest level of dominance with a tiny figure to give the idea of being completely submitted to the emotion, while the highest level of dominance is represented with a large figure that gives the idea of a total control of the situation.

Figure 3.4 shows an example of SAM: from the top line to the bottom one, icons of Valence, Arousal and Dominance are displayed.

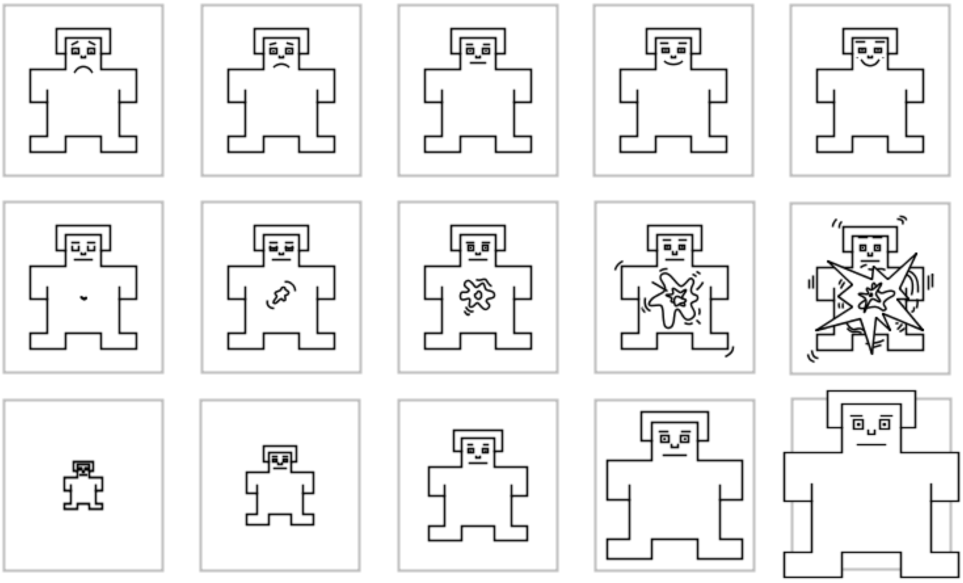


Figure 3.4 - Valence, Arousal, Dominance 5-levels SAM

SAM has been the selected scale representation in this study, also to keep the possibility of integrating the dominance dimension; nonetheless, a 7-level solution instead of a 5-level solution has been preferred, to partially mitigate the drawback of dealing with a discrete scale instead a continuous one. Moreover, the strong correlation between SAM and the Likert scale has been recognized by Huang and Chiang [170].

3.2 Methods

The experiment conducted to arouse emotions in participants has required to face a demanding design phase, which has been detailed in the following subsections.

3.2.1 Participants

First of all, spontaneous reactions have been preferred to forced expressions to maximize the reliability of the facial expressions.

Participants have been selected among students and PhD students of Politecnico di Torino, aged between 18 and 35, for a total number of thirty-five participants, fourteen female and twenty-one male subjects.

The nature of the experiment has required to ensure that participants had at least a standard level of empathy and were not alexithymic, thus each of them had to compile two tests before attending the experiment itself.

Empathy identifies the ability of identifying and understanding others' points of view, thoughts, intentions, and beliefs and is fundamental to build interpersonal relationships [171]. Empathy has two main components: the affective one and the cognitive one. The first refers to the affective reaction to another person's emotional state, the latter refers to the cognitive capacity to take the perspective of the other person [172]. The Balanced Emotional Empathy Scale (BEES) proposed by Mehrabian and translated into Italian by Meneghini et al. [173] has been adopted to evaluate participants' empathy. This test is composed by thirty questions and each answer requires a choice between strongly disagree and strongly agree on a seven-point scale. Results are subdivided in three ranges, namely below the average, standard and above the average, and have been displayed in Figure 3.5.

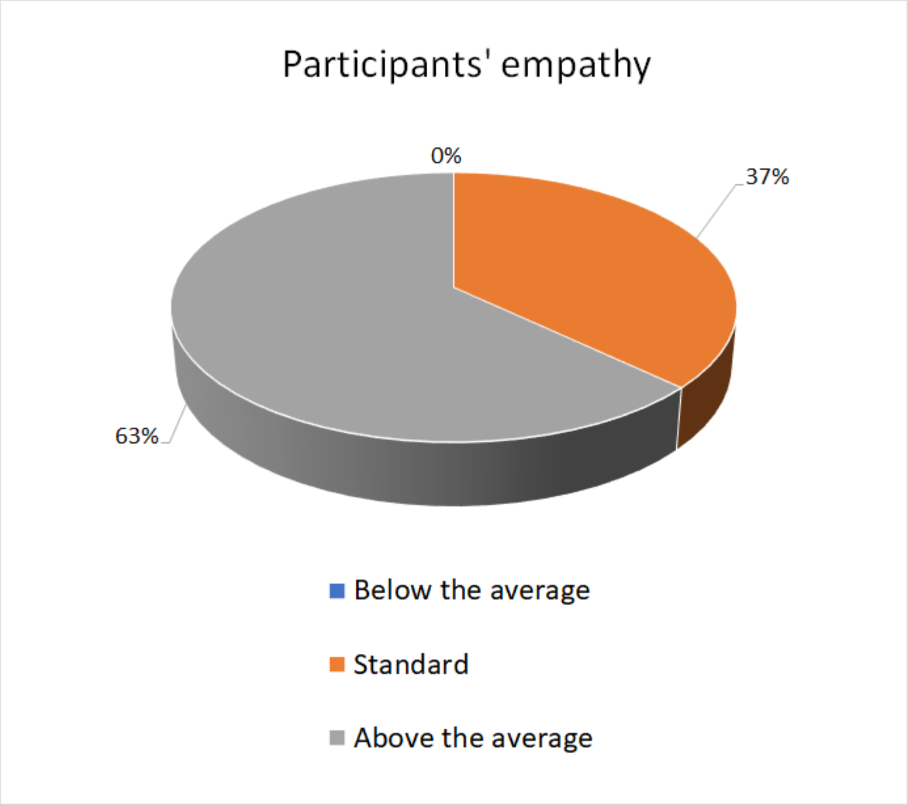


Figure 3.5 - Participants' empathy

Alexithymia identifies a reduced ability in recognizing, describing, and understanding one's own emotions [171]. It goes without saying that alexithymia and empathy are interlinked, because if one has difficulties in recognizing his own emotions, that person will have difficulties in recognizing others' emotions too. The Toronto Structured Interview for Alexithymia (TAS-20) has been adopted to evaluate participants' alexithymia [174]; this test has a 3-factor structure: the first evaluates difficulty in recognizing feelings, the second evaluates the difficulty in describing feelings, the third one considers externally oriented thinking. There are twenty questions, and each answer requires a choice between strongly disagree and strongly agree on a five-point scale. Results are subdivided in three ranges, namely non-alexithymic, borderline and alexithymic, and have been displayed in Figure 3.6.

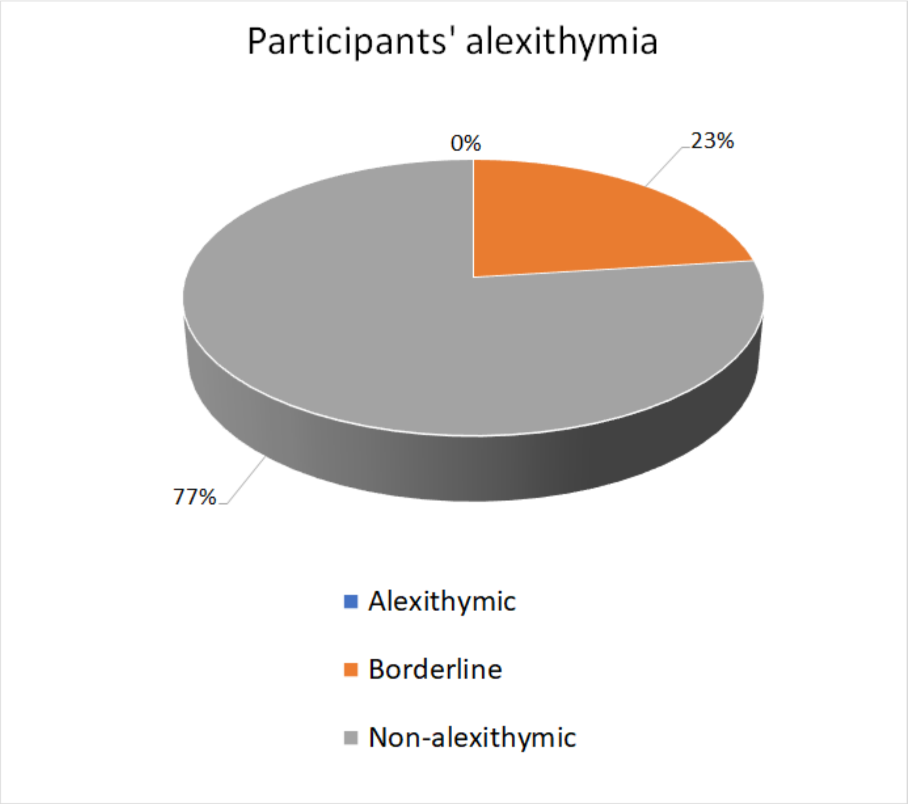


Figure 3.6 - Participants' alexithymia

Empathy and alexithymia tests have been reported in Appendix B and Appendix C respectively.

3.2.2 Picture stimuli

The choice of visual static stimuli, i.e. images, has been done to obtain an important advantage, which is to identify the exact moment during which the image is shown to the participants, namely the moment from which it is reasonable to search for a facial expression in the video acquired with the camera.

Databases from which to gather the pictures have been IAPS and GAPED. This choice has been made after the literature review and, also, a trial day dedicated to identifying weaknesses in the planned design of the experiment. The chosen pictures have been selected to arouse the widest range of stimuli possible and to represent a selection of a comprehensive sample of contents across the entire affective space.

Unfortunately, literature lacks a unique validated system to relate the Russell model (Figure 3.1), the six Ekman's expressions (happiness, sadness, anger, fear, disgust and surprise), and the images stored in the IAPS database, that have been classified according to valence and arousal values. Nonetheless, an attempt to put in relation these dependent dimensions has been done, according to the theory of emotions elaborated by Plutchik [175]. This effort has been done to be able to select the IAPS proper images to arouse specific stimuli without having images organized by emotions (there is no emotion label on IAPS images), but only by valence and arousal values. Results of this mapping operation is shown in Figure 3.7 and in Figure 3.8.

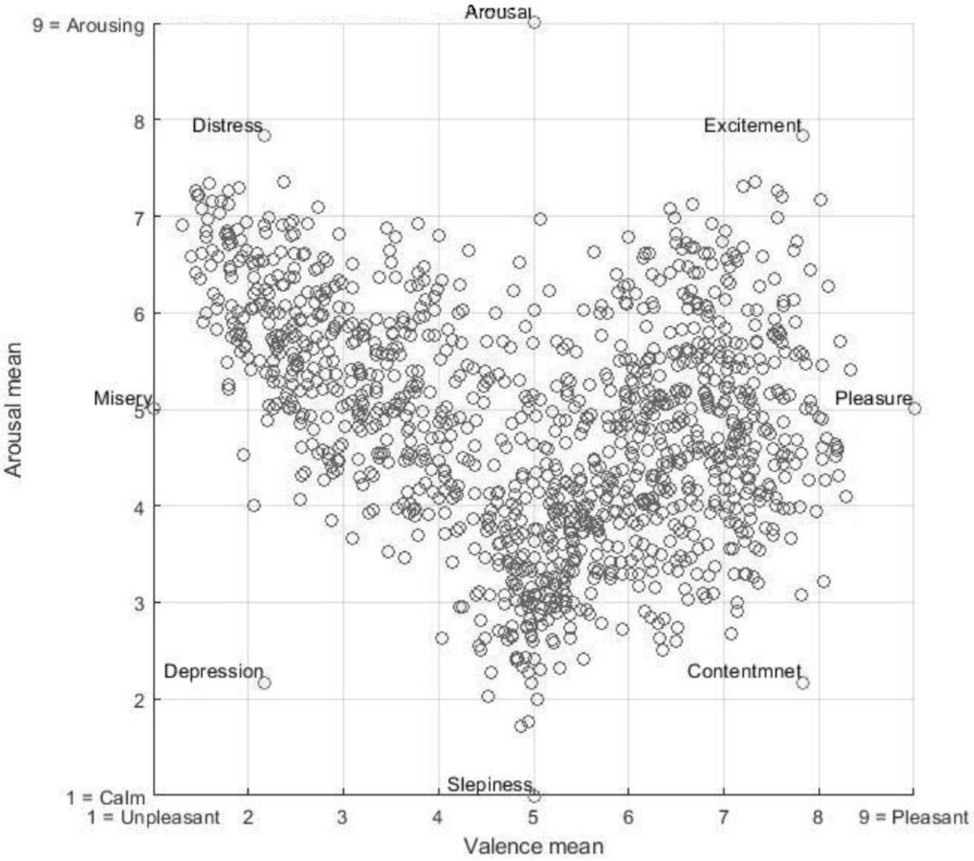


Figure 3.7 – Valence-arousal and Russel's model mapping (affective space)

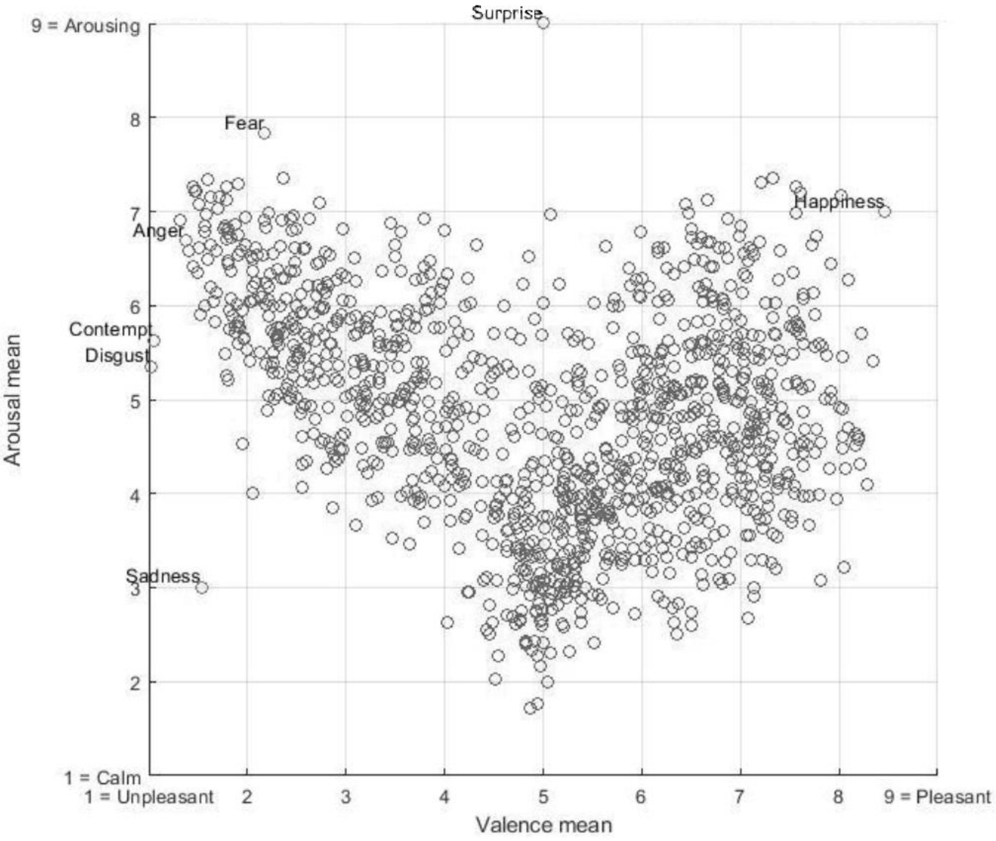


Figure 3.8 - Valence-arousal and Plutchik's theory of emotions mapping

Images stored in the GAPED database had both a labelling system to describe which emotions should arouse and the scores of valence and arousal, even if ranging from 0 to 100, so a normalization has been performed. Despite this operation, a poor correlation between values of IAPS and GAPED associated with positive emotions has been obtained. Some examples of inconsistency are reported in Table 3.1.

Table 3.1 - Valence and arousal comparisons for positive images in IAPS and GAPED databases

	IAPS		GAPED	
Image subject	**Valence**	**Arousal**	**Valence**	**Arousal**
Puppies	8.34	5.41	8.68	3.3
Baby	7.86	5	8.37	3.2
Seal	8.19	4.61	8.61	1.68

These issues in making correspondences between the two datasets can be explained by the different characteristics of the images, IAPS ones looks older, even if GAPED is only six years more recent the most used version of IAPS, but mostly because of cultural factors of people that evaluated the pictures; indeed, IAPS has been developed by the National Institute of Mental Health Center for Emotion and Attention at the University of Florida and images have been evaluated by one hundred people aged 18-24, while GAPED comes from Geneva, Switzerland, and images have been evaluated by sixty people aged 19-34.

To solve the issue, all the images have been carefully selected one-by-one.

GAPED images have been selected more easily since the database is arranged in folders (humans, animals, neutrals, spiders, snakes, and positive categories). Table 3.4 shows the twenty-four GAPED images values.

IAPS images have been selected assigning the correct labels of emotions to each picture considering in a first instance the work of Coan and Allen [176], then, to exclude some contents that could have been resulted ambiguous, the work of Bradley et al. [177]. In practice, the contents that generates the same emotion in the two genders has been considered and the emotion with the major percentage has been taken as predominant (Table 3.2 has been reported from [176]).

Table 3.2 - List of the most frequent specific emotion descriptors selected when viewing different picture contents in the IAPS and the proportion of men and women selecting that specific emotion to describe their affective experience. List of emotion descriptors included: Happy, Loving, Sexy, Excited, Romantic, Satisfied, Comfortable, Free, Amused, Playful, Nurturing, Bored, Confused, Irritated, Sad, Angry, Afraid, Anxious, Pity, Disgusted, Impatient

Picture content	Women		Men	
Erotic couples	Romantic (.41)	Sexy (.37)	Romantic (.47)	Sexy (.44)
Opposite-sex erotica	Amused (.36)	Embarrassed (.22)	Sexy (.50)	Excited (.40)
Same-sex erotica	Bored (.56)	Confused (.26)	Bored (.56)	Confused (.17)
Adventure	Excited (.63)	Free (.66)	Free (.61)	Excited (.55)
Sports	Excited (.69)	Free (.60)	Excited (.55)	Free (.52)
Food	Happy (.37)	Satisfied (.17)	Happy (.27)	Excited (.17)
Families	Happy (.79)	Loving (.78)	Happy (.58)	Loving (.58)
Nature	Free (.76)	Happy (.60)	Free (.56)	Happy (.41)
Pollution	Disgust (.56)	Irritation (.43)	Disgust (.34)	Irritation (.26)
Loss	Sad (.79)	Pity (.56)	Sad (.61)	Pity (.59)
Illness	Pity (.67)	Sad (.69)	Pity (.58)	Sad (.51)
Contamination	Disgust (.88)	Irritation (.50)	Disgust (.78)	Irritation (.40)
Accidents	Sad (.63)	Pity (55)	Pity (.50)	Sad (.49)
Mutilation	Disgust (.81)	Sad (.47)	Disgust (.75)	Pity (.42)
Animal Threat	Afraid (.69)	Anxious (.31)	Afraid (.42)	Anxious (.23)
Human Threat	Afraid (.67)	Angry (.42)	Afraid (.37)	Angry (.35)

A final check with the affective space has been done before the following list of images was confirmed (24 IAPS images in Table 3.3, 24 GAPED images in Table 3.4). Images used in the training phase has not been reported.

Table 3.3 - Selected IAPS images

Description	Valence	Arousal	Emotion
Beaten woman	2.31	6.38	Anger
Soldiers	2.10	6.53	Anger
Soldier	1.51	7.07	Anger
Mutilation #1	1.79	7.26	Disgust
Mutilation #2	1.79	7.12	Disgust
Mutilation #3	1.80	6.77	Disgust
Mutilation #4	1.70	7.03	Disgust
Mutilation #5	1.48	7.22	Disgust
Mutilation #6	1.58	6.97	Disgust
Baby with tumor	1.46	7.21	Disgust
Injury	1.56	6.79	Disgust

Snake	3.79	6.93	Fear
Dog attack	3.09	6.51	Fear
Shark	3.85	6.47	Fear
Kiss	7.27	5.16	Happiness
Mushroom #1	5.42	3.00	Neutrality
Mushroom #2	5.15	3.69	Neutrality
Spoon	5.04	2.00	Neutrality
Bowl	4.88	2.33	Neutrality
Lamp	4.87	1.72	Neutrality
Toddler	1.79	5.25	Sadness
Sad child	1.78	5.49	Sadness
Injured child	1.80	5.21	Sadness
Car accident	2.34	6.63	Sadness

Table 3.4 - Selected GAPED images

Description	**Valence**	**Arousal**	**Emotion**
Animal mistreatment #1	2.12	5.89	Anger
Animal mistreatment #2	2.08	6.46	Anger
Animal mistreatment #3	2.40	6.88	Anger
Animal mistreatment #4	1.15	7.23	Anger
Animal mistreatment #5	1.71	7.46	Anger
Snake #1	4.94	6.09	Fear
Snake #2	2.44	6.5	Fear
Spider #1	4.21	5.44	Fear
Spider #2	4.85	6.4	Fear
Spider #3	3.94	5.63	Fear
Baby #1	8.07	3.38	Happiness
Baby #2	8.03	2.86	Happiness
Baby #3	8.21	2.72	Happiness
Puppies #1	8.19	3.37	Happiness
Puppies #2	8.68	3.3	Happiness
Baby fox	7.83	3.11	Happiness
Kitten	7.77	3.10	Happiness
Antenna	5.4	2.97	Neutrality
Chairs	5.01	2.06	Neutrality
Lamp and sofa	5.84	2.10	Neutrality
Animal in captivity #1	3.20	6.43	Sadness
Animal in captivity #2	1.80	7.48	Sadness
Animal in captivity #3	2.11	6.38	Sadness
Animal in captivity #4	2.08	5.68	Sadness

The final list is also result of an analysis conducted after the trial day; the number of 48 images was defined instead of the initial 60, a trade-off to use the greatest number of pictures preserving the participants' attention, and some images considered too dated (belonging to IAPS database) have been substituted. Images are uniformly distributed among the basic emotions: anger, disgust, fear, happiness, sadness, and neutrality. Moreover, Figure 3.9 identifies the images of the final dataset onto the Valence-Arousal plane. Surprise is not present among the labelled images.

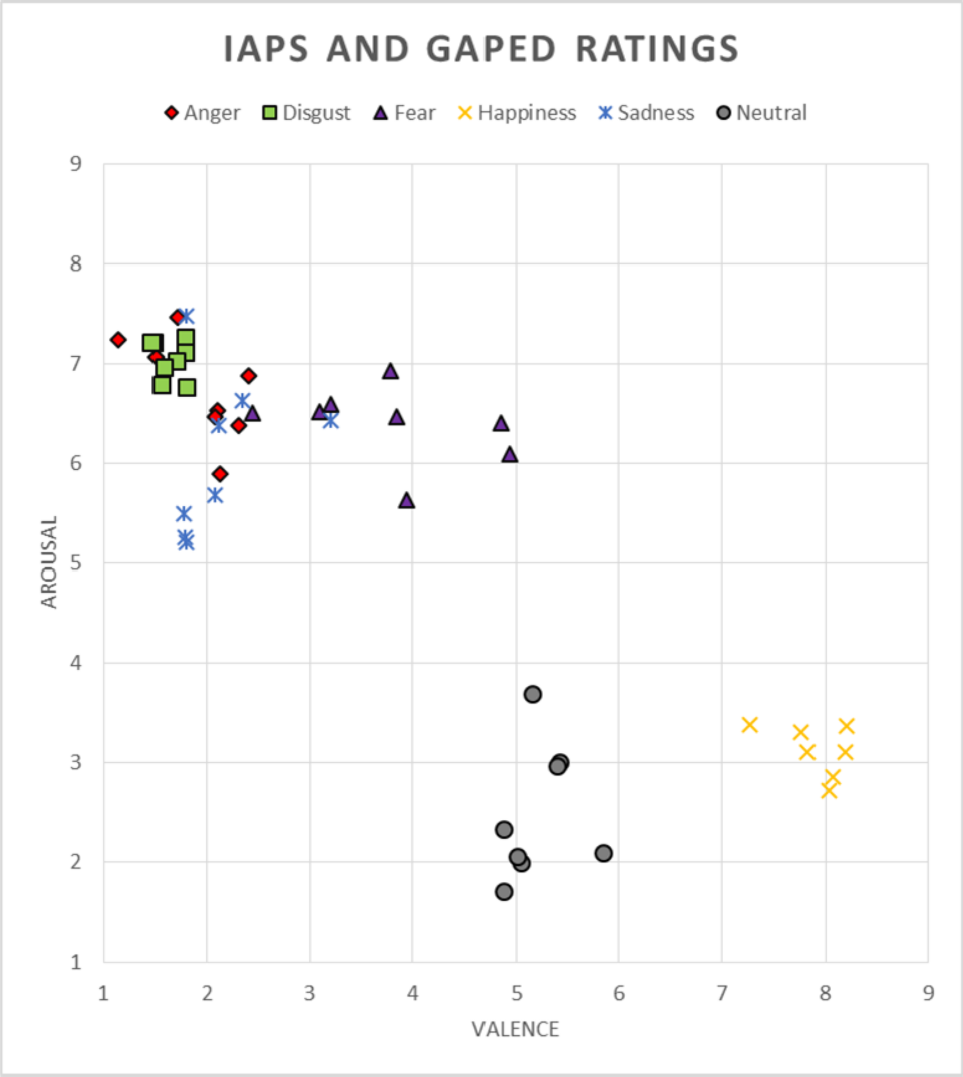


Figure 3.9 - IAPS and GAPED distribution in terms of Valence and Arousal

3.2.3 Experimental protocol and software description

The first step of the experiment has been the empathy and alexithymia test compiling, that participants carried out before coming to the laboratory.

Regarding the part of the experiment held in presence, at the beginning participants have attended a presentation to become familiar with the context and to receive the main indications on what to do during the experiment, without influencing their emotionality in any way not to corrupt the results. They have been warned about the presence of images that could have potentially bothered their sensibility and the possibility of abandon the experiment due to any kind of discomfort has been clarified.

The experiment has taken place in two phases: training and testing. The structure of both the phases has been the same: in a first instance one image provided by affective databases was displayed in full-screen mode (Figure 3.10), then the participants had to fill in the questionnaire about valence, arousal and the prevalent felt emotion (Figure 3.11). It has to be noticed that the label surprise has been inserted in the questionnaire, to let participants free of choosing the most proper basic emotion they felt, independently from the fact that images arousing surprise have not been inserted in the final dataset of 48 images.

Figure 3.10 - Example of full-screen image that aims to arouse happiness

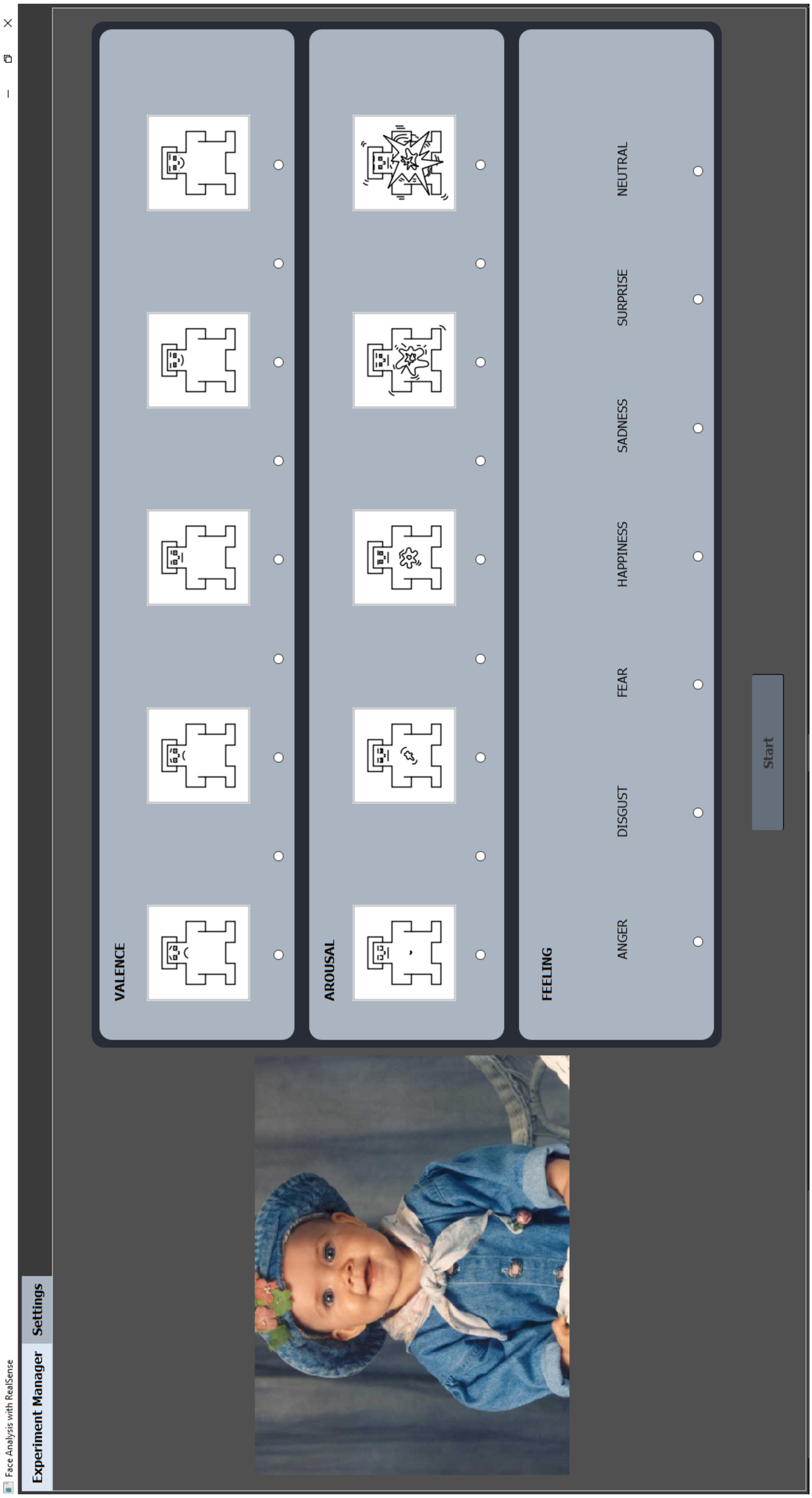


Figure 3.11 - Screenshot representing the questionnaire used for the experiment

The training phase has been useful mainly to get participants familiar with the questionnaire, because answers were forced to be given in no more than 15 seconds, to favor spontaneity. SAM icons are intuitive, but not so easy to interpret if never seen before.

The testing phase is composed by 48 images which are randomized for every participant and lasts about twenty minutes. Every participant has looked at every single image and has answered the relative questionnaire.

An ad-hoc software has been necessary to deal with both the management of images and questionnaire, maximizing the user experience not to distract the user from his task, and the management of the RGB-D camera recording. Indeed, the Intel RealSense SR300 has been connected to the same application using a different thread and has been set up to record user's expression from the moment during which the affective image appears on the screen to two seconds after it disappears, to be sure not to lose any expression. An idle interval of 2 seconds between the affective image and the questionnaire and between the questionnaire and the next affective image has been introduced. The affective image lasts 6 seconds on the screen; nonetheless, a smaller frame has been inserted next to the questionnaire as a reminder for the user.

In Figure 3.12 the experimental setup has been sketched.

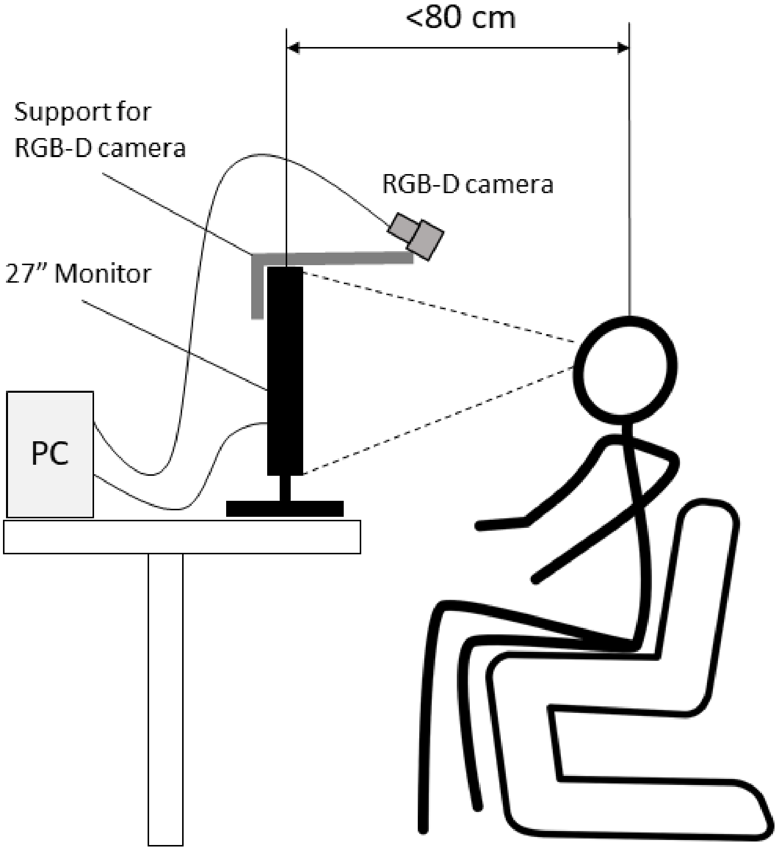


Figure 3.12 - Experimental setup

3.2.4 Face Expression Recognition using Deep Learning

A Convolutional Neural Network (CCN) has been used to perform Face Expression Recognition. This is one of the most used neural networks to identify objects and faces within frames; among the advantages that this deep learning method can provide, it is possible to mention the automatic feature extraction, cutting-edge recognition results and the possibility of retraining existing pre-set networks for other recognition activities.

Data obtained from the acquisitions, i.e. color and depth frames, must be processed before being used as input for the neural network. Three main steps can be identified:

- *Frame capture*: the RGB-D camera provides a double data stream temporally synchronized. Videos are 6 seconds long; it has been chosen to manually extract the most significant frames (Maximum Criterion variation) to be analyzed through the neural network.
- *Color to Depth alignment*: frames are automatically temporally synchronized but need a projection-deprojection process to be spatially synchronized. This way, each pixel on both frames refers to the same point in the space: the color information is stored in the RGB frame, the distance of that point from the camera is specified in the depth frame.
- *Face Detection*: in this context, face detection is not the identification of the bounding box around the face. The face must be identified only in its oval shape, as shown in Figure 3.13, not to mislead the neural network neither in the training nor in the testing phase. This is automatically obtained if the frames are spatially aligned.

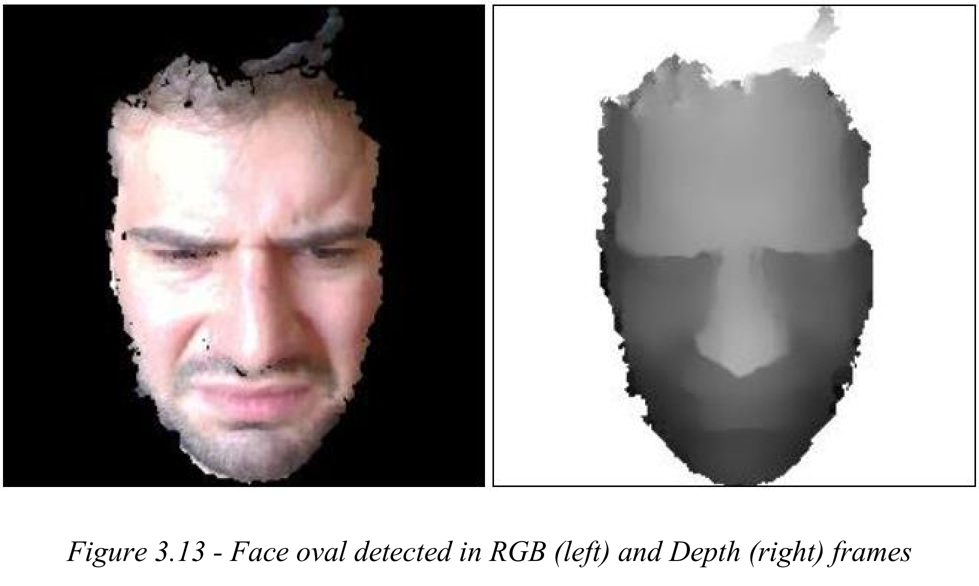


Figure 3.13 - Face oval detected in RGB (left) and Depth (right) frames

Once that these preprocessing steps have been completed, images must be sized adequately, then they are ready to be used as input of the CNN. The size must be predetermined since the VGG16 CNN, a pre-set neural network, has been used [178]. The size is 224x224.

The training phase is the most demanding step when dealing with deep learning: a huge amount of data is required, and the BU-3DFE that has been use for the training phase contains a limited amount of images: 1800 pictures to train 7 classes. Thus, transfer learning techniques had to be applied.

Transfer learning consists in using a model developed for a task as the starting point for a model on a second task. In this context, the model has been trained with some data, in this case images belonging to the BU-3DFE database, then the higher levels have been tuned using another dataset, more similar to the testing one. In other words, the neural network has been trained in two phases: the first one using BU-3DFE images, the second one using a small part of the dataset acquired during the experiment (221 images out of 1616). Labels to these images have been given by visual judgment.

The results of the testing phase are discussed in the next section together with the other experiment results.

3.3 Results and discussion

The main issue in the evaluation of the results has been that participants have showed expressions not always in line with the answers they gave to the questionnaire (Figure 3.14).

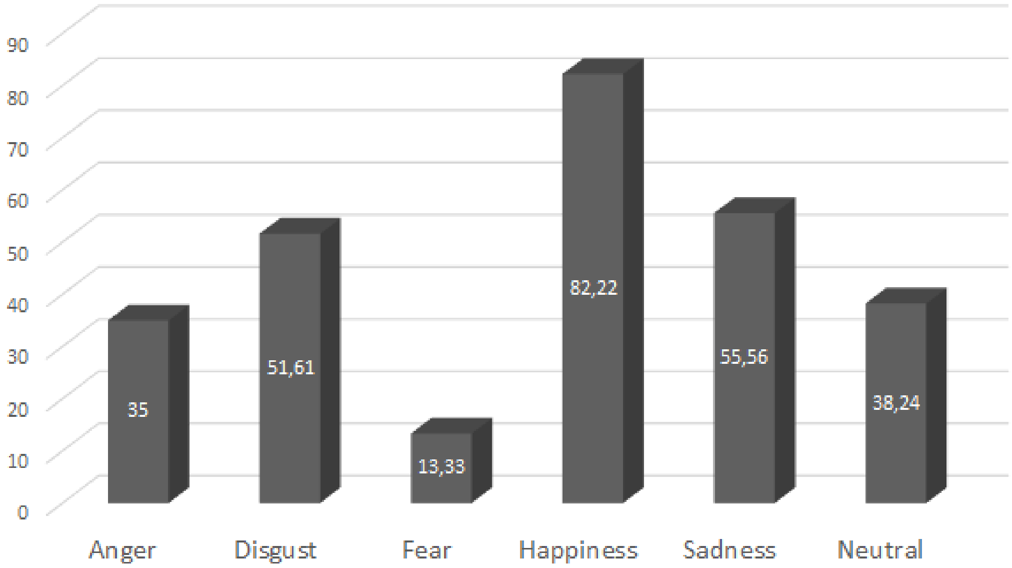


Figure 3.14 - Percentage of agreement between participants' labels an focus group's visual label. To obtain this graph, a focus group has been created to evaluate the expressions of participants.

This way, it is difficult to judge the CNN recognition rate, since the neural network can evaluate only the expressions, not the intentions of the participants. Reasons behind this results are mainly three: the former is the confusion that involves some negative emotions in the affective space, since they are located very close each other, thus valence and arousal values often overlap; the second one is the possibility of pointing out only the prevalent instead of multiple emotions and once again this could have mainly penalized the negative emotions, that can be aroused together; the latter, but the most important one, is the usage of a too weak stimulus to arouse emotions. Even if validated by the literature, visual stimuli are not as effective as past decades since people got used to these kinds of stimuli and need some stronger ones to physically express emotions. Participants have been warned to be as spontaneous as they could, nonetheless nobody has asked them to force expressions not to corrupt the results.

In the light of this, results section has been split in two parts to show the result in a more systematic way.

3.3.1 Expected emotions vs aroused emotions

This subsection compares the emotions expected to be aroused and the emotions pointed out in the questionnaire by the participants. This comparison aims to verify if the images chosen from the affective database have been effective.

In Table 3.5 the six considered emotions are displayed on the first column. From the second column to the last one, the indication of the emotion pointed out by the 35 users has been reported.

Table 3.5 - Comparisons between expected emotions and questionnaire answers

		Emotions reported in the questionnaires						
		Happiness	**Neutrality**	**Sadness**	**Disgust**	**Anger**	**Fear**	**Surprise**
EXPECT.	**Happiness**	79%	16%	1%	0%	0%	0%	4%
	Neutrality	8%	75%	2%	1%	0%	1%	13%
	Sadness	6%	7%	67%	1%	4%	7%	8%
	Disgust	0%	3%	15%	67%	3%	5%	7%
	Anger	0%	9%	23%	14%	44%	3%	7%
	Fear	4%	26%	0%	26%	0%	27%	17%

It can be noticed that the prevalent emotion found in the questionnaires matches with the expected emotion in every case.

To be coherent with the CNN training and the literature, surprise has been maintained as an available option to choose, even if not directly present among the affective images. In some cases, participants have chosen this emotion instead of neutrality because they did not know how to react. Anyways, 75% of matching between expected and aroused neutrality is remarkable, as well as the 79% of happiness.

Obtained results are perfectly consistent with Table 3.2. For instance, mutilations should arouse disgust both for men and women, then sadness in women and pity in men. Pity is not a basic emotion, the closest one is sadness, and in our study the mutilations that have been chosen to arouse disgust, have aroused disgust in the 67% and sadness in 15% of the participants.

Anger images have been mostly evaluated as anger (44%) or sadness (23) or disgust (14%) confirming the not so clear area of the affective space occupied by these three emotions.

Fear has been the emotion aroused with less success (27%). According to Edwards et al. [179] disgust can be part of the emotional reaction to certain phobic stimuli, this explains why it has been chosen from the 26% of the participants, as well as the neutrality, simple to explain that the 26% of the participants have not felt these images frightful enough.

After that emotions have been analyzed, a comparison between valence and arousal values expected from one side, valence and arousal pointed out in the questionnaires on the other side has been carried on.

The 48 images have been represented in the affective space (singularly in Figure 3.15, compacted in Figure 3.16), both with valence and arousal values reported in affective databases and with valence and arousal values given by participants' answers to the questionnaire. In this last case, valence and arousal have been averaged among the 35 participants for every image, and to choose the emotion that each valence-arousal couple represents, the most selected emotion by the participants has been used.

Surprise has not been reported in the graphs because, as expected, it has been chosen a few times by the users and not significative for this comparison.

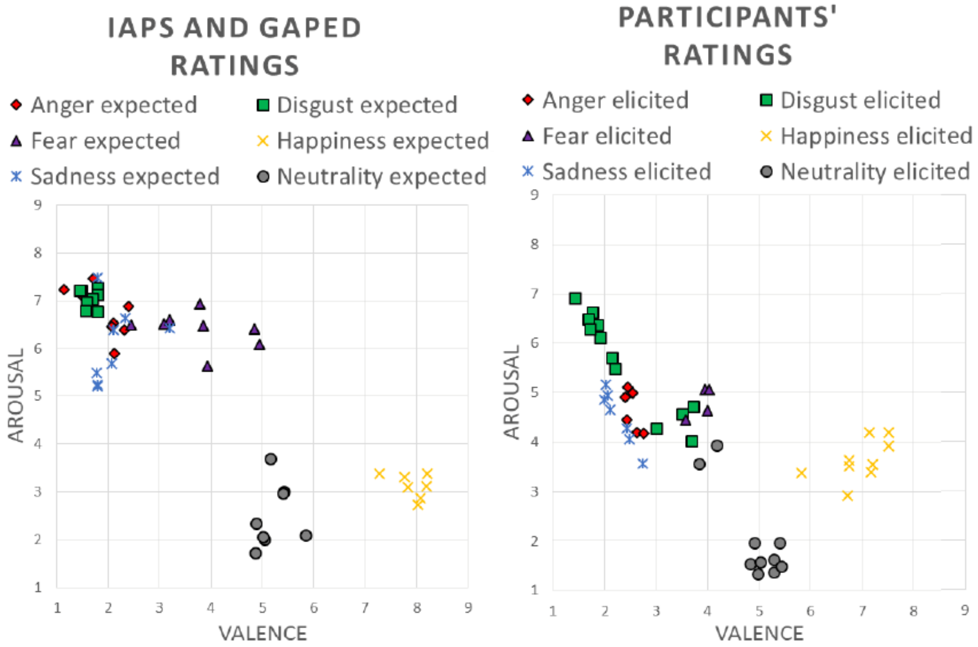


Figure 3.15 - Valence and arousal values comparisons in affective databases and obtained during the experiment

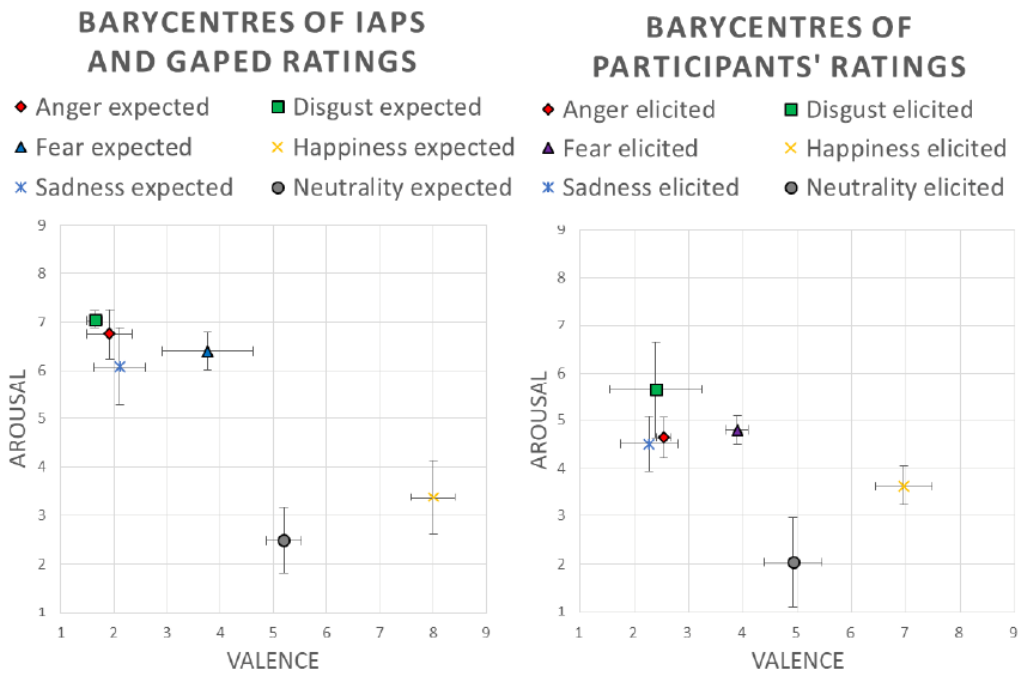


Figure 3.16 - Valence and arousal barycenter comparisons in affective databases and obtained during the experiment. Lines represent the standard variation.

It can be noticed that characteristics of IAPS and GAPED values are maintained in the values retrieved with the experiment. The most evident difference is a translation along the y-axis, i.e. the arousal dimension, which testifies that aroused emotions have been less powerful in terms of activation (Table 3.6).

Table 3.6 - Valence and arousal values reported from the affective databases (expected values) and obtained during the experiment (averages)

	Values expected		Values obtained	
	Valence	**Arousal**	**Mean valence**	**Mean arousal**
Anger	1.92	6.74	2.54	4.64
Disgust	1.65	7.05	2.39	5.63
Fear	3.76	6.39	3.89	4.8
Happiness	8.01	3.38	6.95	3.63
Sadness	2.11	6.07	3.82	3.06
Neutrality	5.2	2.48	4.91	2.04

The translation towards smaller arousal values can be ascribed to a fact previously mentioned: nowadays it is more difficult to arouse emotions in people that are continuously submerged by different and strong stimuli. In this context, a flat environment such as the one where the experiment has taken place, could have also negatively contributed to this aspect.

Figure 3.17 and Table 3.7 summarize the results just explained.

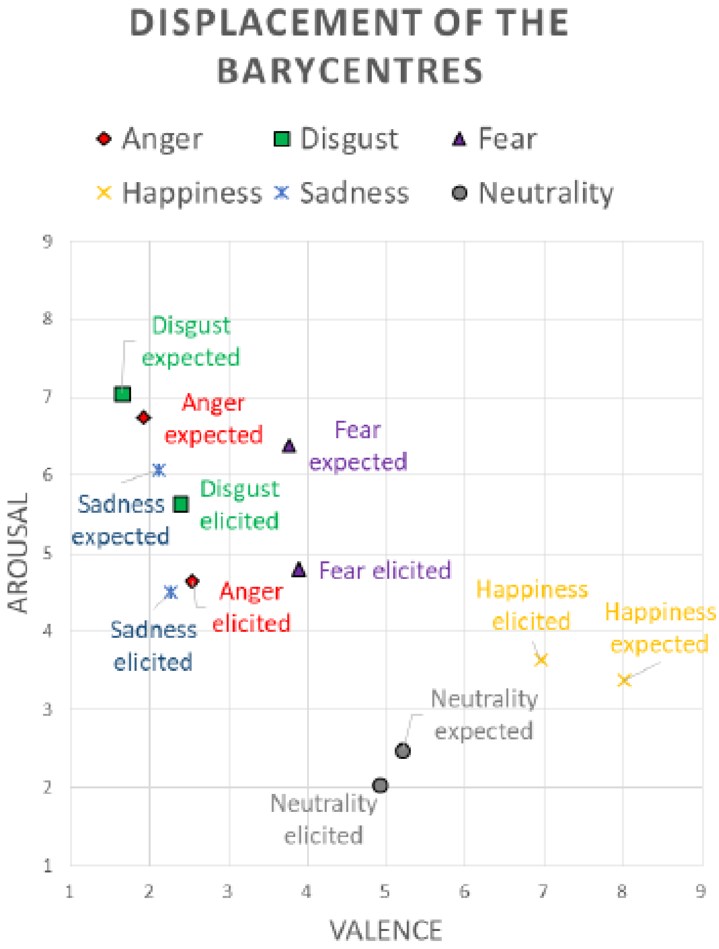


Figure 3.17 - Displacement of the barycenters

Table 3.7 - Displacement of the barycenters: values

	Euclidean distance
Sadness expected → Sadness aroused	3.46
Anger expected → Anger aroused	2.19
Disgust expected → Disgust aroused	1.6
Fear expected → Fear aroused	1.6
Happiness expected → Happiness aroused	1.09
Neutrality expected → Neutrality aroused	0.53

3.3.2 Expressions vs recognized emotions

This section compares the emotions acquired with the RGB-D camera evaluated manually and recognized by the CNN during the testing phase.

Emotions are recognized focusing on specific areas that can change considering the different features that the neural network has learned to evaluate during the training phase.

Figure 3.18 shows the result of the testing phase for some images. It has to be noticed that colored areas identify the part of the image on which the CNN has focused; moreover, each label (whatever it is: anger, neutral, fear, sadness, disgust, happiness) has a percentage above it: indeed, the CNN always computes the probability of each emotion. To display only the most probable has been a design choice for the sake of synthesis.

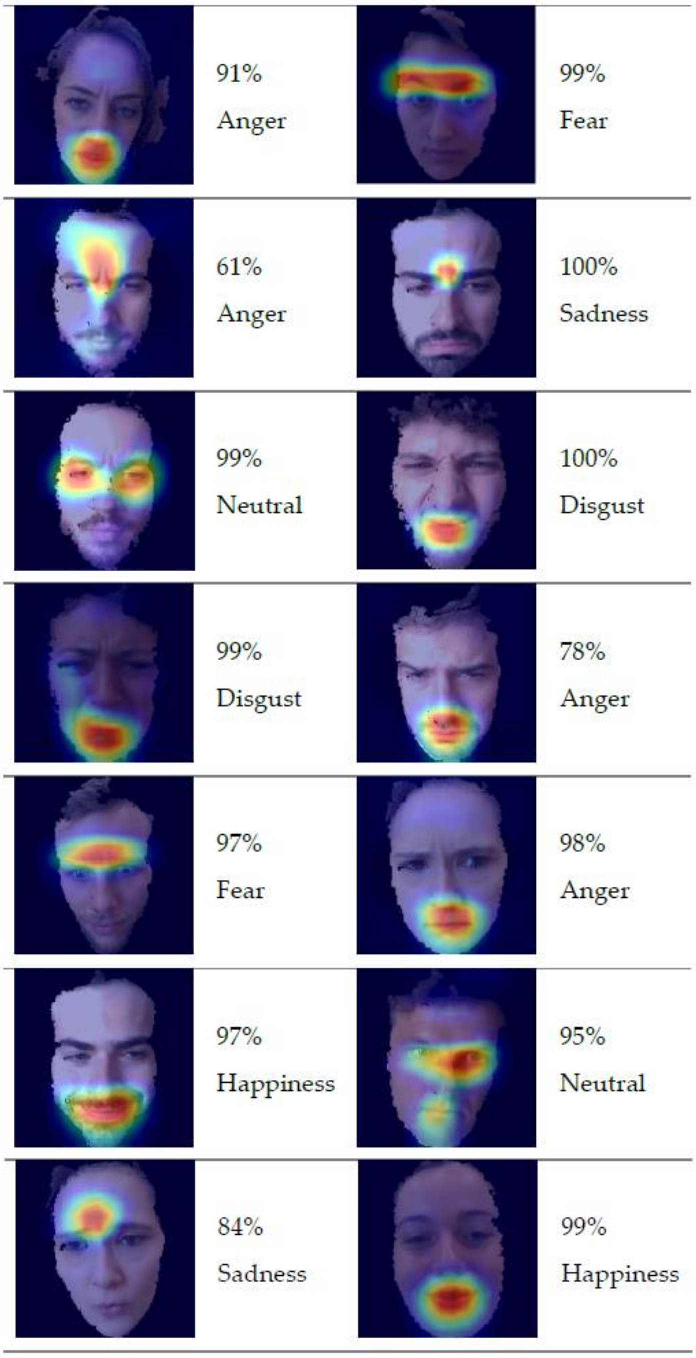


Figure 3.18 - CNN activation

Emotions showed by users in front of the camera have been evaluated and compared with the results provided by the neural networks. Best results have been obtained considering only RGB images, correctly recognized 3 times out of 4 (75,02%). Results considering only depth images have been not satisfactory (55,20%), while the combined usage of RGB and Depth has led to a 72,65% of agreement. Some considerations about these results must be done.

Head orientation of the participants was not the most suitable one as can be seen in Figure 3.12. The RGB-D camera has been positioned not in the center of the field of view of the participants (the area delimited with the dashed lines), otherwise the monitor would have not been visible. The result is that faces have been framed from the top; in particular, the CNN trained with depth images labelled a lot of images as *angry*, since areas highlighting anger have been pointed out this way (wrinkles between eyes).

This consideration leads to the second one: dataset for training is too limited, especially for the negative emotions. The training set is uniformly split, nonetheless negative emotions are really close each other in the affective space. This means that the CNN, as well as most of classifiers, needs more images to properly run, especially those images belonging to classes critical to be discriminated.

Finally, to maintain consistency, every component asked to evaluate emotions (participants to the experiment, human judges, and also the CNN has been forced to do it) assigned one single label to every image; nonetheless, some images could have led to feel more than an emotion as discussed in previous sections.

Regardless these considerations, the final recognition rate of 75,02% can be considered very satisfactory since it has been obtained using a neural network that can be further improved by targeted interventions in the next future.

3.4 Conclusions and future work

The number and the quality of the color images acquired with the RGB-D camera are not the only parameters that can be enhanced to improve the performance of the CNN. Depth images can be used as well to provide a different, integrable information source. Chapter 5 shows that CNN can be used also with the depth channel. For the present work there are two solutions to successfully integrate depth frames: the former is to use a smaller monitor during the experiment, so that the camera would be less misaligned with respect to participant's field of view; the latter is to exploit the intrinsic characteristic of 3D: to rotate the depth information pretending to align camera and face in post-processing. This would be easier to be done with a complete 3D model, depth map are screenshots, and some areas would be missing after the rotation process, nonetheless this is a solution to take in serious consideration if the rotation angle is sufficiently low, as in the case of the experiment already completed.

Another aspect on which we have already started to work on is the creation of immersive virtual reality environment to arouse specific emotions. The goal is to substitute images of affective databases to obtain more pronounced and clear expressions, especially for those emotions that occupy similar areas in the affective space (the negative ones: anger, disgust, fear, and sadness).

The satisfactory results, obtained using a neural network that can be improved training it not only with a database that prove to be too limited in terms of amount of images (BU-3DFE), leads both research groups to believe that to focus on the individual is essential to obtain better results.

The building of an ecological dataset goes exactly in this direction, aiming not only to provide quantity, but quality images to train the neural network with images describing realistic and truthful emotions.

Chapter 4

Real-time Face Expression Recognition

Real-time is an essential requirement for many applications and face expression recognition is not an exception. Nonetheless, it is not always clear what real-time exactly refers to.

According to literature, real-time systems must guarantee a response within a specified range of time, beyond which the response itself is useless. The time constraint is also referred as *deadline* and its value depends on the application that it is going to be designed.

Real-time requirement must be defined during the design phase of an application and can vary a lot depending on the considered application. For instance, thermal datalogger, which are devices designed both to measure and to store temperature data, acquire the temperature every 3 minutes, while Anti-Lock Braking Systems (ABSs), which are anti-skid braking system mounted on modern vehicles, can apply or release braking pressure fifteen times per second, and also the human reaction time is very low, on the order of a quarter of a second (250 milliseconds).

In our case, real-time has been defined considering the fact that the fastest voluntary movement that a person can perform is the eyes blinking. A human can blink up to five times per second. Consequently, this application is designed to perform Face Expression Recognition at a rate not lower than 5 frames per second (FPS).

This chapter aims to describe a procedure to recognize facial expression of a subject in front of an RGB-D camera using all the information provided by this sensor, namely both the color and the depth streams, and working automatically to give the opportunity of obtaining real-time.

4.1 Introduction

Real-time is an indispensable requirement for the vast majority of face expression recognition applications, such as live videos [180], facial tracking and animation [181], tutoring systems for children with autism spectrum disorder [182]. Awareness of the numerous fields of application faced in the previous chapters and the need of achieving the real-time requirement have led us to develop an automatic procedure to perform face expression recognition in real-time.

Deep learning methodologies proved to provide state-of-art results in the context of object recognition, both using only color images and employing 3D facial data [183]; among the other evidence in literature, Eitel et al. [184] have proposed a novel architecture focused on learning imperfect sensor data for object recognition in real-world applications, Li et al. [185] and Li et al. [186].

The literature works just described have a common trait; employed neural networks have been developed using both RGB and depth information, following a multimodal approach to improve final performances. Due to these results, the procedure of the data processing described in this chapter has been designed to deal with both the data streams provided by the selected RGB-D camera.

4.1.1 RGB-D camera employed in the study

The research has proceeded step-by-step, which are going to be showed in the present chapter.

The RGB-D camera choice has been fallen to the Intel RealSense SR300 (Figure 4.1). The criterion that has led to this choice has been to acquire color and depth information under the best possible conditions. The greatest concern was to maximize the accuracy of the depth information since the sensor depth resolution is usually much lower the sensor color resolution. To put the camera as close as possible to the subjects' faces is the most immediate solution to increment details acquisition, provided that the employed depth acquisition technology is different from stereoscopy.

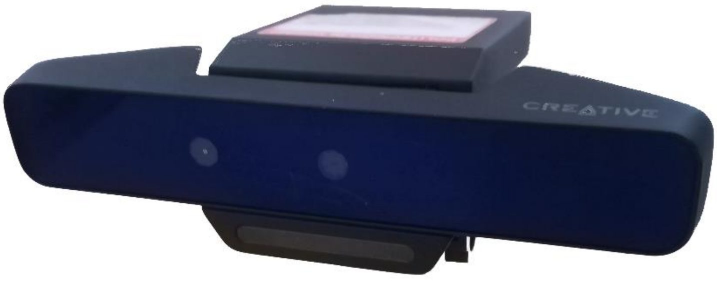


Figure 4.1 - Intel RealSense SR300

Intel RealSense SR300 is a coded light camera designed to work between 0.2m and 1.5m in indoor environments and controlled lighting situation due to the usage of a projector to acquire the depth information. It can work in totally dark (0 Lux) environments thanks to the infrared light.

Every image provided by Intel RealSense devices is a 2D matrix with *w* columns (width) and *h* rows (height). Each cell in the matrix represents a pixel and is identified by two indices. The pixel with coordinates [0, 0] is the top left pixel in the image referring to the center, while the pixel with coordinates [w-1, h-1] is bottom right one. Coordinates identified in this space are referred to as *pixel coordinates*.

Frames example are shown in Figure 4.2.

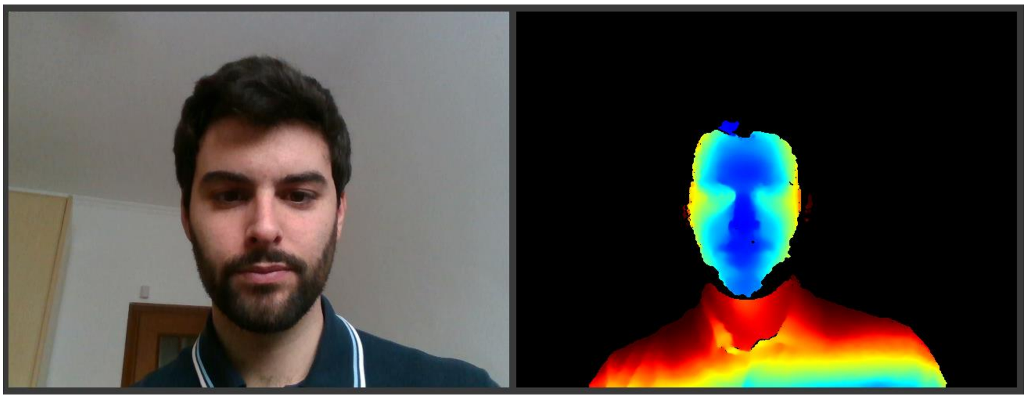


Figure 4.2 - Color frame (on the left) and depth frame (on the right)

Images are also associated with another coordinate system, that allows to identify *point coordinates* in real space, with distance values expressed in meters. The origin [0, 0, 0] refers to the center of physical imager. The positive x-axis points to the right, the y-axis towards down, and the z-axis towards forward (Figure 4.3).

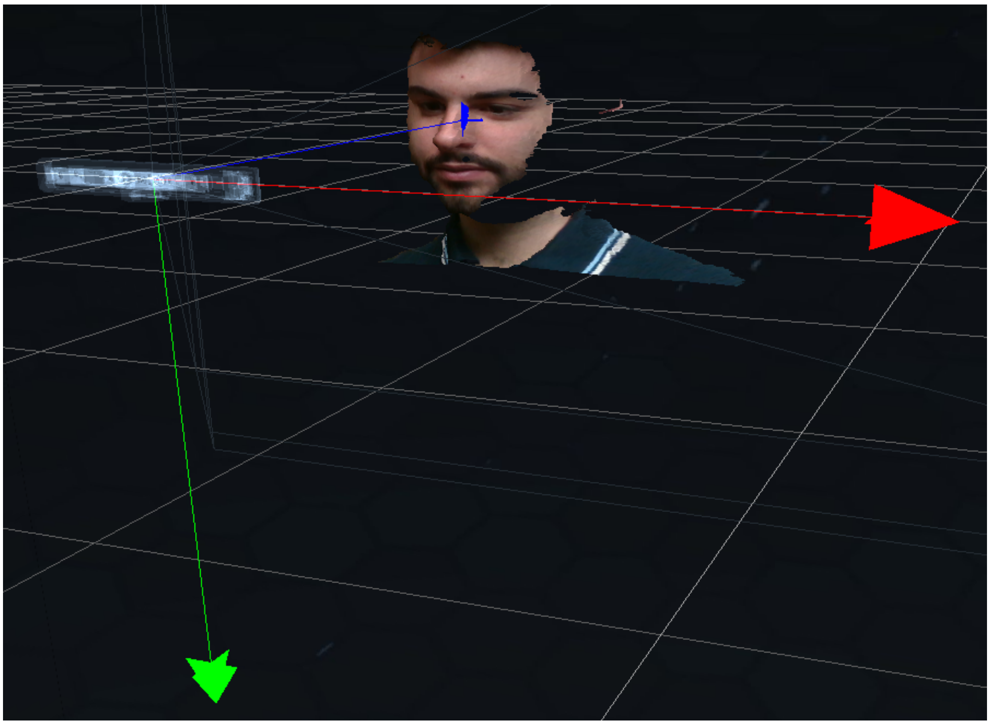


Figure 4.3 - Textured point cloud in real space

The mapping between pixel coordinates and point coordinates can be done knowing intrinsic camera parameters. These parameters vary according frame size and frame rate and can be accessed through the advanced mode. Once that all the parameters are available, two mapping operations can be performed: projection and deprojection. Projection maps a 3D coordinate space point to a 2D coordinate space pixel location. Deprojection maps a 2D coordinate space pixel to a 3D coordinate space point location.

Intrinsic camera parameters are:

- *Width* and *height*: row and column numbers within a frame of the video stream.

- The *focal length*: it is the distance between the optical center of the lens and the focal point of entering ray lights, emitted from a point located at infinity. In other terms it is a value that measures how strongly the lens converges or diverges the light. It is expressed by *fx* and *fy*, which can be slightly different and are multiple of pixel width and pixel height.

- The *center of projection*: *ppx* and *ppy* are the x-coordinate and the y-coordinate. This information is essential since they do not necessarily match with the center of the image, even if they are usually very close to it.

- The *model* describing the distortion introduced in the video streams with the related distortion *coefficients* (up to five). The model employed in Intel RealSense SR300 is the inverse Brown-Conrady, which provides a closed-form formula to map distorted points to undistorted points. Quite the opposite, iterations or lookup tables are required to map undistorted points to distorted points.

These parameters have been introduced because they are essential in the alignment process, the first main step in the data processing.

4.2 Methods

The developed procedure to perform face expression recognition in real-time aims to perform the following operations: color to depth alignment, holes region filling, face detection (performed on depth map), cropping, resizing (Figure 4.4).

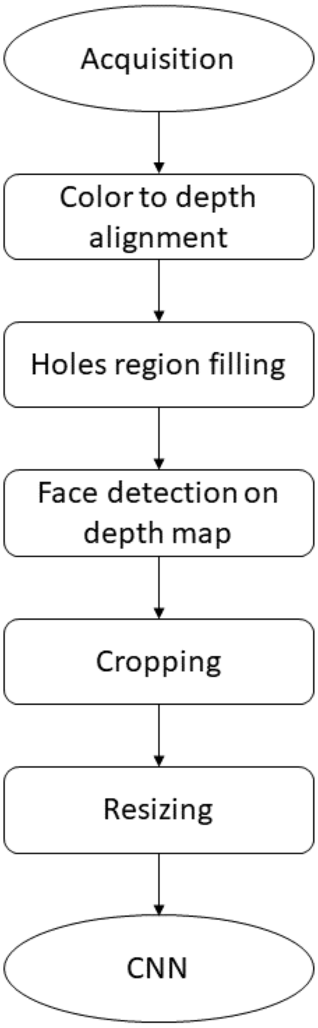


Figure 4.4 - Data processing steps

4.2.1 RGB to Depth Alignment

Infrared camera and color camera are very close each other on the depth module of Intel RealSense SR300, but they are not coincident (Figure 4.5).

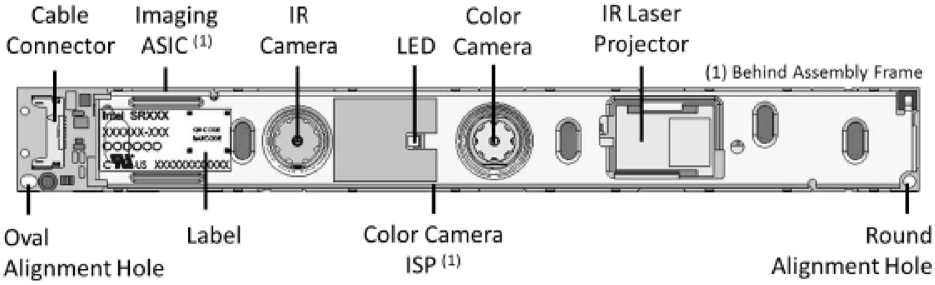


Figure 4.5 - Intel RealSense SR300 depth module [187]

This disparity leads to have color and depth frame misaligned, even if points of view of the two cameras, i.e. color and depth, only slightly differ from each other.

In order to have the images aligned, an alignment process is inevitable and involves two operations: projection and deprojection.

- Projection allows to select a point in the coordinate space, representing the real world, and to map this point to a pixel location on the 2D stream, which could be the color one or the depth one.
- Deprojection makes available the opposite operation; indeed, through the deprojection, it is possible to map the location of a 2D pixel, belonging to the color frame or to the depth frame, into its location in the 3D space.

The alignment process can be performed in two directions:

- The depth frame is aligned to the color frame
- The color frame is aligned to the depth frame

In both cases, the operations order is the same.

As shown in

Figure 4.6, deprojection is performed first, to map a 2D pixel belonging to a frame (color or depth) into its corresponding 3D point into the space (real-world); then, the point is projected from the real-world system to the other frame (depth if the previous frame was color or vice versa), to identify the correct dual pixel.

In other words, in case of the color to depth alignment, each color pixel from the color frame is transformed so that it matches the corresponding pixel in the depth frame.

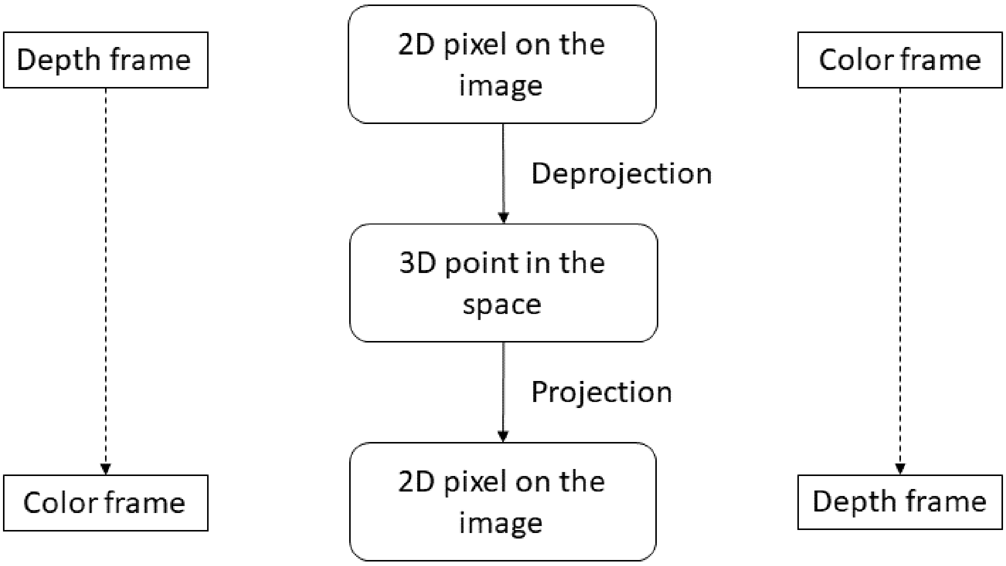


Figure 4.6 - Operation order for the alignment process

In this research, since the alignment of depth to color frames results in a loss of resolution, the chosen alignment direction has been color to depth, namely each color frame is aligned to the corresponding depth frame.

Result of this transformation is shown in Figure 4.7.

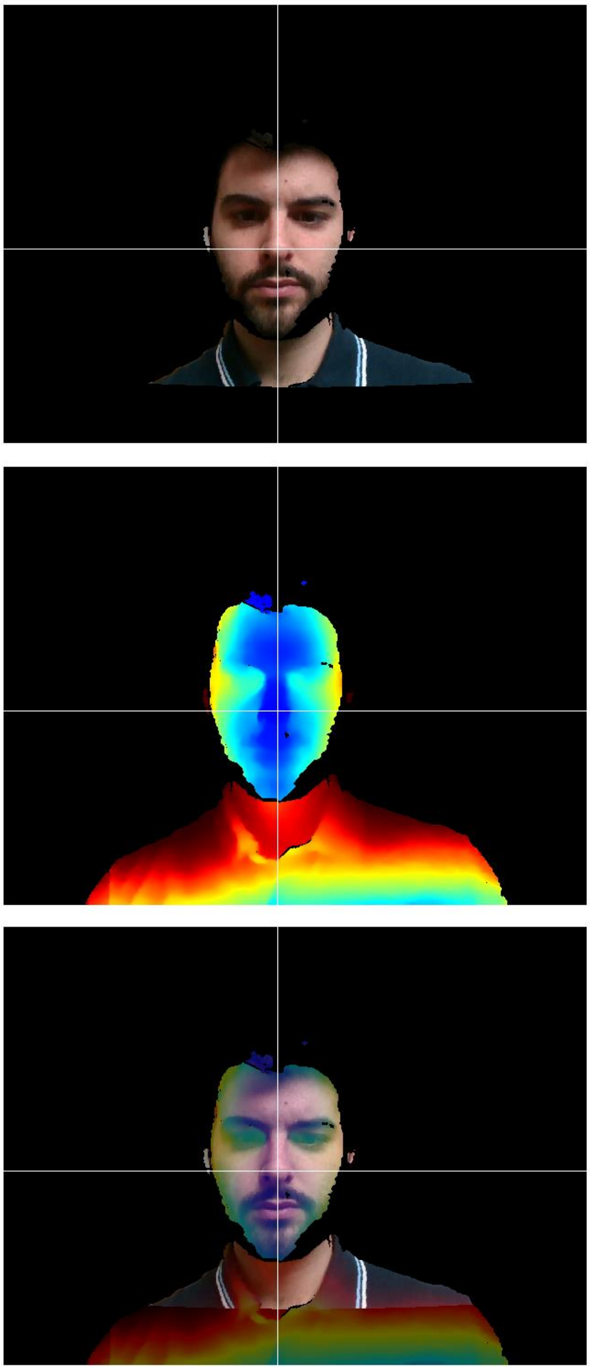


Figure 4.7 – The first image shows the color frame after the alignment with the depth frame color frame; the second image shows the depth frame; the third image shows the aligned color frame and depth frame overimposed to highlight the pixel-by-pixel correspondence between RGB and Depth information

From a comparison with Figure 4.2, it is evident that depth frame is not modified by the color to depth alignment process. Only the pixels belonging to the color frame have been changed, so that every pixel in both frames refer to the same point in the space (one carrying the color information, the other carrying the depth information).

Visually, the main difference is the removal of the background in the color frame. This is due to the fact that no information about the background is present in the depth frame, since the structured light depth acquisition technology has a limited operational range, so pixels in the color frame have no corresponding pixels in the depth frame (the background is displayed black, which means no information available for this part of the image).

This is an advantage, because aligning depth frames to color frames means also to discard some information not useful for further processing; indeed, the data processing aims to provide only the oval shape of the face to the neural network. This way, a first delimitation of the region of interest (ROI) is performed.

4.2.2 Holes region filling

The coded light depth acquisition technology sometimes fails in identifying all the point distances from the camera, even if they are within the operation range. The problem is ascribable to the pattern projected onto the scene and becomes evident in presence of surfaces that absorb or twist infrared wavelength, such as beard, eyelashes, eyebrows, and hair.

In the case small holes are present on the depth map, it is possible to resort to some techniques to partially restore the image. Due to the holes nature that are usually small details, among the other techniques to average not corrupted pixels has provide good results in covering these holes and to obtain a graceful degradation.

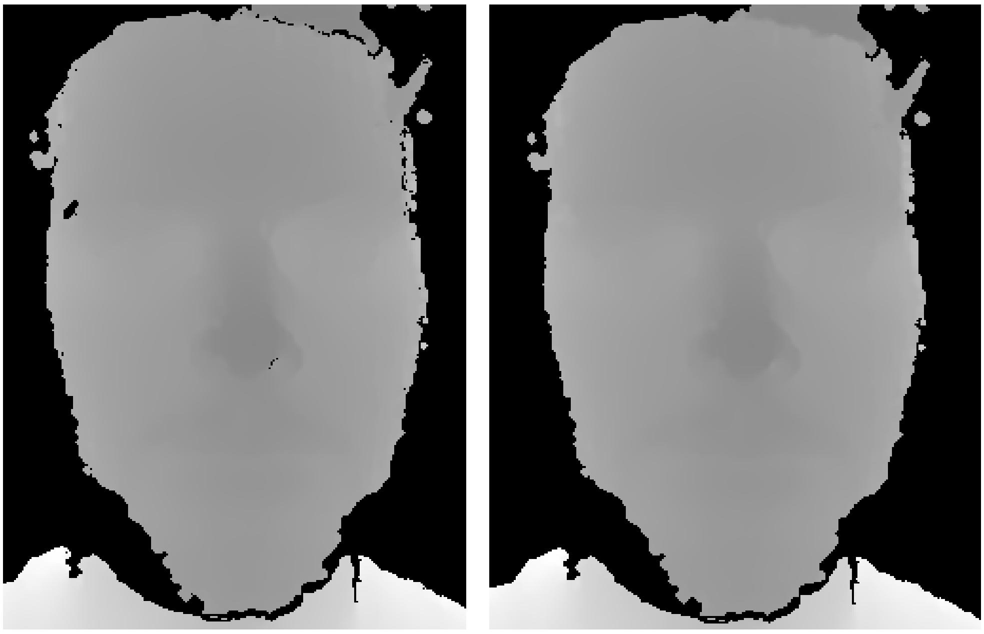


Figure 4.8 – On the left the depth map with holes, on the right holes are filled with the information provided by the neighboring pixels. Depth maps are zoomed to better visualize the holes.

It has been noticed that these artifacts are more critical in manual feature extraction methods than deep learning techniques. For instance, CNN focuses on specific areas of the face, while an SVM method for face expression recognition could need the Euclidean distance between specific landmarks, hence those specific facial points should always be present within the frame. If one on these landmarks is missing, the result is unpredictable.

4.2.3 Face detection on depth frame

Within this procedure, face detection is the operation that allows to identify the ROI on the depth frame. The ROI is composed by all the points belonging to the face and is identified by using a pixel clustering technique: the k-means.

The rough idea behind this implementation is to partition the image in some clusters and to identify the cluster to which facial points belong according to some pretty strong presuppositions, valid in the context of a controlled experiment, which is our situation since this part of the research is ongoing.

The main steps of face detection on the depth map are shown in Figure 4.9.

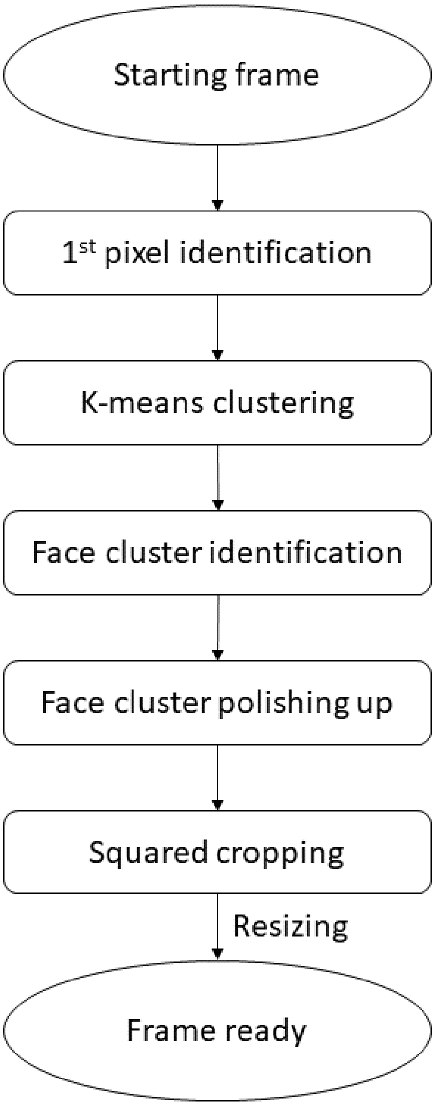


Figure 4.9 - Data processing: face detection on depth frame

To better explain every decision taken to perform face detection on a depth map, it is useful to consider an example, for instance Figure 4.10.

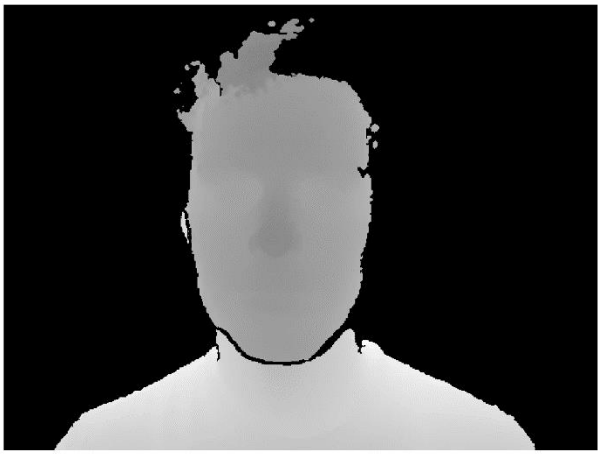


Figure 4.10 - Face detection: starting depth frame

Every depth frame of a subject in front of the camera in a controlled environment can be similar to the image above: the background, identified by the *0* value, and the foreground with pixels of different values identified by different grayscale levels.

Initially, the frame must be analyzed from left to right and from top to bottom. The first pixel not equal to *0*, i.e. belonging to the foreground, must be stored.

Secondly, k-means clustering subdivides the pixels of the depth image in three clusters. The choice of *3* as number of clusters refers to the three main area that can be distinguished on the depth map: the background, the face and the area that includes the neck and the chest. The result is shown in Figure 4.11.

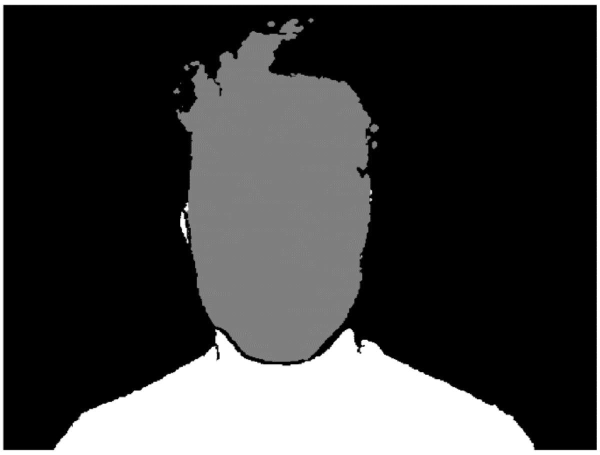


Figure 4.11 - Face detection: clustered depth frame

Since clusters label are randomly assigned to the clusters, to identify the label assigned to the face it is necessary to get back the location of the first pixel that has been memorized as first operation. The value of that pixel in the clustered image is the value of all the pixels belonging to the subject's face.

The success of this face detection algorithm depends on too strong assumptions, even in a controlled environment, so a polishing up step has been introduced to improve robustness. An analysis of the possible scenarios during which the algorithm could run into criticalities has led to consider the situation in which the subject is further from the camera (Figure 4.12).

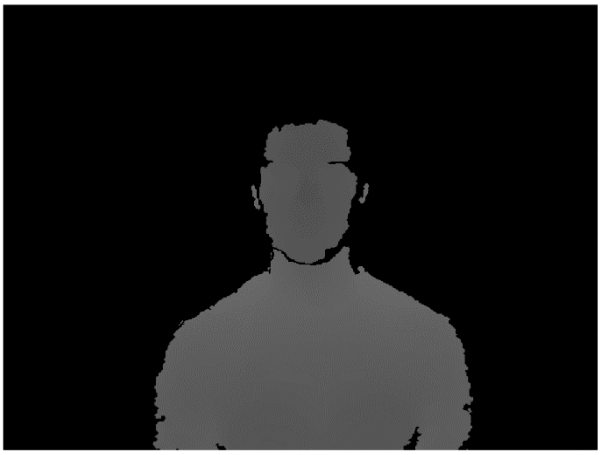


Figure 4.12 - Subject positioned far from the camera

In this example, other pixels not referring to the face are part of the same cluster (Figure 4.13).

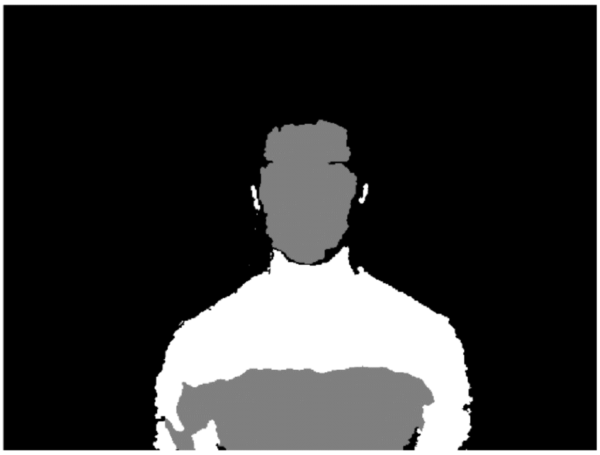


Figure 4.13 - Face detection: lower pixels labelled as facial pixels

Pixels referring to the lower part of the chest have been labelled with the same values of the facial pixels. This could be a common situation when the subject is further than 40 cm from the camera.

The solution to this issue is the following: firstly, only the cluster to which the face belongs must be considered, i.e. the cluster to which both the connected components identified by the upper region (face) and the lower region (lower chest) belong, as shown in Figure 4.14. The cluster to which the shoulders area belong and the background are discarded.

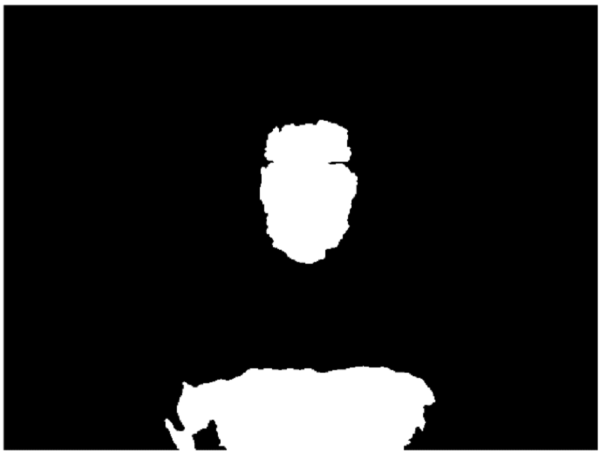


Figure 4.14 - Face detection: all the pixels belong to the same cluster, the face's one

Secondly, the proper connected component must be identified: this can be done selecting the connected component to which the first pixel that has previously memorized belong, which is the upper one.

Final images referring to the examples just explained are shown in Figure 4.15.

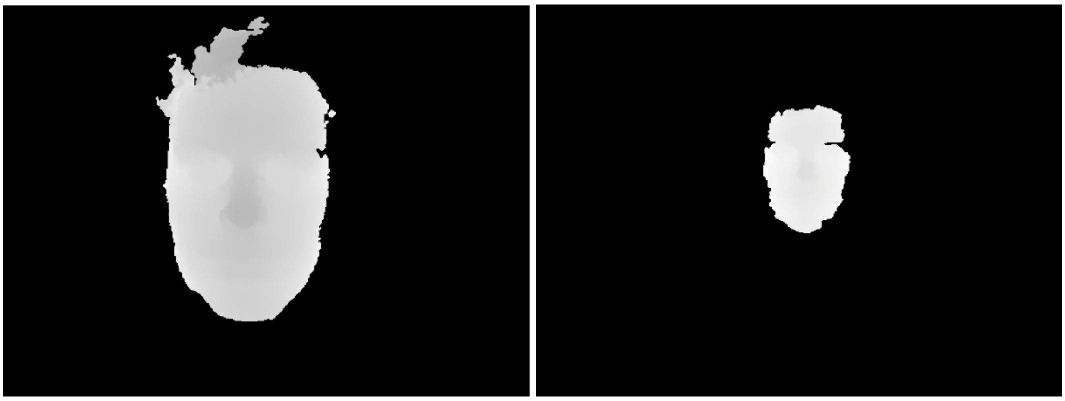


Figure 4.15 - Face detection: final images with detected faces. On the left the final result regarding the subject close to the camera, on the right the final result regarding the subject far from the camera

Since all the images are aligned, it is possible to retrieve the pixels describing the face on the depth frame and also on the color frame. Once that the oval of the face has been identified, both the depth frame and the corresponding, aligned color frame are properly cropped so that a global square shape for the image has obtained.

The final data processing step is to resize the obtained images in order to fit the input requirements of the VGG16 CNN, that expects 224x224 images. This is the reason why the cropping has been done square-shaped.

If the neural network changes, the image dimensions will change accordingly.

4.3 Convolutional Neural Network

VGG16 CNNs have been used for performing face expression recognition, namely, to implement the last step of the procedure.

In order to understand strength and weaknesses, three architectures have been set up. A VGG16 CCN requires 224x224 images on three channels so the first neural network deals with RGB images (channels: Red, Green and Blue components of the image); the second one deals with depth images (channels are the Z component replicated three times); the third architecture is composed by two VGG16 CNNs, one dealing with color images, one with depth images and combined the results together to take advantage of a multimodal approach.

Output of these CNNs is the probability of membership of the data to seven categories: anger, disgust, fear, happiness, sadness, surprise, and neutrality, namely the six basic emotions and the neutral state.

At present, neural networks have been trained using the public database BU-3DFE. This database contains 2500 images of people performing facial expressions at different levels of activation, uniformly distributed.

The 2500 images have been subdivided as follows: 1800 images have been used for the training step, 450 for the validation and 250 for the testing.

4.4 Results and discussion

Regarding the real-time constraint, an evaluation of the procedure has been done, to understand which are the most expensive steps from the computational time point of view.

A first scenario has considered the implementation of these processing steps: color to depth alignment, face detection on depth map, cropping, resizing, and face expression recognition. Results are reported in Table 4.1.

Table 4.1 – Frame rate computation considering all the designed data processing steps

Operation	Frame rate
Acquisition	30 FPS
RGB to depth alignment	30 FPS
Face detection on depth map (k-means)	9 FPS
Cropping	9 FPS
Resizing	8 FPS
Face Expression Recognition (CNN)	4 FPS

The acquisition frame rate is set to 30 FPS. The alignment step does not show to decrease this value since it is not a bottleneck for the procedure. Quite the opposite, the face detection algorithm has a heavy toll on framerate, decreasing it up to 9 FPS; at current state, this strong degradation is due to the fact the k-means algorithms, core of this step, has been implemented with a routine that run on a single-thread. The cropping phase is not a demanding operation, indeed there is no further degradation in performing this step.

A consideration must be done in evaluating the resizing operation; both depth and color frames have been acquired at a 640x480 resolution and the target size for each acquired frame is 224x224. If the face detection and the consequently cropping are performed, frame size become closer to the target size and the resizing operation is not computationally expensive. The result is a very slight downgrade in the performances (from 9 to 8 FPS).

The CNN has a strong impact, so that the overall performances using all these steps is smaller than 4 FPS.

In order to respect the real-time requirement selected as constraint (5 FPS), another scenario has been considered, in which the face detection on depth frame is set aside (Table 4.2).

Table 4.2 - Frame rate computation with limited data processing

Operation	**Frame rate**
Acquisition	30 FPS
RGB to depth alignment	30 FPS
Cropping	30 FPS
Resizing	22 FPS
Face Expression Recognition (CNN)	9 FPS

In this scenario, the real-time constraint is largely respected, achieving an overall performance of 9 FPS.

Nothing has changed for the alignment operation. Frames must be properly dimensioned to fit with the input required by the CNN. For this reason, the frame is directly cropped to a trade-off squared size (from 640x480 to 400x400) and resized to fit with 224x224. This time, since the difference between the cropped frame and the target frame is higher, resizing operation results more demanding and the downgrade stands at 22 FPS. The implementation of the CNN leads to an overall performance of 9 FPS.

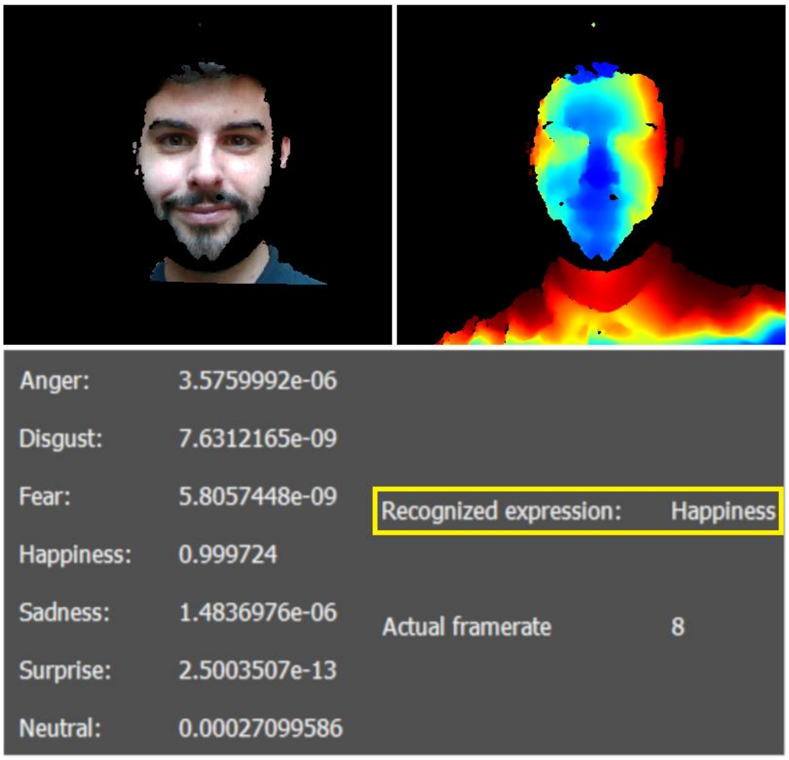
The main drawback of this scenario is that the user must keep his/her position in front of the camera. This way, the ecological validity of the results could be compromised, hence the first scenario is preferable.

Regarding the holes filling, the operation is necessary in those situations during which depth frames are corrupted due to the technological limit introduced by using structured light. This step can be considered not essential in the context of a real-time application, during which the next frame is available soon and the user can move in front of the lens to meet the need of the sensor, allowing to compute the results with little adjustments so that the pattern can be projected on the whole surface of interest. It goes without saying that in case of extremely corrupted depth maps, more drastic solutions are inevitable, such as to change RGB-D cameras in situation where the short-range requirement is not respected anymore.

CNNs used in these trials have been trained with images belonging to the BU-3DFE database. As seen in the previous chapter, BU-3DFE has a limited number of images, so transfer learning techniques must be applied outside the context of BU-3DFE database.

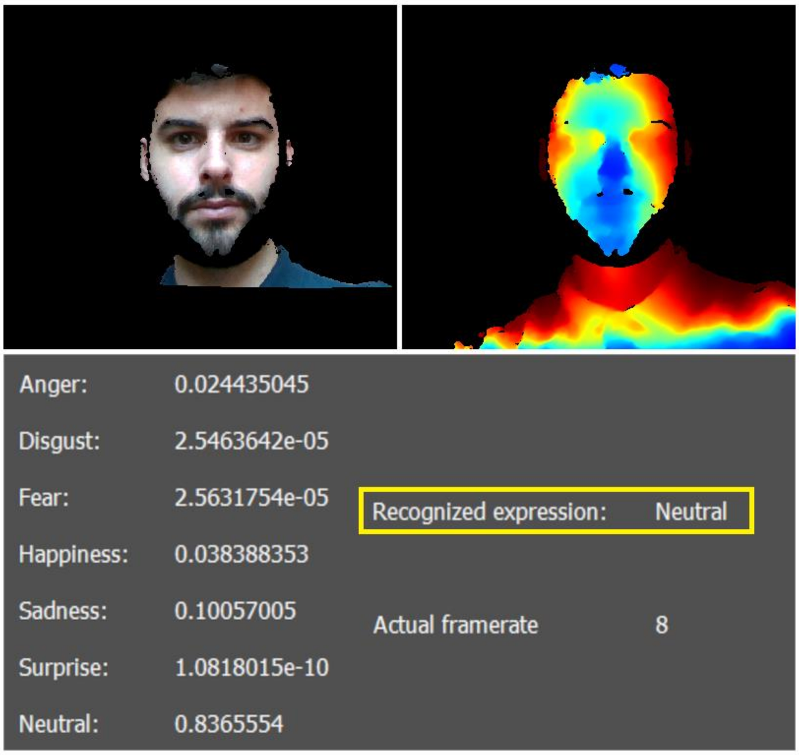
An analysis within the context of the BU-3DFE have been conducted, splitting the 2500 images in three groups: 1800 images for the training, 250 images for the validation, and 450 images for the testing phase. The limited number of images used for the training is balanced by using images provided by the same database and the recognition rates are here discussed: the RGB-only neural network has reached the 76.8%; the Depth-only neural network has reached the 78.8%; the RGB-D architecture the 79.2%. These results support the theory according to which the usage of both color and depth information can improve overall performances. Another expected confirmation has arrived outside the context of BU-3DFE: face expression recognition results more complicated when negative expressions have to be recognized.

Even if CNN has been trained using a database with a limited number of images, accentuated happiness and neutrality are recognized (Figure 4.16 and Figure 4.17).



```
Anger:        3.5759992e-06
Disgust:      7.6312165e-09
Fear:         5.8057448e-09     Recognized expression:    Happiness
Happiness:    0.999724
Sadness:      1.4836976e-06     Actual framerate          8
Surprise:     2.5003507e-13
Neutral:      0.00027099586
```

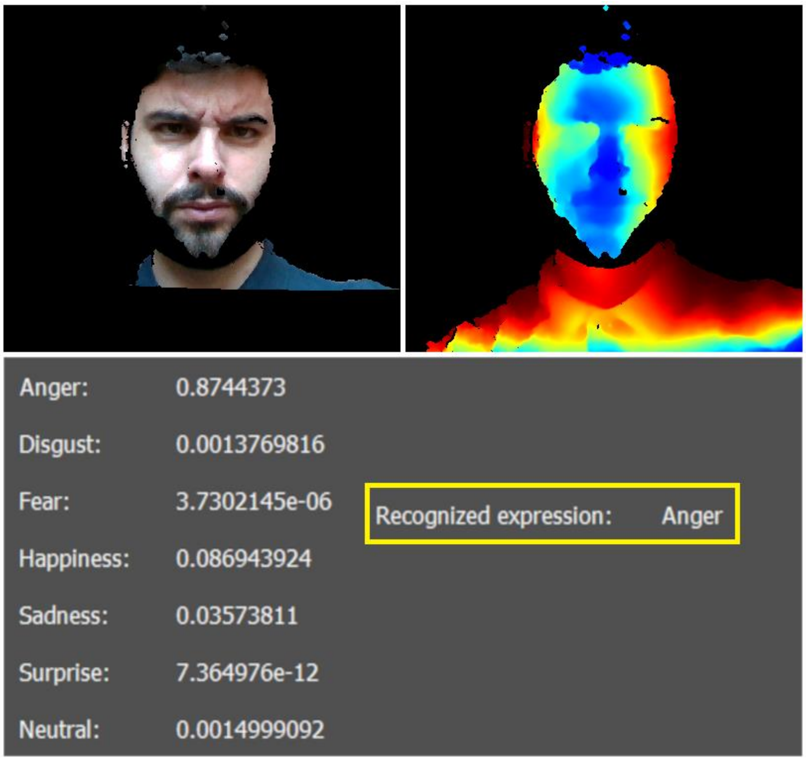
Figure 4.16 - Happiness recognition



```
Anger:        0.024435045
Disgust:      2.5463642e-05
Fear:         2.5631754e-05    Recognized expression:    Neutral
Happiness:    0.038388353
Sadness:      0.10057005       Actual framerate          8
Surprise:     1.0818015e-10
Neutral:      0.8365554
```

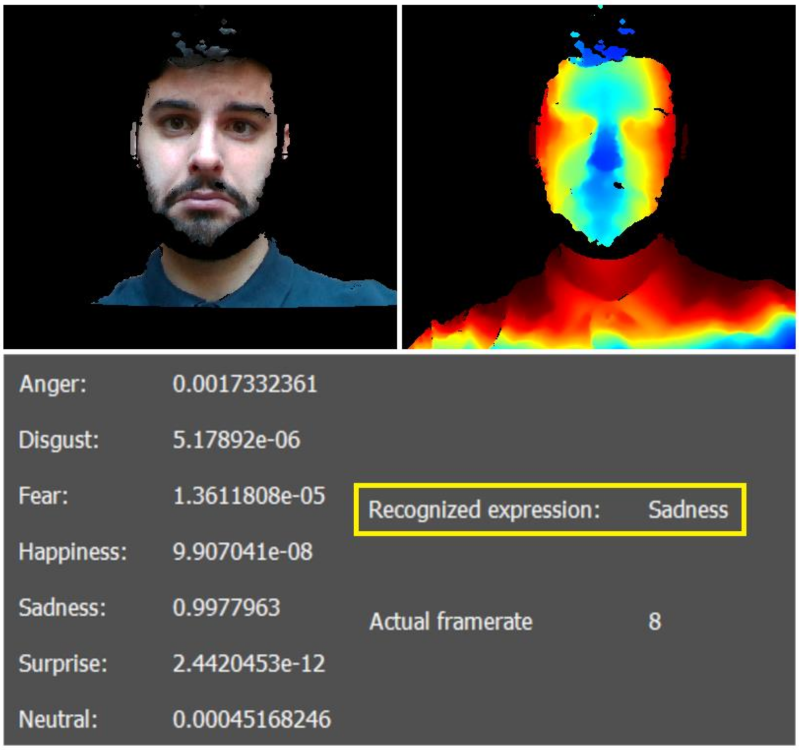
Figure 4.17 - Neutral expression recognized

This is consistent with the analysis conducted in the previous chapter, according to which negative emotions, in particular anger, disgust and fear (Figure 4.18), have similar valence and arousal values, so the difficulties in identifying those emotions increase; sadness is the least critical emotion to be recognized (Figure 4.19).



```
Anger:        0.8744373
Disgust:      0.0013769816
Fear:         3.7302145e-06    Recognized expression:    Anger
Happiness:    0.086943924
Sadness:      0.03573811
Surprise:     7.364976e-12
Neutral:      0.0014999092
```

Figure 4.18 - Anger recognition



Anger: 0.0017332361

Disgust: 5.17892e-06

Fear: 1.3611808e-05

Happiness: 9.907041e-08

Sadness: 0.9977963

Surprise: 2.4420453e-12

Neutral: 0.00045168246

Recognized expression: Sadness

Actual framerate 8

Figure 4.19 - Sadness recognized

These considerations are translated into the need of more training images regarding negative expressions, while for happiness images this need is less strong.

All the trials have been done on a system equipped with Intel Xeon E-2186M. The CPU specs has been reported since GPU potentialities have not been explored yet.

The application developed to implement data processing steps and neural network testing is reported in Appendix D.

The procedure is ongoing and currently has been set up only in its preliminary version, hence the results can be strongly improved and must be interpreted in this light.

4.5 Conclusions and future work

RGB-D cameras continuously evolve and could provide the opportunity to obtain better quality starting data and to fit more scenarios. For instance, a more conspicuous usage of active stereoscopy in order to increment the operational range is highly predictable; this is the reason why the application developed for the real-time procedure is fully compatible with Intel RealSense D435, the current state-of-art Intel active stereoscopy camera.

However, the awareness of a technological improvement must be supported by an adequately algorithm development, which is the core of such an application proper functioning.

At the current state, the steps that have been implemented up to now have been useful to focus on specific issues. The final aim must be not to sacrifice any step to provide to the CNN the best data possible and to preserve the quality of the data acquired with the RGB-D camera, so the first scenario drawn in the previous section should be considered, i.e. the implementation of all the following steps: color to depth alignment, face detection on depth map, cropping, resizing, and face expression recognition.

Computational time performances will be improved optimizing the code to complete the data processing steps. The focus will be moved on the face detection algorithm since it is the main responsible of the frame rate downgrade. More specifically, an efficient way to run the k-means operation should be fine-tuned through a multi-thread approach.

Accuracy can be improved mainly with a substantial increment of training images for the CNN. From this perspective, the building of the ecological dataset debated in the previous chapter is considered essential, not only to increment the number of images (an empirical rule is that every class should be trained with at least 1000 images, even if this number depends on a variety of parameters among which the application final goal and the difficulties in recognizing the membership to a certain class), but also to use spontaneous, more realistic expressions and, consequently, more quality images.

Chapter 5

Perspective Morphometric Criteria for Facial Proportion Assessment

The human face is the type of data which this whole work is focused on. Indeed, face analysis is a key step for gaining a full understanding of the framework of this thesis.

In particular, the study behind this chapter has been carried on in collaboration with a research group of EURECOM, in France, during a visiting period. The focus of this work has been moved on faces' proportions, to make the most of the know-how of both research groups: French group was familiar with female attractiveness [188] and facial applications [189], while the Italian group has conducted several studies regarding 3D features on human faces, in particular on geometrical descriptors and landmarking [190].

Furthermore, the idea for this research project has also been conceived to make a socially responsible contribution to this field. In past months, a partnership with the maxillofacial department of Molinette hospital in Turin has led to review all findings linked

This work can be addressed to those studies that need the development of an average face in the context of facial recognition and facial expression recognition but could also be helpful in the medical field, in particular to facial reconstruction surgical interventions due to pathological problems. Very general guidelines for tissue reconstruction after a surgical intervention exist, but they are rarely a valuable auxilium, and physicians need more accurate indications. Obviously, the reconstruction target is to obtain a good-looking face, and the present work has been directed on that need. Nonetheless, it should be emphasized that for this kind of applications surgeons' experience is irreplaceable and their supervisions will always be essential.

This part of the work has been published in Ulrich et al. [191].

5.1 Introduction

Beauty and proportion of human face have always been object of interest through the years, as evidenced by Greek sculptures, ancient Egyptian paints and even in prehistory [192]. Some studies on the topic have been carried on during the Renaissance period by well-known artists, such as Leonardo da Vinci with his

Vitruvian man [193] and Michelangelo Buonarroti, which studied human anatomy deeply [194], but it is from the 20th century that systematic studies focused on objective assessments began.

Several studies about facial aesthetics focused on specific aspects: Jefferson [195] has claimed that an ideal facial proportion exists, regardless population, age and gender and it is related to the concept of divine proportion, Schmid et al. [196] have built an index based on neoclassical canons, symmetry and golden ratios to compute face attractiveness, Baker et al. [197] have affirmed that divine proportion is not sufficient to plan surgical operations on faces, Holland et al. [198] have strongly criticized the mask representing ideal facial archetype derived by Stephen Marquardt from golden ratio and approved by some cosmetic surgeons, showing how it leads to a model of masculinized European women, while Pallett et al. [199] have found quantitative values that claimed to be *new golden ratios* as opposed to the traditional one. Furthermore, Zhang et al. [200] have investigated beauty and proportions of Chinese male and female faces via automatic shape analysis, demonstrating that, under certain limitations, the more beautiful a face is, the closer it is to an average face shape, the same conclusion that have reached Edler et al. [201], broadening the concept to the different populations to which an individual belongs and also Valenzano et al. [202] have found that attractiveness and averageness are strongly correlated.

Cultural influence on beauty assessment plays an important role, as testified by different canons that have been adopted through the years to evaluate female facial proportions. During the Paleolithic period statuettes of *Venus* were sculpted in such a way that they looked full-figured to symbolize fecundation, fertility, and regeneration [203]. Ancient Egyptians considered a large forehead and well-defined mandibles attractive [204], whereas Greeks preferred an oval facial shape for both men and women, and a forehead as small as possible to highlight the hair [205]. In the middle age, there is evidence of a preference for larger foreheads and absence of wrinkles, even if the sign at the time was not negatively considered, as testified by contemporary positive reflections on grey hair [206]. Cultural differences can also be found within the same period. In recent times, debates over femininity depicted by the media have been widely discussed [207], suggesting that beauty relies on the eye of the beholder; nonetheless, there are several experiments suggesting that beauty is assessed through quantitative tips, even if hidden and not directly perceived, especially regarding the face. The presence of a strong objective component in beauty assessment has been clearly showed, among the others, by the following experiments: Iliffe [208] asked 4355 readers of a London newspaper to rank 12 women's pictures taken in the same conditions, showing that the results were extraordinarily similar among all the answers even if people differed by gender, age and origin within United Kingdom, Udry [209] expanded this survey, asking more than 100000 American people to rank the same photographs reaching the same conclusions; in particular, the three top choices were exactly the same both for the British and the American set of people and the ranking difference considering the other choices were very limited. Furthermore, Cunningham's [210] asked male subjects to evaluate 50 female photographs and, in parallel, 24 facial features of the same photographs were computed; results have confirmed the correlation between feature measurements and attractiveness.

Sforza et al. have focused on analyzing databases of attractive and common individuals, from the identification of facial esthetic canons in Italian children in the deciduous and early mixed dentition [211] to the soft-tissue analysis of adolescent boys' and girls' faces [212]. There have also been studies on people of non-Caucasian ethnicity, such as Jayaratne et al. [213].

Several works show comparisons between *normal* and *attractive* women, the gender most widely studied in literature. The *normal* term has been used to identify common, non-selected women, whereas *attractive* is typically used to identify good-looking women, typically chosen among actresses, such as Ferrario et al. [214] or beauty contest participants, such as Sforza et al. [215] and Galantucci et al. [216]. Results appearing in these studies have once more confirmed the presence of objective elements defining attractiveness related to the concept of proportions between different parts of the face in terms of Euclidean distances and angular measurements.

The medical field is one of the most interested discipline in studying this topic, since some branches of surgery must intervene directly on face by modifying the shape both for merely aesthetical and pathological reasons [217]; therefore, it is not surprising that the study of two orthodontists, Peck and Peck [218], was one of the first works aimed to discover and gather facial features. Furthermore, over the last two decades there has been an incremental increase of 3D imaging, such as MRI, CT [219], and also of stereophotogrammetry, which has been used by Plooij et al. [220] and Deli et al. [221] to acquire and to reproduce soft tissues of the face, and 3D modelling, which one example is the creation of the virtual patient by Kau [222].

Before the advent of 3D tools, face operation planning evaluations relied on two-dimensional images acquired on sagittal, coronal, and axial planes [223], and a quantitative analysis of proportions was critical [224]. Geometrical descriptors [225] and landmarks [226] proved to be effective tools to study human face proportion, since they allow to gather those common traits that everybody share. They have provided the possibility of analyzing point clouds reproducing patients' faces for diagnosis, Hammon et al. [227] and Nanda et al. [228] are two examples of this application, and surgical intervention planning, for which handbooks such as Proffitt [229] are nowadays widely used, but they have also allowed to build virtual faces from scratch, as testified by Fan et al. [230]. Nevertheless, medical field is not the only interested field, indeed Average Face Models (AFMs) can be used in face analysis applications as preprocessing step to align faces [231], but can also be employed to improve robustness of face recognition and face expression recognition algorithms; indeed, Dagnes et al. [232] have introduced a new geometrical descriptor called personal Shape Index which highlights differences between a face and the average face derived from a set of 100 neutral faces belonging to Bosphorus database.

Some databases of human faces have been built to be analyzed and to provide new suggestions for further feature extraction and proportions studies, but also to validate results already obtained. Two available databases are the BU-3DFE [233] and the Bosphorus [234], employed in the development of the current research work, which provides faces belonging to more than one hundred people in various poses, expressions and different types of occlusions.

The present work gathers facial measures that have been identified in previous studies as representative of Caucasian female face attractiveness with the purpose of classifying Caucasian female faces in terms of proportions. After that, the Bosphorus, meaning a database of *normal* women, has been used to verify if the final set of canons is suitable and sufficient for the proportions evaluation of female faces.

Results confirm that selected measures evaluation gets close to human's assessment, providing the opportunity of quantitatively analyzing Caucasian women's facial proportions; moreover, a ranking showing the influence of each measure for Caucasian women has been drawn up. The importance of ratios between measures and the higher relevance of the vertical measures compared to the horizontal have been highlighted.

5.2 Materials and Methods

Face analysis is the discipline that studies human faces based on the identification of landmarks, specific points common to everyone that can be identified on the face. Landmarks can be recognized on the hard tissue through palpation or on the soft tissue through observation, even if some of the landmarks positioned on the soft tissue depend on landmarks positioned on the hard tissue. In order to evaluate women's proportions, an expanded set of measures, relying on landmark positions, has been defined.

As seen in the previous section, significant experiments proving the presence of a strong objective component have been conducted, nonetheless the most difficult step in evaluating female beauty is to identify a ground truth that allows to compare measures of women's faces in terms of proportions. In the past decades, several works aiming to establish which are the human face traits that influence an observer's assessment on proportions have been conducted. Works considered as the most incisive in this field, thus taken into greater account in the present work, have been carried out by Farkas et al. [235], Ferrario et al. [214], Sarver et al [229], Sforza et al. [215] and Galantucci et al. [216].

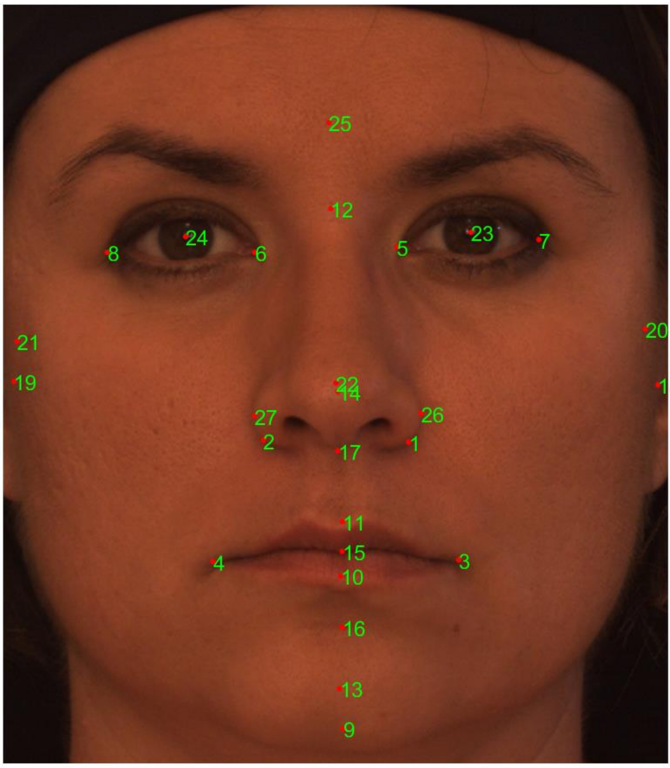
Each of those works has been carried out live-positioning landmarks on female subjects before the acquisition of the point cloud required to compute measures. Manual allocation directly on subjects has been chosen in order to achieve the best accuracy possible. Acquisitions have been made using RGB-D camera, so that both color image and depth information are stored for the further processing steps. Subjects were actresses, participants to beauty contests and common women; the latter have been chosen to validate the results.

All the information found by those studies have been gathered to obtain an expanded set of measures able to evaluate women's attractiveness; then, the Bosphorus database has been used to validate the expanded set of measures comparing it with qualitative evaluations issued by human observers. There are 110 subjects in Bosphorus database, but only the 44 women have been analyzed in this work.

The landmark framework considered in this work is reported in Table 5.1 and shown in Figure 5.1.

Table 5.1 – Landmark framework. The third column reports a description for each landmark [236].

Landmark	**Abbreviation**	**Description**
Alar curvature point	ac	Point located at the facial insertion of each alar base.
Cheilion	ch	Point located at each labial commissure.
Endocanthion	en	Soft tissue point located at the inner commissure of each eye fissure.
Exocanthion	ex	Soft tissue point located at the outer commissure of each eye fissure.
Gonion (or Menton)	gn (or me)	Most inferior midpoint on the soft tissue contour of the chin.
Labiale inferius	li	Midpoint of the vermilion line of the lower lip.
Labiale superius	ls	Midpoint of the vermilion line of the lower lip.
Nasion	n	Midpoint on the soft tissue contour of the base of the nasal root at the level of the frontonasal suture.
Pogonion	pg	Most anterior midpoint of the chin.
Pronasale	prn	Most anterior midpoint of the nasal tip.
Stomion	sto	Midpoint of the horizontal labial fissure.
Sublabiale	sl	Most posterior point on the labiomental soft tissue contour that defines the border between the lower lip and the chin.
Subnasale	sn	Midpoint on the nasolabial soft tissue contour between the columella crest and the upper lip.
Tragion	t	Point located at the upper margin of each tragus.



Landmark number	Landmark symbol
1	ac_l
2	ac_r
3	ch_l
4	ch_r
5	en_l
6	en_r
7	ex_l
8	ex_r
9	gn
10	li
11	ls
12	n
13	pg
14	prn
15	sto
16	sl
17	sn
18	t_l
19	t_r
20	zy_l
21	zy_r
22	M(t_l-t_r)
23	pupil_l
24	pupil_r
25	g
26	al_l
27	al_r

Figure 5.1 – Figure showing landmarks studied in the current work. Landmarks 20 (zy_l), 21 (zy_r), 22 (midpoint of Tragi), 25 (g), 26 (al_l) and 27 (al_r) have been discarded due to lack of available measures involving them. Suffixes _l and _r state that the landmark considered is respectively the left or the right one.

All the measures considered in this analysis have been acquired employing the above-mentioned landmarks and can be subdivided into three categories: linear, angular and ratios.

Linear measures described in Table 5.2 and shown in Figure 5.2 are Euclidean distances between two landmarks or between a landmark and another specific point. More specifically, one of those specific points is the point on the *E-line* that minimizes the distance with *Labiale superius* (or *Labiale inferius*), where the *E-line* is the line passing through the *Pronasal* and the *Pogonion*.

Table 5.2 - Euclidean linear distances. The third column reports the work from which the measure has been taken.

Measure	Description	References
n-pg	Facial line	Ferrario et al. [214], Galantucci et al. [216]
n-sn	Anterior upper facial 2° third height	Ferrario et al. [214], Galantucci et al. [216]
ch_r-ch_l	Oral length	Farkas et al. [235], Sarver et al. [229], Galantucci et al. [216]
ex_r-ex_l	Upper facial width	Ferrario et al. [214], Sforza et al. [215], Galantucci et al. [216]

sn-pg	Anterior lower facial height	Galantucci et al. [216]
t_r-t_l	Middle facial width	Galantucci et al. [216]
ls-(prn-pg)	Upper lip to E-line distance	Galantucci et al. [216]
li-(prn-pg)	Lower lip to E-line distance	Sforza et al. [215], Galantucci et al. [216]
ls-li	Vermilion height	Galantucci et al. [216]
en_r-en_l	Intercantal distance	Farkas et al. [235], Sarver et al. [229]
ac_r-ac_l	Width nose base	Farkas et al. [235], Sarver et al. [229]
n-gn	Facial height	Farkas et al. [235], Sarver et al. [229]
sn-gn	Lower third facial height	Farkas et al. [235], Sarver et al. [229]
ls-sto	Upper vermilion	Farkas et al. [235], Sarver et al. [229]
li-sto	Lower vermilion	Farkas et al. [235], Sarver et al. [229]

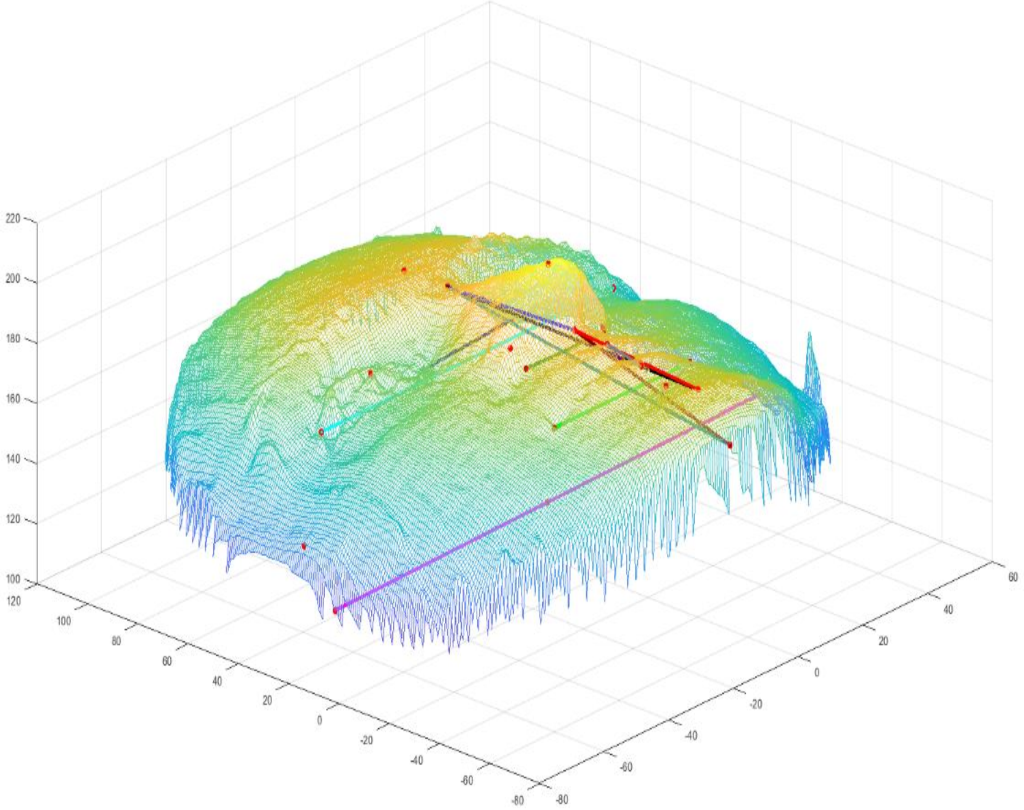


Figure 5.2 - Euclidean distances

Angular measures described in Table 5.3 and showed in Figure 5.3 are angles subtended by a vertex identified by three landmarks or, exceptionally for the

Interlabial distance, by two lines lying on the same plane and identified by four landmarks, two for each line.

Table 5.3 - Angular measures. The third column reports the work from which the measure has been taken.

Measure	Description	References
n-sn-pg	Facial convexity excluding the nose	Ferrario et al. [214], Galantucci et al. [216]
sl-n-sn	Maxillary prominence	Ferrario et al. [214], Galantucci et al. [216]
prn-sn-ls	Nasolabial	Farkas et al. [235], Sarver et al. [229], Galantucci et al. [216]
n-prn-pg	Nasion – Pronasal - Pogonion	Galantucci et al. [216]
ex_l-n-ex_r	Left Exocanthion – Nasion - Right Exocanthion	Sforza et al. [215]
pg-n-ls	Maxillo-facial angle (mf)	Galantucci et al. [216]
en_l-n-en_r	Left Endocanthion-Nasion-Right Endocanthion	Ferrario et al. [214]

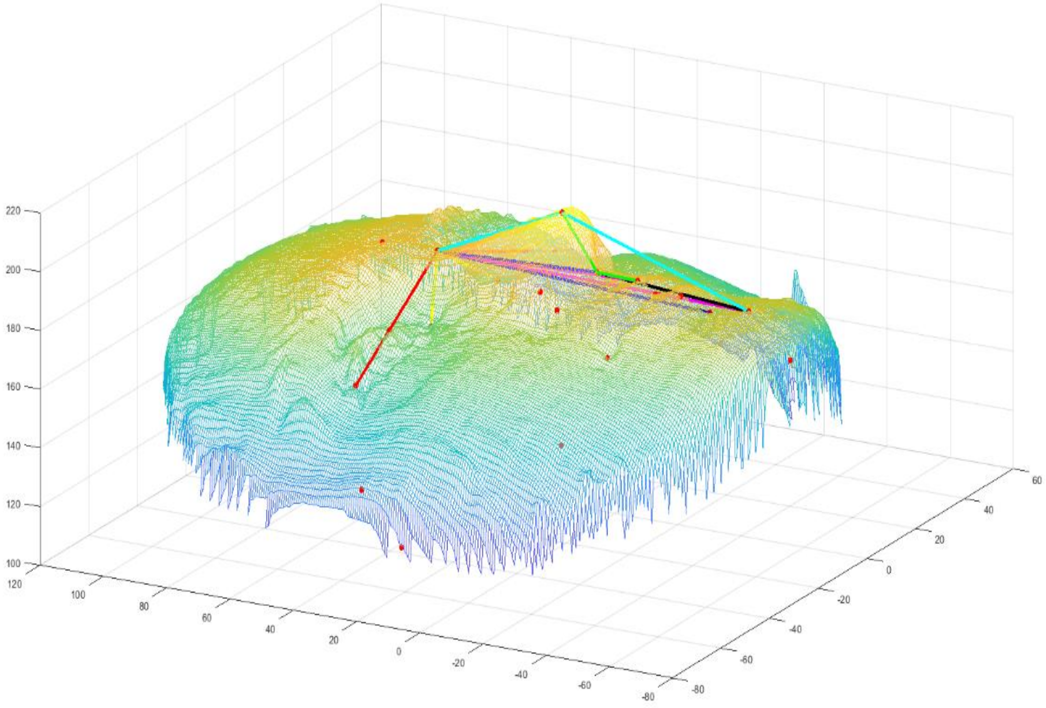


Figure 5.3 - Angular measures

Ratios between linear distances (Table 5.4) allow to perform quantitative evaluations about proportions. Face analysis moves the focus from the local to the global point of view, since not only the absolute value of one single measure is considered, but rather the overall effect of two measures. Intuitively, ratios are the quantitative way to represent the big picture.

Table 5.4 - Ratios of Euclidean distances. The third column reports the work from which the measure has been taken.

Measure	Description	References
(t_r-t_l)/(n-pg)	Middle facial width to facial height	Galantucci et al. [216]
(n-sn)/(n-pg)	Nasion - Subnasale / Nasion - Pogonion	Ferrario et al. [214], Galantucci et al. [216]
(sn-pg)/(n-pg)	Subnasale - Pogonion / Nasion - Pogonion	Ferrario et al. [214], Galantucci et al. [216]
(t_r-n)/(t_r-sn)	Right Tragi-Nasion / Right Tragi-Subnasale	Galantucci et al. [216]
(sn-pg)/(n-sn)	Lower to upper facial height	Galantucci et al. [216]
(sn-gn)/(n-gn)	Lower third / facial height	Farkas et al. [235], Sarver et al. [229]
(sto-gn)/(sn-gn)	Mandibula / lower third	Farkas et al. [235], Sarver et al. [229]

As a result of the literature review, twenty-nine measures have been identified. The present work employs data from the Bosphorus database, namely women's pictures, 3D models and relative landmark coordinates. Because all the studies previously cited are carried on by research groups with different expertise, slightly different sets of landmarks have been adopted. A landmarking expert identified some missing landmarks on 2D pictures and on 3D models on the Bosphorus database, in order to complete the landmark framework. Nonetheless, some landmarks had to be discarded because they relied on the hard tissue and the only way to identify them was through palpation. Thus, it has not been possible to include some measures into the expanded set. An example is the *zygion* (*zy*), which is the most lateral point on the soft tissue contour of each zygomatic arch.

The works of Galantucci et al. [216], Farkas et al. [235], Sarver et al. [229], Ferrario et al. [214] and Sforza et al. [215] provided mean value and standard deviation for every measure. Some measures are common to different sources, even if related mean value and standard deviation are slightly different depending on the study. Indeed, one of the purposes of the present study is to merge the information coming from these different sources (Figure 5.4) and build an overall measuring methodology.

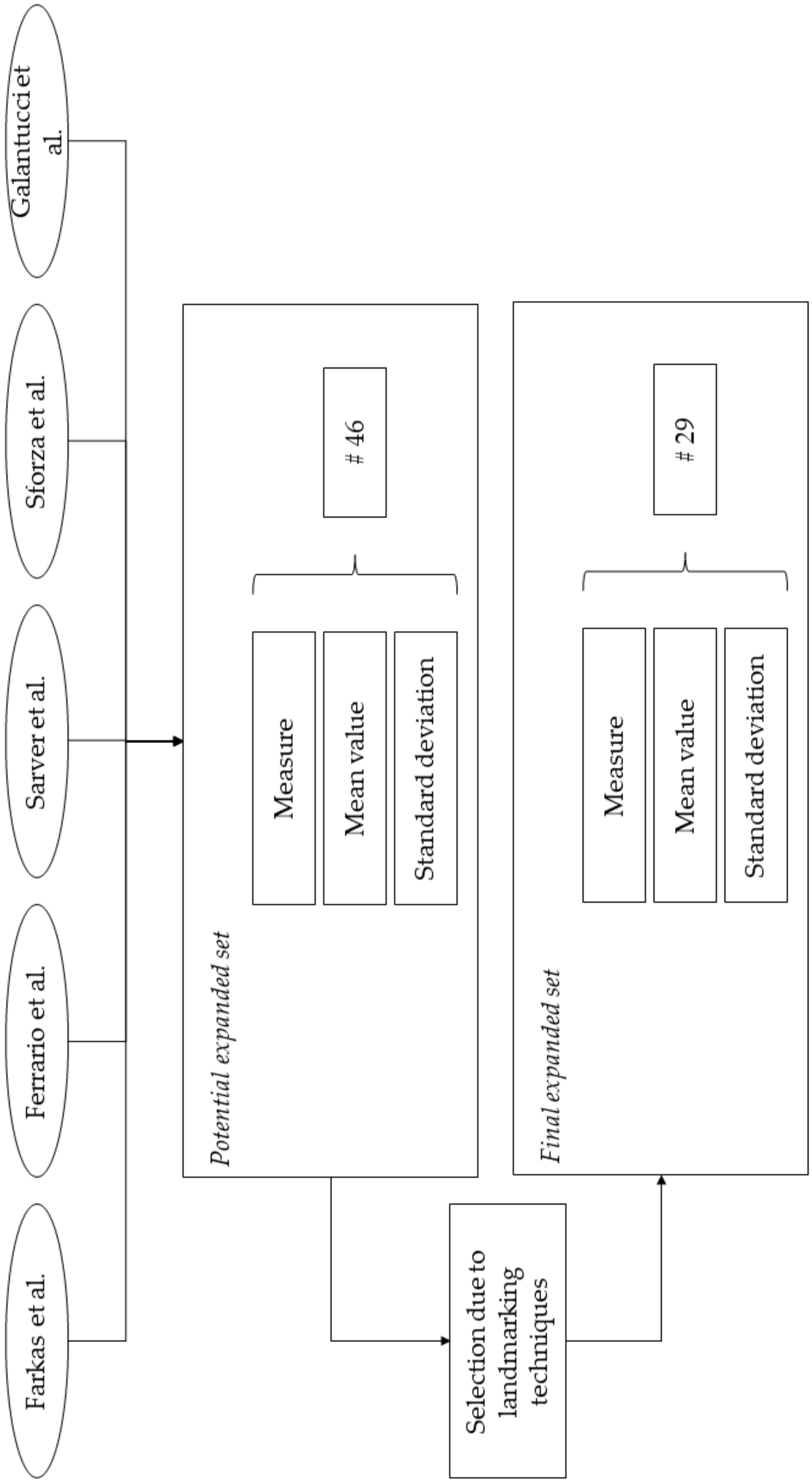


Figure 5.4 - Definition of the expanded set of measures.

A score, result of sum of penalties, has been computed for each woman present in the Bosphorus database. For each measure, if the value of the considered woman was within the range *mean value ± standard deviation,* $\mathrm{mean value} \pm \mathrm{standard deviation}$ no penalty has been added. Conversely, if the value was out of range, a penalty has been added, and the amount of penalty (1) has been computed as the ratio between the distance of the measure from the mean value normalized with the mean value.

$$\mathrm{penalty} = \mathrm{weight} * \frac{|\ \mathrm{current\ woman's\ measure} - \ \mathrm{mean\ value}\ |}{\mathrm{mean\ value}} \qquad (1)$$

In the case of more than one mean value, since there is more than one source in literature that refers to the same measure, the so computed penalty value has been then multiplied for a weight (2).

$$\mathrm{weight} = \frac{\#\ \mathrm{women\ in\ current\ study}}{\#\mathrm{women\ in\ all\ the\ studies\ referring\ to\ the\ same\ measure}} \qquad (2)$$

This weight has been introduced to consider the different degree of confidence assigned to different studies found in literature. Remembering that in those studies canons have been extracted from sets of attractive women, it has been considered essential to evaluate more robust the analysis with a greater number of subjects. Thus, the weights have been computed as the ratio between the number of women involved in a single study and the sum of all the women involved in all the studies related to the same measure. For instance, consider a hypothetic measure identified by two studies A and B, which involves x and y women, respectively. The weight related to the study A will be x/(x+y), while the weight related to the study B will be y/(x+y).

After all the scores have been obtained, a cluster analysis has been performed through the usage of k-means methodology, subdividing the dataset in five classes. The purpose of this step was to identify which were the women closer to the well-proportioned standard face in terms of compliance with the measures in the expanded set; in other words, considering the descending order adopted, faces belonging to class 5 are closer to the well-proportioned standard face than the faces belonging to class 1. The number of classes has been chosen to have some correspondence with the Likert scale. Indeed, in parallel, a qualitative evaluation of the women present in Bosphorus has been made by a focus group using precisely the Likert scale. A comparison between the results obtained using the developed methodology and the qualitative evaluation performed by human observers has been performed (Figure 5.5).

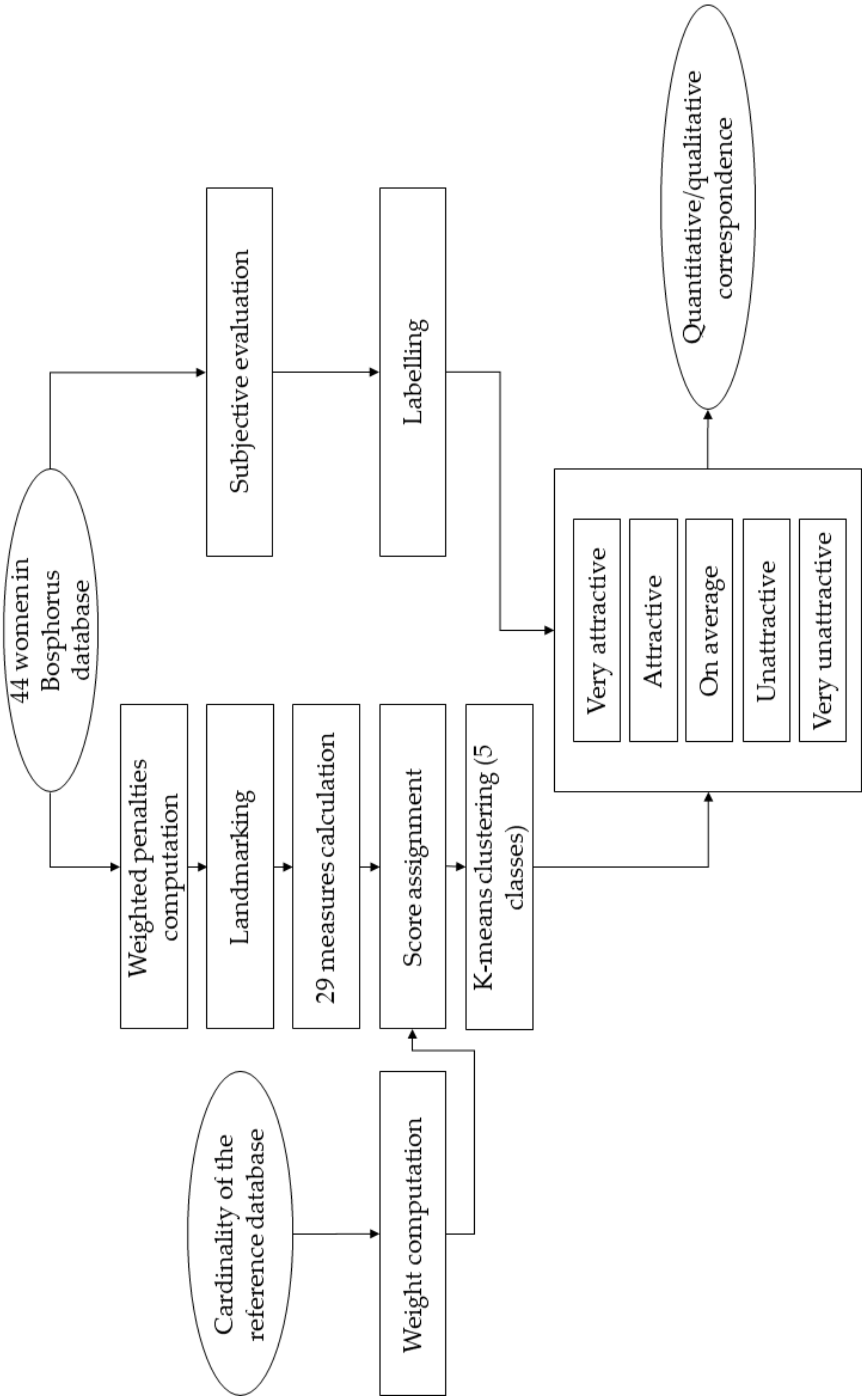

44 women in Bosphorus database
Subjective evaluation
Labelling
Weighted penalties computation
Landmarking
29 measures calculation
Score assignment
K-means clustering (5 classes)
Cardinality of the reference database
Weight computation
Very attractive
Attractive
On average
Unattractive
Very unattractive
Quantitative/qualitative correspondence

Figure 5.5 – Methodology to compare quantitative and qualitative evaluation.

Finally, the results of the developed methodology in terms of cluster subdivision have been analyzed to rank the influence of each measure of the expanded set in women's facial proportion assessment.

5.3 Results

The measures selected from literature and included into the expanded set are reported in Table 5.5. The table is set up so that there is a measure for every row and mean values and standard deviations on the columns. Since it is possible to have more than one mean value and one standard deviation due to the possibility of retrieving the same measure from different sources in literature, the last column reports the reference number of the considered source.

Table 5.5 - - List of the final set of 29 measures. Reference [24] analyzes two different beauty contests (Miss Italia 2006 and Miss Italia 2007) that provide slightly different values. "2006" and "2007" has been added in some references to distinguish between those two contests.

Measure	Mean value	Standard deviation	Reference
n-pg	97.34	4.03	[216]
	99.06	5.54	[214]
n-sn	50.29	2.43	[216]
	52.38	2.76	[214]
ch_r-ch_l	45,7	2.77	[216]
	50	3.2	[214]
ex_r-ex_l	84,01	2.98	[216]
	101.59	4.97	[214]
	95.2	3.3	[215] – 2006
	92.5	3.5	[215] – 2007
sn-pg	48.14	2.65	[216]
t_r-t_l	131.52	4.19	[216]
ls-(prn-pg)	3.95	1.8	[216]
li-(prn-pg)	2.32	1.31	[216]
	2.4	1.4	[215] – 2006
	1,7	1.3	[215] – 2007
ls-li	17.97	2.29	[216]
en_r-en_l	32	2.4	[229] [235]
ac_r-ac_l	31	1.9	[229] [235]
n-gn	112	5.2	[229] [235]
sn-gn	66	4.5	[229] [235]
ls-sto	8.4	1.3	[229] [235]
li-sto	9.7	1.6	[229] [235]
n-sn-pg	163.55	4.37	[216]
	164.02	3.71	[214]
sl-n-sn	9.33	1.98	[216]

	8.48	1.72	[214]
prn-sn-ls	123.12	9.53	[216]
n-prn-pg	131.12	4.07	[214]
ex_l-n-ex_r	124.9	3.5	[215] – 2006
	120.8	4.2	[215] – 2007
pg-n-ls	7.93	2.1	[216]
en_l-n-en_r	120.29	5.24	[214]
(t_r-t_l) / (n-pg)	1.35	0.06	[216]
(n-sn) / (n-pg)	0.52	0.02	[216]
	0.5341	0.017	[214]
(sn-pg) / (n-pg)	0.49	0.02	[216]
	0.4659	0.017	[214]
(t_r-n) / (t_r-sn)	0.96	0.02	[216]
(sn-pg) / (n-sn)	0.9588	0.0600	[216]
(sn-gn) / (n-gn)	0.586	0.029	[229] [235]
(sto-gn) / (sn-gn)	0.691	0.028	[229] [235]

After the expanded set definition, each of the forty-four women in Bosphorus database has been measured and the scores have been computed summing all the penalties. Scores are subdivided into five clusters, so that it is immediate to identify women's faces closer to the well-proportioned standard face, remembering that clusters are arranged in descending order, namely faces belonging to cluster 5 are the closest to the well-proportioned standard face, conversely faces belonging to cluster 1 are the furthest.

Hereafter, a focus group has met up to evaluate Bosphorus women's attractiveness. The final aim was to compare the developed methodology with the human judgment. The results showed that the cluster label and the Likert's scale label are the same, namely the difference between those two values equals to 0, for 15/44 women (very high correspondence); the difference is 1 for 16/44 women (high correspondence); the difference is 2 for 9/44 women (moderate correspondence); the difference is 3 for 3/44 women (low correspondence); the difference is 4 for 1/44 women (very low correspondence).

Very high correspondence means that the result of the method matches the focus group's outcome (a woman with *very high correspondence* label has been judged in the same way both by the method and by the focus group).

All those results are reported in Table 5.6.

Table 5.6 - Scores, cluster labels, qualitative evaluation labels and correspondence between quantitative and qualitative evaluation for each woman.

Subject	Score	Cluster label	Qualitative evaluation label	Correspondence
1	0.7608	5	4	High
2	0.7868	5	3	Moderate
3	5.9095	1	2	High

4	2.2549	3	3	Very high
5	3.3508	2	2	Very high
6	0.8030	5	4	High
7	2.8187	2	2	Very high
8	1.3236	4	3	High
9	0.7395	5	3	Moderate
10	1.0387	4	4	Very high
11	2.9316	2	1	High
12	1.5750	4	3	High
13	2.6975	2	2	Very high
14	2.0539	3	3	Very high
15	1.4707	4	2	Moderate
16	2.6951	2	1	High
17	1.5603	4	1	Low
18	0.4124	5	4	High
19	0.7826	5	4	High
20	2.2078	3	2	High
21	4.1747	1	1	Very high
22	3.4563	2	2	Very high
23	3.2831	2	2	Very high
24	2.5096	3	1	Moderate
25	1.8256	3	2	High
26	4.5707	1	1	Very high
27	2.3485	3	2	High
28	4.3690	1	3	Moderate
29	1.3455	4	1	Low
30	5.4693	1	1	Very high
31	3.4337	2	1	High
32	4.2277	1	2	High
33	2.9829	2	2	Very high
34	1.4366	4	2	Moderate
35	1.9479	3	1	Moderate
36	2.7541	2	2	Very high
37	0.7544	5	1	Very low
38	1.9969	3	1	Moderate
39	1.7917	3	2	High
40	1.1677	4	1	Low
41	1.4717	4	2	Moderate
42	1.4425	4	3	High
43	3.3575	2	2	Very high
44	2.1768	3	3	Very high

Finally, results obtained through cluster analysis have been more deeply analyzed. The purpose of this step was to rank the measures to understand which are the most important in facial women’s proportion assessment. In particular, the focus has been moved on cluster 4 and cluster 5, which contain *attractive* and *very attractive* women, respectively. Knowing the cardinality of each cluster, that is 17 if clusters 4 and 5 are taken together, values out of range have been computed for each measure; a measure must be considered influential if the smallest possible number of women has that measure out of range. For instance, referring to Table 5.7, it is possible to notice that the Euclidean distance *ls-sto* is the most influential measure because only two women (12%) belonging to cluster 4 and cluster 5 are out of range, conversely *ch_r-ch_l* is one of the least influential measure because only 5 out of 17 women are within range.

Table 5.7 - Measures ranking. Cardinality of each cluster is reported in the header of the table. The first column lists the measures, the second column reports the number of women belonging to cluster 4 or 5 that are out of range. For the sake of completeness, in the other columns the number of women belonging to each cluster has been reported.

Measures	**Cluster 4+5 (#17)**	**Cluster 5 (#7)**	**Cluster 4 (#10)**	**Cluster 3 (#10)**	**Cluster 2 (#11)**	**Cluster 1 (#6)**
ls-sto	2 (12%)	0 (0%)	2 (20%)	4 (40%)	11 (100%)	5 (83%)
(sn-pg) / (n-sn)	2 (12%)	1 (14%)	1 (10%)	6 (60%)	1 (9%)	5 (83%)
n-sn	3 (18%)	2 (29%)	1 (10%)	6 (60%)	3 (27%)	5 (83%)
sn-gn	3 (18%)	0 (0%)	3 (30%)	4 (40%)	11 (100%)	5 (83%)
(t_r-n) / (t_r-sn)	3 (18%)	0 (0%)	3 (30%)	4 (40%)	4 (36%)	3 (50%)
ex_r-ex_l	4 (23%)	1 (14%)	3 (30%)	5 (50%)	6 (54%)	5 (83%)
sl-n-sn	4 (23%)	2 (29%)	2 (20%)	5 (50%)	6 (54%)	2 (33%)
(t_r-t_l) / (n-pg)	4 (23%)	0 (0%)	4 (40%)	4 (40%)	6 (54%)	3 (50%)
n-gn	5 (29%)	0 (0%)	5 (50%)	6 (60%)	6 (54%)	5 (83%)
li-sto	5 (29%)	2 (29%)	3 (30%)	3 (30%)	7 (64%)	5 (83%)
n-prn-pg	5 (29%)	1 (14%)	4 (40%)	5 (50%)	3 (27%)	2 (33%)
(sn-gn) / (n-gn)	5 (29%)	1 (14%)	4 (40%)	7 (70%)	3 (27%)	2 (33%)
en_l-n-en_r	6 (35%)	3 (43%)	3 (30%)	6 (60%)	6 (54%)	4 (67%)
(sn-pg) / (n-pg)	6 (35%)	2 (29%)	4 (40%)	3 (30%)	4 (36%)	2 (33%)
n-pg	7 (41%)	3 (43%)	4 (40%)	7 (70%)	6 (54%)	5 (83%)
pg-n-ls	7 (41%)	0 (0%)	7 (70%)	3 (30%)	6 (54%)	4 (67%)

(n-sn) / (n-pg)	7 (41%)	3 (43%)	4 (40%)	4 (40%)	7 (64%)	3 (50%)
ls-li	8 (47%)	2 (29%)	6 (60%)	6 (60%)	5 (45%)	4 (67%)
en_r-en_l	8 (47%)	4 (57%)	4 (40%)	7 (70%)	4 (36%)	4 (67%)
n-sn-pg	8 (47%)	5 (71%)	3 (30%)	8 (80%)	8 (73%)	5 (83%)
ac_r-ac_l	9 (53%)	3 (43%)	6 (60%)	6 (60%)	9 (82%)	4 (67%)
(sto-gn) / (sn-gn)	9 (53%)	3 (43%)	6 (60%)	8 (80%)	6 (54%)	2 (33%)
sn-pg	10 (59%)	3 (43%)	7 (70%)	3 (30%)	6 (54%)	3 (50%)
ex_l-n-ex_r	10 (59%)	5 (71%)	5 (50%)	6 (60%)	6 (54%)	5 (83%)
prn-sn-ls	11 (65%)	6 (86%)	5 (50%)	9 (90%)	10 (91%)	6 (100%)
ch_r-ch_l	12 (71%)	4 (57%)	8 (80%)	8 (80%)	9 (82%)	6 (100%)
li-(prn-pg)	13 (76%)	4 (57%)	9 (90%)	7 (70%)	5 (45%)	5 (83%)
ls-(prn-pg)	15 (88%)	7 (100%)	8 (80%)	8 (80%)	9 (82%)	3 (50%)
t_r-t_l	16 (94%)	6 (86%)	10 (100%)	7 (70%)	10 (91%)	6 (100%)

5.4 Discussion

Results provided by literature analysis have led to the first outcome of this research, namely the expanded set of measures, each of which characterized by mean value and standard deviation. In literature, a measure is considered relevant for women's facial proportions if a statistically significant number of *attractive* women possesses similar values of the same measure and, conversely, that measure assumes different values in subjects belonging to *normal* woman set. The critical point is to define a ground truth, i.e. a set of measures distinctive for *attractive* women. All the studies carried out in this field agree upon considering *attractive* those women's faces that are commonly positively evaluated in terms of facial proportions, thus famous actresses or beauty contests participants, especially those that move on to the final stage of national competitions.

The present work has focused on gathering measures validated in past studies, building the expanded set of measures defined on facial landmarks and analyzing a public database of *normal* women, the Bosphorus. Unfortunately, some of the landmarks used in literature were not present in the set of data of the Bosphorus database. For this reason, an expert has manually added those missing landmarks lying on soft tissue, but some of the hard tissue landmarks have not been considered due to the impossibility of identifying them without live palpation. Consequently, a limited set of measures that could have been included into the expanded set has been discarded.

In order to analyze the Bosphorus, a methodology able to integrate information provided by different sources was required. Thus, the algorithm based on the penalty mechanism has been developed and the results have been clustered to provide the possibility of classifying faces into 5 different levels. Cluster numbering is from 5 to 1, that means from the most compliance to the well-proportioned standard face to the least. Cluster numbering is chosen this way so that the opportunity of comparing quantitative outcome and qualitative assessment is guaranteed. Likert's 5-level scale has allowed to analyze the correspondence between the developed methodology and people assessment.

Obtained results displayed in Figure 5.6 show levels of correspondence in women's evaluation from very high to very low, namely from a 4-level difference to a 0-level difference between quantitative and qualitative evaluation. The 31/44, that means slightly more than 70%, obtained summing high and very high correspondence, justifies the theories mentioned in the introduction section stating the presence of objective elements that are unconsciously but incontrovertibly considered in evaluating women's attractiveness.

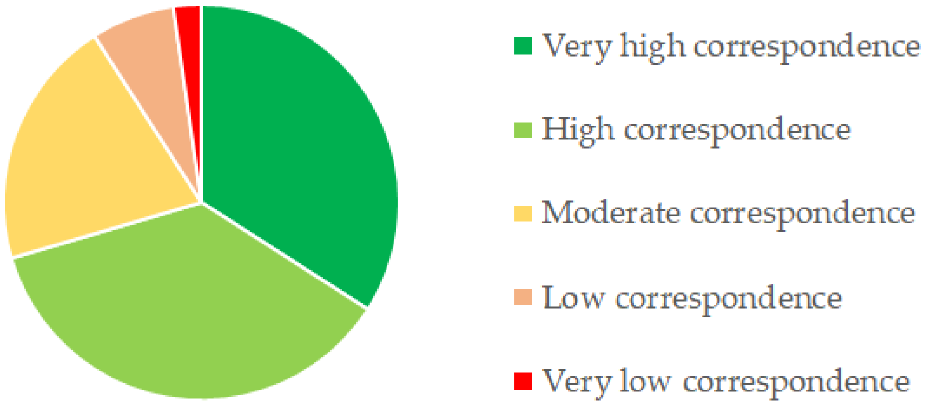


Figure 5.6 - Correspondences between quantitative and qualitative evaluation.

Cluster analysis has led this study to discover part of those elements in terms of proportions between significant measures, ranking them from the most to the least influential. Some interesting observations have arisen from the ranking analysis. Firstly, vertical measures are typically more meaningful than the horizontal: Some examples of influential vertical measures are the thickness of the upper lip (ls-sto), the height of the central part of the face (n-sn), and the height of the lower part of the face (sn-gn). Going deeper into the detail, the lower part of the face in attractive women resulted as higher than the central part, coherently with the fact that if they have identical values, faces appear rounded, a characteristic not considered attractive. The upper lip was more meaningful than lower lip, but this does not mean it should be greater; rather, it means that it is more ordinary to have a full lower lip, while to have also a full upper lip is more uncommon, thus it is a peculiar feature of attractiveness. The angle between the two exterior corners of the eyes and the nasion, i.e., the point which separates the upper third of the face and the middle third of the face, (ex_l-n-ex_r), and the mouth width (ch_r-ch_l) at the end of the table are some examples of less meaningful horizontal measures. This does not retract the importance of mouth width in women's faces assessment; simply, a not-so-relevant difference in mouth width between attractive and normal women has been shown by the analysis. Secondly, beauty turned out to be strictly connected to proportions; 5 out of 7 ratios present in the expanded set of measures are on the top half of the ranking, confirming that women's beauty and well-proportioned faces are also given by relationships between measures, thus they involve a holistic process. Another indication is the relatively poor importance of the facial width absolute value, t_r-t_l, compared with the ratio between facial width and facial height, (t_r-t_l)/(n-pg), ranked in the top 10 most influential measures.

In total, 70% of high and very high correspondence between quantitative and qualitative evaluation is a not neglectable result, but the 30% of non-similarity of the results needs to be investigated in the future research. A bigger set of *attractive* and *normal* women would allow to enlarge the expanded set of measures, as well as the opportunity of live acquiring all the needed landmarks would allow to have uniform data in term of colors, pose and expressions that could be more properly evaluated by humans. Moreover, texture analysis in terms of eye color and skin imperfection could lead to further thin that percentage of non-correspondence and bridge the gap between attractiveness and facial proportions.

5.5 Conclusions

The present study has been carried on considering several experimental evidences proving that beauty assessment is not only subjective but rely on objective elements.

Several previous studies have identified relevant measures to assess facial women's beauty and proportion, comparing sets of *attractive* and *normal* women by measuring Euclidean distances, angular and ratio values; measures have been computed relying on landmarks live identified on subject faces. All these works have considered different sets of measures to find differences in *normal* and *attractive* population.

The current work has defined an expanded set of measures gathering all those information present in literature and a methodology to merge the results and to classify faces has been developed. For each woman, the methodology provides a score, computed as a sum of penalties, and each penalty is given when a woman's measure is not in the range identified by mean value and standard deviation. That procedure has been then used on the set of *normal* women of the Bosphorus public database, and after a cluster analysis it has been possible to classify women's faces considering their distance from the well-proportioned standard face, i.e. the expanded set of measures.

After that, a deeper investigation on results provided by the cluster analysis has permitted to rank the measures from the most to the least influential, to understand which are the most considered measures in women's proportions assessment. Vertical Euclidean distances showed to be very impactful, as well as ratio measures; indeed, the direct comparison between measures resulted to be even more significant than the single measure value, as proved by the facial width. Upper lip thickness proved to be the most meaningful measure in the mouth area.

The present work has been designed to give an important contribution in female facial proportion assessment and aims to be a key point both for further investigations about proportions of human face. Maxillofacial surgery could be an application field for this study, indeed the establishment of a set of guidelines to reconstruct faces affected by pathological problems is strongly requested by physicians to operate at the best of their ability. Another field of application could be face analysis, indeed Average Face Models (AVMs) can be used as preprocessing step to align faces and to improve robustness of face recognition and face expression recognition algorithms

Further work could focus on a greater number of individuals and different subjects in terms of gender, age, and population in order to cover the whole set of possible subjects.

Conclusions

This thesis aimed to develop an automatic procedure to investigate face expression recognition (FER) methodologies using RGB-D cameras.

The research has begun identifying the most suitable depth acquisition technologies to be used for different scenarios. RGB-D cameras are not born to be specifically used for facial applications, nonetheless a strong interest in these sensors has grown during the years, given the great opportunity of placing side by side color and depth information. In the preliminary part of the thesis, coded-light 3D cameras have been identified as the most suitable for short-range applications, hence the Intel RealSense SR300 has been selected to be used for the further experiments.

A first employment of this camera has been the recording of a subject submitted to an interview, to evaluate his emotional activation level through the analysis of his facial expressions. Facial expressions have been analyzed with a Support Vector Machine algorithm set up for that purpose, which features have been selected among Euclidean distances between landmarks and geometrical descriptors of the face. Data considered have been provided by the depth frames recorded during the interview. The developed algorithm has a recognition rate of 81%, improvable by tuning the chosen features and by incrementing the variability and the quantity of training image.

The face expression recognition application in its more canonical sense, i.e. the identifications of emotions, has been faced using a deep learning approach.

From one side, an ecological valid database has been built. The *ecological* term refers to the condition according to which a subject truly feels an emotion and consequently makes an expression without acting or being distracted by boundary conditions. One of the most used methods to achieve this goal is to arouse emotions in subjects through the vision of images belonging to affective databases. 48 combined images of IAPS and GAPED have been used for the purpose, distributed among the following emotions: anger, disgust, fear, happiness, neutrality, sadness. The experiment has proved that some emotions have been correctly elicited in most cases (79% for happiness), while others need a stimulus different from the static visual one, i.e. not images. Literature shows that some negative emotions can often be twisted (fear images have aroused fear in the 27% of cases and disgust in the 26%); hence, future work will provide virtual reality environments designed to arouse specific emotions and to minimize the contamination between different feelings. The virtual reality component should be introduced to increase the effectiveness of the perceived stimulus; indeed, this experiment has proved that as the years go by it is harder and harder to arouse specific emotions using the same stimuli. Furthermore, the experiment has been the opportunity of testing a Convolutional Neural Network trained on the BU3DFE database and tuned by transfer learning technique. In this case, the best result has been obtained using the only-RGB CNN with a 75.02% of recognition rate, downgraded to 72.65% when using both RGB and Depth information. The issue is clearly ascribable to a lack of accuracy in depth frames, since the CNN trained and tested with this kind of information has provided the 55.20% in terms of recognition rate.

In order to face with criticalities encountered using depth frames and also to introduce the real-time constraint, a preliminary automatic procedure has been developed. Steps involved in the procedure can be gathered into acquisition, data processing, and deep learning FER. Regarding the CNN, the future work on the ecological valid dataset aims to solve the weaknesses linked to the training phase, that is limited by the amount of images provided by the BU3DFE database. Regarding the data processing phase, at current state color to depth alignment, cleaning of the acquired frames (holes filling), face detection, cropping, and resizing have been developed to be used in a controlled environment. Future work will be focused on improving and speeding these operations, not to adversely affect the overall procedure in terms of framerate.

This research is addressed to those applications where the recognition of the user's mood is necessary to perform a task accordingly to the aroused emotions.

Human-Computer Interaction could benefit from face expression recognition to improve the safety of a human operator working in close proximity with a robot, adapting its behavior to the human's one; people suffering by affective disorders could be supported by an artificial intelligence tool to understand and to describe emotions; marketing strategies could be better addressed adapting the advertisement contents to the user's state of mind; furthermore, face expression recognition could be used within a product lifecycle management in the beginning of life phase to contribute to the design of a product or a service more into line with users' needs and expectations.

These are only some examples because the opportunity of acquiring data to interpret the users' mood can be extended to all those applications within which an RGB-D camera can be involved.

In last years, a progress in the tools towards the adoption of both color and depth information has been successfully done also in the context of face expression recognition. Nonetheless, overall performances are rarely the best possible due the continuous evolution of acquisition systems and the usage of data to train the classifiers not fully reliable since clear, strong, and spontaneous facial expressions are difficult to be aroused.

This work tries to address these issues by partially and preliminarily bridging the gap.

Appendix A – Geometrical descriptors

Depth maps represent surfaces, which differentiability/derivability is studied by a branch of Geometry called Differential Geometry, and in this context represent human faces, defined as *free-form* surfaces. The human face has no known equation but can be broken down into subdomains traceable to known geometries (cones, cylinders, paraboloids, saddles).

Geometrical descriptors are geometrical features extracted from depth maps, that aim to be representative of the surface and to speed up the processing in facial applications based on feature extraction; furthermore, they should tolerate within-class variations in FR applications and between-class variation in FER applications.

The starting point in the definition of geometrical descriptors are the first, the second and the mixed derivatives of a surface h, that are h_x, h_y, h_{xx}, h_{yy} and h_{xy} (which is equal to h_{yx}) with respect of x and y directions; h is obtained from the depth map provided by the depth camera (Figure 0.1).

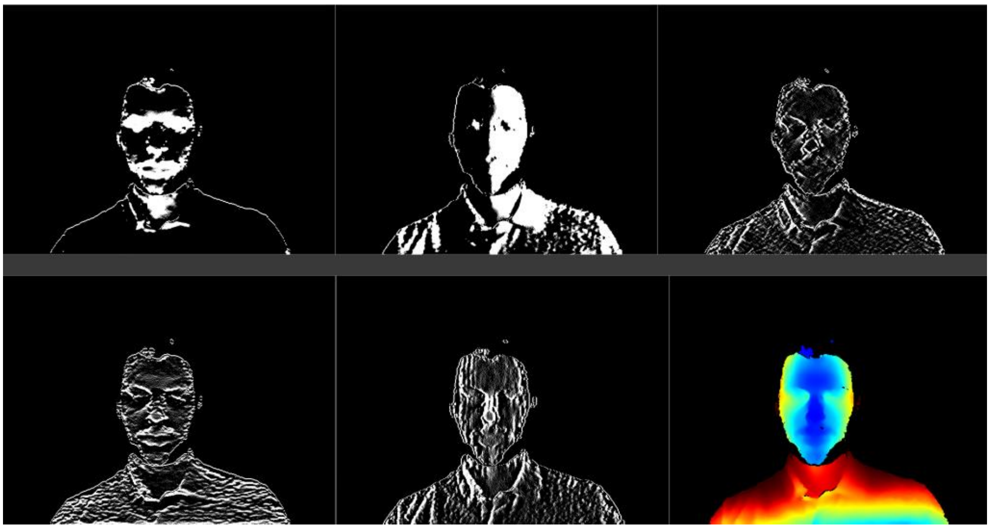


Figure 0.1 - Example of gradients. The depth map representing the surface h is on the bottom-right. In the first line, from left to right: h_x, h_x, h_{xy}. In the second line, from left to right: h_{xx}, h_{yy}, the depth map.

A patch or local surface is a differentiable mapping $x: U \rightarrow \mathbb{R}^n$, where U is an open subset of $\mathbb{R}^2$.

Given that a patch can be written as an n-tuple of functions:

$$x(u, v) = (x_1(u, v), \dots, x_n(u, v)) \tag{1}$$

the partial derivative of x with respect to u can be defined by

$$x_u = \left(\frac{\partial x_1}{\partial u}, \dots, \frac{\partial x_n}{\partial u}\right) \tag{2}$$

The first and second fundamental forms provide the first six descriptors of the set. Their definitions rely on the possibility of measuring distances on surfaces.

In Euclidean space $\mathbb{R}^n$, if $\underline{p} = (p_1, \dots, p_n)$ and $\underline{q} = (q_1, \dots, q_n)$ are points in $\mathbb{R}^n$, then the distance s from $\underline{p}$ to $\underline{q}$ is given by

$$s^2 = (p_1 - q_1)^2 + \dots + (p_n - q_n)^2. \tag{3}$$

Given that a general surface is curved, the distance on it is not the same as in Euclidean space; in particular, the form above is in general false however the coordinates are interpreted.

To describe how to measure distance on a surface, the concept of *infinitesimal* is required. The infinitesimal version of that for $n = 2$ for a surface is

$$ds^2 = Edu^2 + 2Fdudv + Gdv^2 \tag{4}$$

called first fundamental form, or Riemann metric. This is the classical notation for a metric on a surface. E, F, G are functions $U \rightarrow \mathbb{R}$ such that:

$$E = \|x_u\|^2, \tag{5}$$
$$F = \langle x_u, x_v \rangle, \tag{6}$$
$$G = \|x_v\|^2, \tag{7}$$

and are called *coefficients of the first fundamental form*.

These coefficients are given by inner products of the partial derivatives of the surface. Therefore, the first fundamental form is merely the expression of how the surface inherits the natural inner product of $\mathbb{R}^3$.

Geometrically, the first fundamental form allows to make measurements on the surface (lengths of curves, angles of tangent vectors, areas of regions) without referring back to the ambient space $\mathbb{R}^3$ where the surface lies.

To introduce the second fundamental form, the definitions of Gauss map must be given.

For an injective patch $x\colon U \longrightarrow \mathbb{R}^n$ the unit normal vector field or surface normal N is defined by

$$N(u,v) = \frac{x_u \times x_v}{|x_u \times x_v|}(u,v) \tag{8}$$

at those points $(u,v) \in U$ at which $x_u \times x_v$ does not vanish.

The map that assigns to each point p on a surface the point on the unit sphere $S^2(1) \subset \mathbb{R}^3$ that is parallel to the unit normal $N(p)$, or N_p, is called the Gauss Map.

Let $x\colon U \longrightarrow \mathbb{R}^n$ be a regular patch. Then

$$e = -\langle N_u, x_u \rangle = \langle N, x_{uu} \rangle, \tag{9}$$
$$f = -\langle N_v, x_u \rangle = \langle N, x_{uv} \rangle = \langle N, x_{vu} \rangle = -\langle N_u, x_v \rangle, \tag{10}$$
$$g = -\langle N_v, x_v \rangle = \langle N, x_{vv} \rangle \tag{11}$$

are called the *coefficients of the second fundamental form* of x, and $edu^2 + 2fdudv + gdv^2$ is the second fundamental form of the patch x.

Very often a surface is given as the graph of a differentiable function $z = h(x,y)$, where (x,y) belong to an open set $U \longrightarrow \mathbb{R}^2$. It is, therefore, convenient to be provided by formulas for the relevant concepts in this case. To obtain such formulas let us parameterize the surface by

$$x(u,v) = \big(u, v, h(u,v)\big), \qquad (u,v) \in U, \tag{12}$$

where $u = x,\ v = y$. A simple computation shows that

$$x_u = (1,0,h_u), \tag{13}$$
$$x_v = (0,1,h_v), \tag{14}$$
$$x_{uu} = (0,0,h_{uu}), \tag{15}$$
$$x_{uv} = (0,0,h_{uv}), \tag{16}$$
$$x_{vv} = (0,0,h_{vv}). \tag{17}$$

Thus,

$$N(x,y) = \frac{(-h_x, -h_y, 1)}{\sqrt{1 + h_x^2 + h_y^2}} \tag{18}$$

is a unit normal field on the surface, and the coefficients of the second fundamental form in this orientation are given by:

$$e = \frac{h_{xx}}{\sqrt{1+h_x^2+h_y^2}}, \tag{19}$$

$$f = \frac{h_{xy}}{\sqrt{1+h_x^2+h_y^2}}, \tag{20}$$

$$g = \frac{h_{yy}}{\sqrt{1+h_x^2+h_y^2}}. \tag{21}$$

From the above expressions, any needed formula can be easily computed. For instance, the Coefficients of the first fundamental form are obtained [237]:

$$E = 1 + h_x^2, \tag{22}$$

$$F = h_x h_y, \tag{23}$$

$$G = 1 + h_y^2. \tag{24}$$

E, F, G, e, f, and g are six of the twelve primary descriptors. The other six are:

- $\mathrm{K} = \frac{\mathrm{h_{xx}h_{yy} - h_{xy}^2}}{(1+\mathrm{h_x^2}+\mathrm{h_y^2})^2} = \frac{eg-f^2}{EG-F^2}$

  K is the Gaussian curvature and highlights local maximum and minimum points (vertex curvature).
  If a point has K>0 that point is elliptical.
  If a point has K<0 that point is hyperbolic.
  If a point has K=0 and only one of the two principal curvatures is null, that point is parabolic.
  If a point has K=0 and both principal curvatures are null, that point is planar.

- $\mathrm{H} = \frac{(1+\mathrm{h_x^2})\mathrm{h_{yy}} - 2\mathrm{h_x h_y h_{xy}} + (1+\mathrm{h_y^2})\mathrm{h_{xx}}}{(1+\mathrm{h_x^2}+\mathrm{h_y^2})^{\frac{3}{2}}} = \frac{eG-2fF+gE}{2(EG-F^2)}$

  H is the mean curvature and highlights maximum and minimum curvature regions (edge curvature). The behavior of H is smoother than the behavior of K.

- $\mathrm{k_1} = \mathrm{H} + \sqrt{\mathrm{H}^2 - \mathrm{K}}$

  k1 is the first principal curvature and describes the surface inclination variation.

- $k_2 = H - \sqrt{H^2 - K}$

  k2 is the second principal curvature and highlights regions with strong concavity.

- $S = -\frac{2}{\pi}\arctan\frac{k_1+k_2}{k_1-k_2}, S \in [-1,1], k_1 \geq k_2$

  S is the Shape Index [238] and gives an information on what is the reference surface for the point taken into consideration (spherical cup, trough, rut, saddle rut, saddle, saddle ridge, ridge, dome, spherical cap) as shown in Figure 0.2. The number of intervals has been subsequently reduced to 7, merging spherical cup and trough into cup and dome and spherical cap into dome.

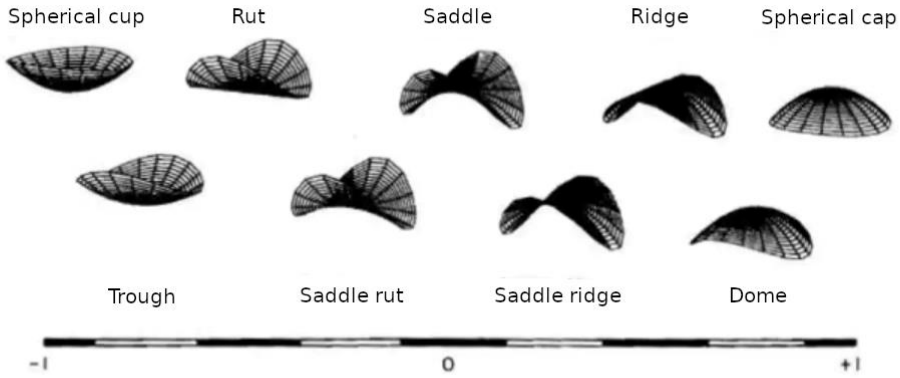


Figure 0.2 - Reference surfaces for Shape Index

- $C = \sqrt{\frac{k_1^2+k_2^2}{2}}$

  C is the curvedness and is null only in planar points. It has been designed to overcome some problems emerged using Gaussian and mean curvature, in which a null value is assigned also to parabolic and minimum local points respectively, resulting not intuitive for the observer.
  It does not discriminate between concavity and convexity.

A recap of the last six descriptor is shown in Figure 0.3.

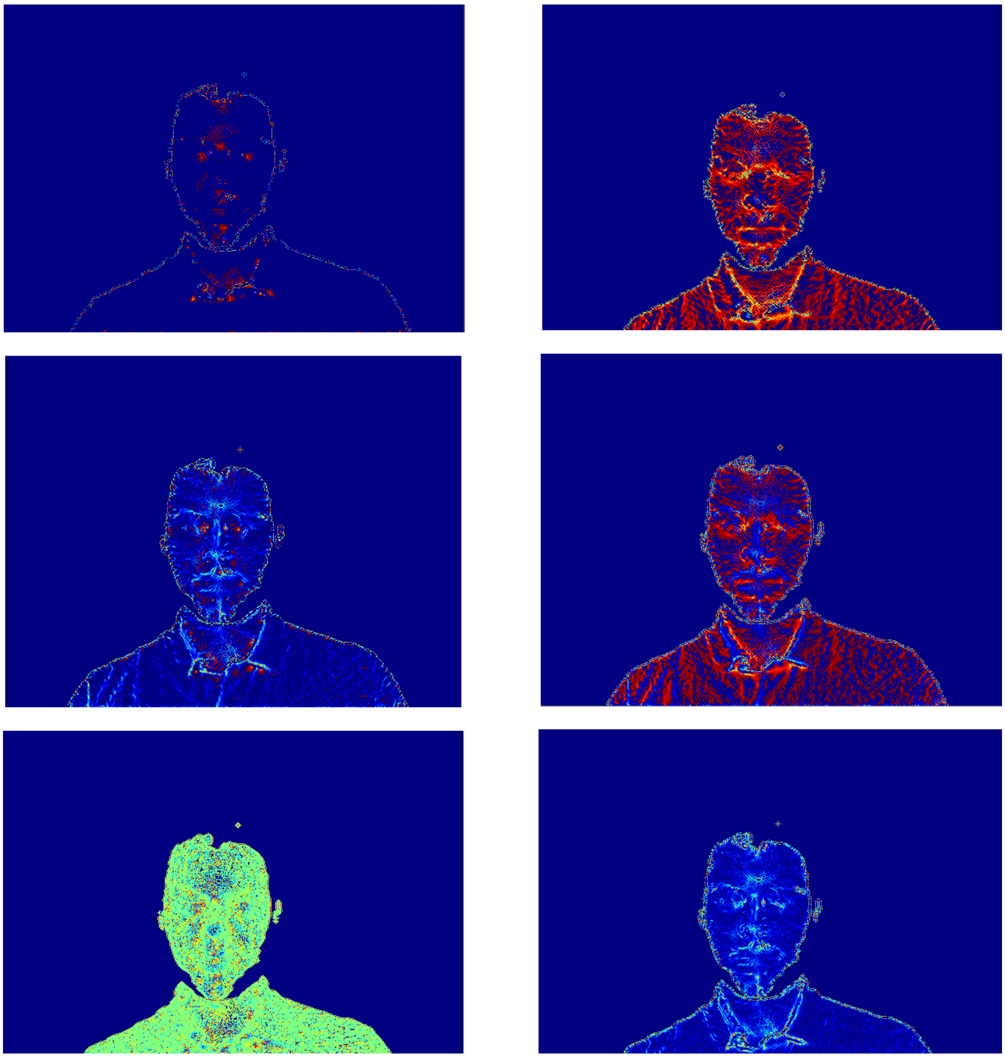


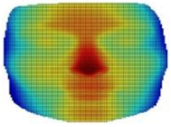
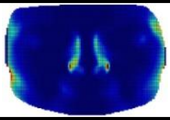
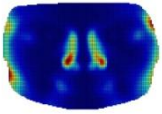
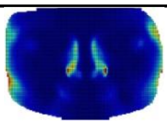
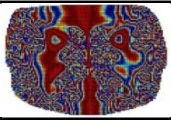
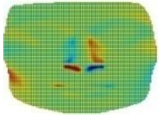
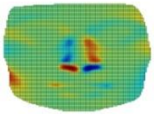
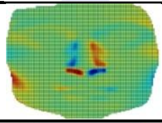
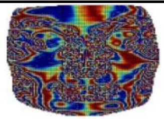
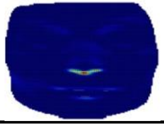
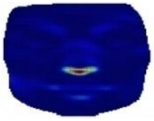
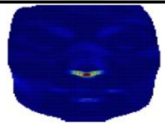
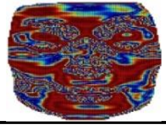
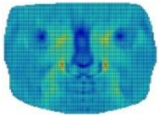
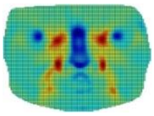
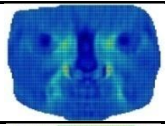
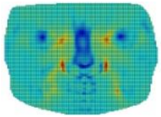
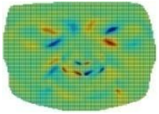
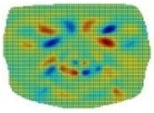
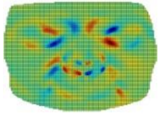
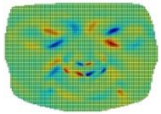
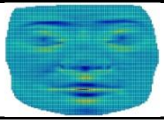
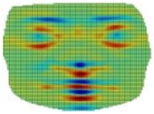
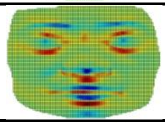
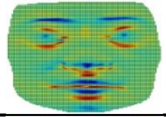
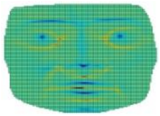
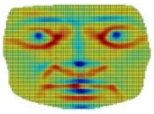
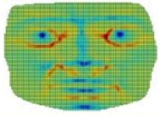
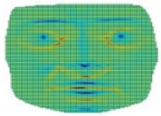
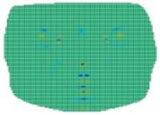
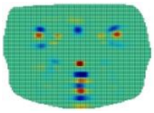
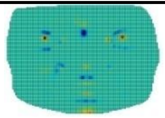
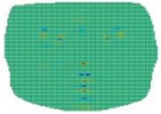
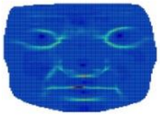
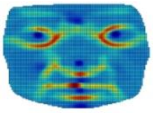
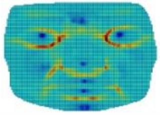
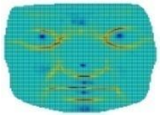
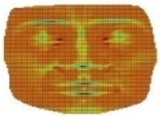
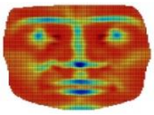
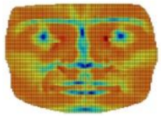
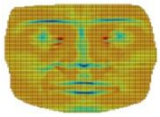
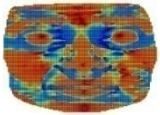
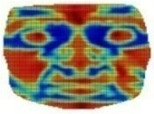
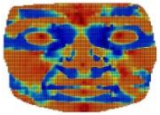
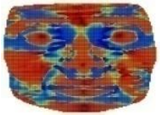
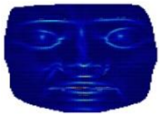
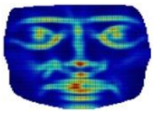
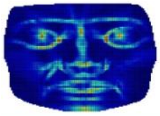
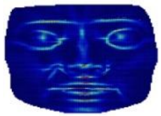
Figure 0.3 - From left to right, then from top to bottom: K, H, k1, k2, S, C

Starting from the twelve primary descriptors, it is possible to obtain *derived* and *composed* descriptors.

Derived descriptors are those entities which are built from the application of a single standard function such as mean, median, sine, cosine, logarithm. These classic functions are directly applied to the primary descriptors to generate the derived one.

Table 0.1 refers to a subject acquired with Konica Minolta Vivid laser scanner. Starting depth map is shown in the first cell of the table.

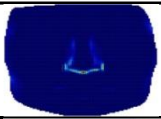
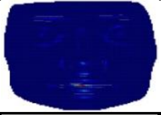
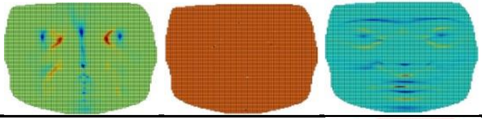
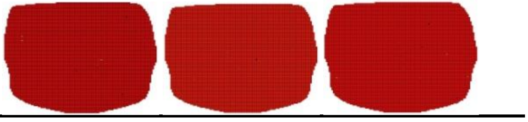
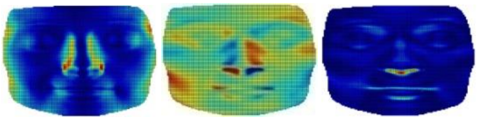
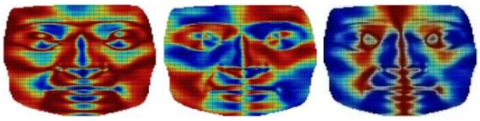
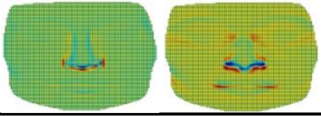
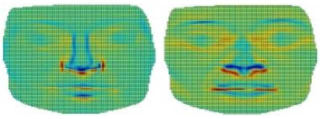
Table 0.1 - Derived descriptors

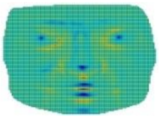
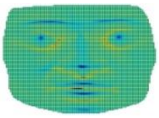
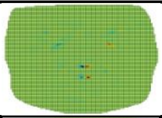
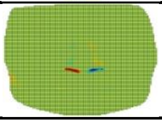
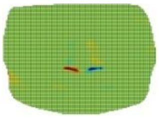
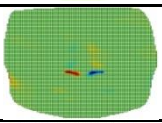
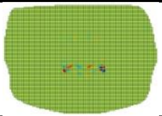
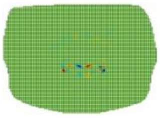
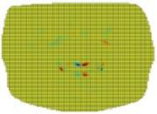
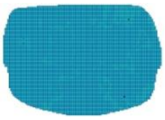
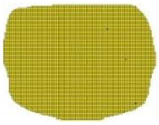
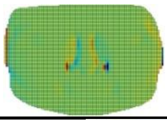
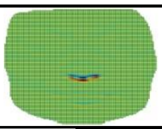
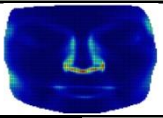
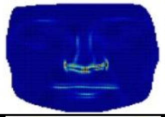
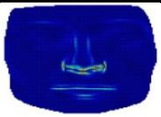
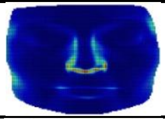
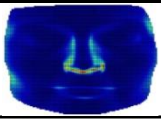
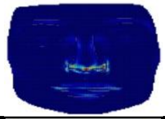
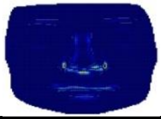
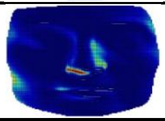
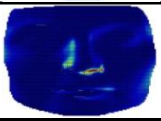
	**Primary descriptor**	**mean**	**median**	**sin**
E				
F				
G				
e				
f				
g				
H				
K				
$\boldsymbol{k_1}$				
$\boldsymbol{k_2}$				
S				
C				

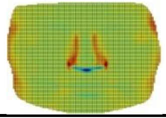
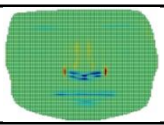
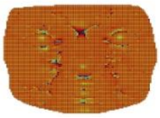
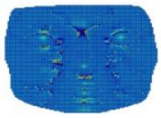
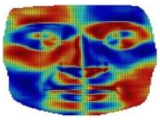
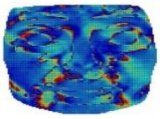
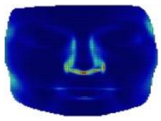
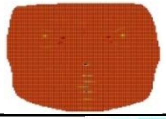
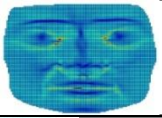
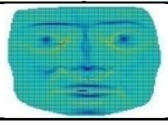
Composed descriptors have been designed by adopting standard mathematical operations such as combinations, fractions, products, special products of primary descriptors to gain novel facial representations.

These descriptors are shown in Table 0.2.

Table 0.2 - Composed descriptors

Composed descriptor(s)	**Map(s)**
$\boldsymbol{ellipsoid_1 = E^2 + F^2 + G^2}$	
$\boldsymbol{ellipsoid_2 = e^2 + f^2 + g^2}$	
$\boldsymbol{ellipsoid_i = \left(\frac{e}{E}\right)^2 + \left(\frac{f}{F}\right)^2 + \left(\frac{g}{G}\right)^2}$	
$\boldsymbol{ellipsoid_{ii} = \left(\frac{E}{e}\right)^2 + \left(\frac{F}{f}\right)^2 + \left(\frac{G}{g}\right)^2}$	
$\boldsymbol{eE = \frac{e}{E}}$ $\boldsymbol{fF = \frac{f}{F}}$ $\boldsymbol{gG = \frac{g}{G}}$	
$\boldsymbol{Ee = \frac{E}{e}}$ $\boldsymbol{Ff = \frac{F}{f}}$ $\boldsymbol{Gg = \frac{G}{g}}$	
$\boldsymbol{E_{den} = \frac{E}{\sqrt{1+h_x^2+h_y^2}}}$ $\boldsymbol{F_{den} = \frac{F}{\sqrt{1+h_x^2+h_y^2}}}$ $\boldsymbol{G_{den} = \frac{G}{\sqrt{1+h_x^2+h_y^2}}}$	
$\boldsymbol{E_{den2} = \frac{E}{1+h_x^2+h_y^2}}$ $\boldsymbol{F_{den2} = \frac{F}{1+h_x^2+h_y^2}}$ $\boldsymbol{G_{den2} = \frac{G}{1+h_x^2+h_y^2}}$	
$\boldsymbol{EeFfGg = E \cdot e + F \cdot f + G \cdot g}$ $\boldsymbol{EgFfGe = E \cdot g + F \cdot f + G \cdot e}$	
$\boldsymbol{EeFfGg_{den} = \frac{E \cdot e + F \cdot f + G \cdot g}{\sqrt{1+h_x^2+h_y^2}}}$ $\boldsymbol{EgFfGe_{den} = \frac{E \cdot g + F \cdot f + G \cdot e}{\sqrt{1+h_x^2+h_y^2}}}$	

$EeFfGg_{den2} = \frac{E\cdot e+F\cdot f+G\cdot g}{1+h_x^2+h_y^2}$ $EgFfGe_{den2} = \frac{E\cdot g+F\cdot f+G\cdot e}{1+h_x^2+h_y^2}$	 
$efg = e\cdot f\cdot g$	
$EFG = E\cdot F\cdot G$	
$EFG_{den} = \frac{E\cdot F\cdot G}{\sqrt{1+h_x^2+h_y^2}}$	
$EFG_{den2} = \frac{E\cdot F\cdot G}{1+h_x^2+h_y^2}$	
$second\ =\ h_{xx}\ \cdot h_{xy}\cdot h_{yy}$	
$second_{den} = \frac{h_{xx}\ \cdot h_{xy}\cdot h_{yy}}{\sqrt{1+h_x^2+h_y^2}}$	
$second_{den2} = \frac{h_{xx}\ \cdot h_{xy}\cdot h_{yy}}{1+h_x^2+h_y^2}$	
$x = \frac{h_x}{h_{xx}}$ $y = \frac{h_y}{h_{yy}}$	 
$xx = h_x\, h_{xx}$ $yy = h_y\, h_{yy}$	 
$cl =\ h_x^2\ +\ h_y^2\ +\ h_{xy}^2\ +\ h_{xx}^2\ +\ h_{yy}^2$	
$pnb_{AA+} = h_{xx}^2 +\ 2\cdot h_{xy} +\ h_{yy}^2$ $pnb_{AA-} = h_{xx}^2 -\ 2\cdot h_{xy} +\ h_{yy}^2$	 
$pnb_{A+}\ = h_x^2 +\ 2\cdot h_{xy} +\ h_y^2$ $pnb_{A-}\ = h_x^2 -\ 2\cdot h_{xy} +\ h_y^2$	 
$pnb_{BB+} = h_{xx}^2 +\ 2\cdot h_{xx}\cdot h_{yy} +\ h_{yy}^2$ $pnb_{BB-}\ = h_{xx}^2 -\ 2\cdot h_{xx}\cdot h_{yy} +\ h_{yy}^2$	 
$pnb_{B+}\ = h_x^2 +\ 2\cdot h_x\cdot h_y +\ h_y^2$ $pnb_{B-}\ = h_x^2 -\ 2\cdot h_x\cdot h_y +\ h_y^2$	 

$pndp_A = h_x^2 - h_y^2$ $pndp_{AA} = h_{xx}^2 - h_{yy}^2$	 
$newS_I = -\frac{2}{\pi} arctan \frac{K+H}{K-H}$ $newS_{II} = -\frac{2}{\pi} arctan \frac{K+H}{H-K}$	 
$newC = \sqrt{\frac{K^2+H^2}{2}}$	
$Sfond_1 = -\frac{2}{\pi} arctan \frac{E+F+G}{E+G-F}$ $Sfond_2 = -\frac{2}{\pi} arctan \frac{e+f+g}{e+g-f}$	 
$Cfond_1 = \sqrt{\frac{E^2+F^2+G^2}{2}}$ $Cfond_2 = \sqrt{\frac{e^2+f^2+g^2}{2}}$	
$newGaussian = K \cdot H$ $newMean = \frac{K \cdot H}{2}$	 
$thecurvature = \frac{k_1 + k_2 + K + H}{4}$	

Appendix B – Empathy test

In this module you will be asked questions to be answered on a scale from 1 to 7, where:

(1) = strongly disagree
(2) = disagree
(3) = slightly disagree
(4) = neither agree nor disagree
(5) = slightly agree
(6) = agree
(7) = strongly agree

1. I very much enjoy and feel uplifted by happy endings.

2. I cannot feel much sorrow for those who are responsible for their own misery.

3. I am moved deeply when I observe strangers who are struggling to survive.

4. I hardly ever cry when watching a very sad movie.

5. I can almost feel the pain of elderly people who are weak and must struggle to move about.

6. I cannot relate to the crying and sniffing at weddings.

7. It would be extremely painful for me to have to convey very bad news to another.

8. I cannot easily empathize with the hopes and aspirations of strangers.

9. I do not get caught up easily in the emotions generated by a crowd.

10. Unhappy movie endings haunt me for hours afterwards.

11. It pains me to see young people in wheelchairs.

12. It is very exciting for me to watch children open presents.

13. Helpless old people do not have much of an emotional effect on me.

14. The sadness of a close one easily rubs off on me.

15. I do not get overly involved with friends' problems.

16. It is difficult for me to experience strongly the feelings of characters in a book or movie.

17. It upsets me to see someone being mistreated.

18. I easily get carried away by the lyrics of love songs.

19. I am not affected easily by the strong emotions of people around me.

20. I have difficulty knowing what babies and children feel.

21. It really hurts me to watch someone who is suffering from a terminal illness.

22. A crying child does not necessarily get my attention.

23. Another's happiness can be very uplifting for me.

24. I have difficulty feeling and reacting to the emotional expressions of foreigners.

25. I get a strong urge to help when I see someone in distress.

26. I am rarely moved to tears while reading a book or watching a movie.

27. I have little sympathy for people who cause their own serious illnesses (e.g., heart disease, diabetes, lung cancer).

28. I would not watch an execution.

29. I easily get excited when those around me are lively and happy.

30. The unhappiness or distress of a stranger are not especially moving for me.

Appendix C – Alexithymia test

In this module you will be asked questions to be answered on a scale from 1 to 5, where:

(1) = strongly disagree
(2) = disagree
(3) = neither agree nor disagree
(4) = agree
(5) = strongly agree

1. I am often confused about what emotion I am feeling.

2. It is difficult for me to find the right words for my feelings.

3. I have physical sensations that even doctors do not understand.

4. I am able to describe my feelings easily.

5. I prefer to analyze problems rather than just describe them.

6. When I am upset, I do not know if I am sad, frightened, or angry.

7. I am often puzzled by sensations in my body.

8. I prefer to just let things happen rather than to understand why they turned out that way.

9. I have feelings that I cannot quite identify.

10. Being in touch with emotions is essential.

11. I find it hard to describe how I feel about people.

12. People tell me to describe my feelings more.

13. I do not know what is going on inside me.

14. I often do not know why I am angry.

15. I prefer talking to people about their daily activities rather than their feelings.

16. I prefer to watch “light” entertainment shows rather than psychological dramas.

17. It is difficult for me to reveal my innermost feelings, even to close friends.

18. I can feel close to someone, even in moments of silence.

19. I find examination of my feelings useful in solving personal problems.

20. Looking for hidden meanings in movies or plays distracts from their enjoyment.

Appendix D – Real-time FER application

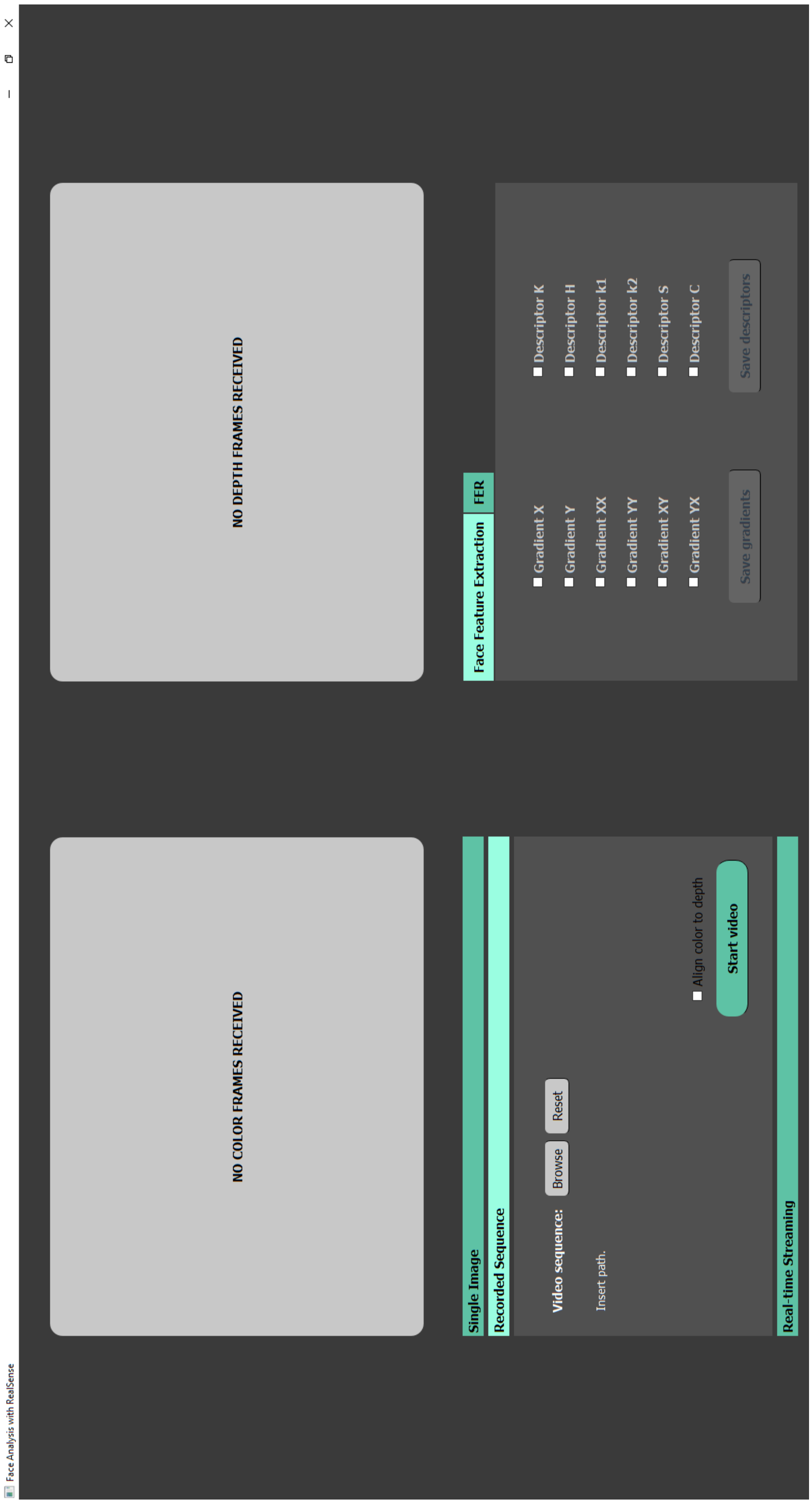


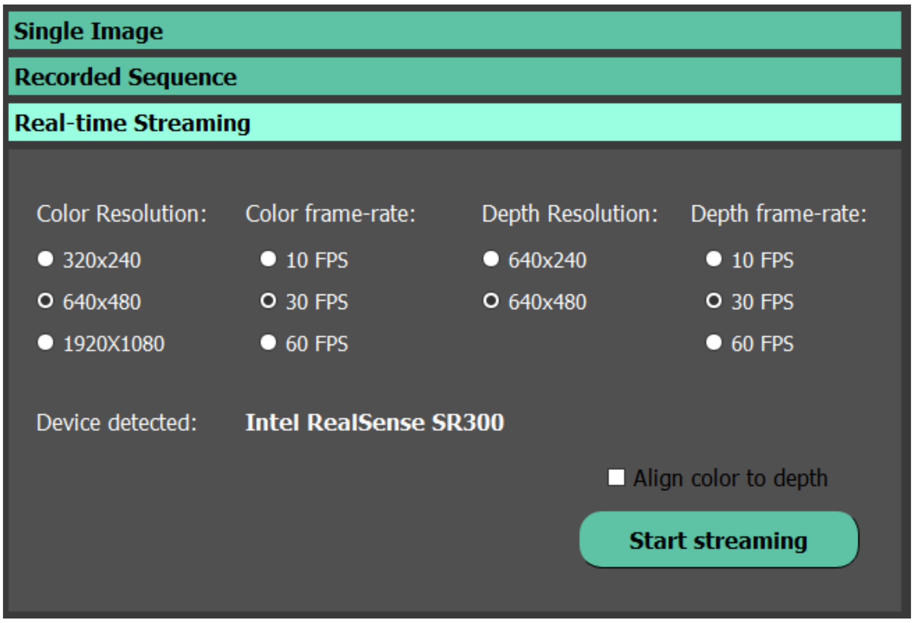
Figure 0.1 - rsGUI main screen

Figure 0.1 shows the main screen.

In the upper half of the screen there are two boxes that display video streams frame-by-frame; on the left the RGB stream, on the right the Depth stream.

In the bottom-left of the screen there is a panel dedicated to the input data that must be processed. It is possible to select single images, recorded videos (Recorded Sequence), or to use the live stream (Real-time Streaming) acquired with an Intel RealSense camera, if connected and compatible with librealsense2, that is the last version of the Software Development Kit provided by Intel.

The real-time streaming tab inside the input panel allows to make some choices regarding the video stream parameters Figure 0.2.



Color Resolution: ● 320x240 ○ 640x480 ● 1920X1080

Color frame-rate: ● 10 FPS ○ 30 FPS ● 60 FPS

Depth Resolution: ● 640x240 ○ 640x480

Depth frame-rate: ● 10 FPS ○ 30 FPS ● 60 FPS

Device detected: **Intel RealSense SR300**

☐ Align color to depth

Figure 0.2 - rsGUI real-time streaming tab

Four groups of radio buttons allow to choose one option regarding the resolution of the color stream, for the color frame rate, for the depth resolution and for the depth frame rate. In this application only the main options are available for two reasons: not all the options are available for all the RGB-D cameras. Experiments have been conducted with Intel RealSense SR300, nonetheless the possibility of using SR305 and D435 is maintained preserving the options available for all these RGB-D cameras. Moreover, sensors' firmware upgrade quite often, enabling some new features and disabling some others lesser used by the users (the firmware is the software that controls the device's hardware at low-level).

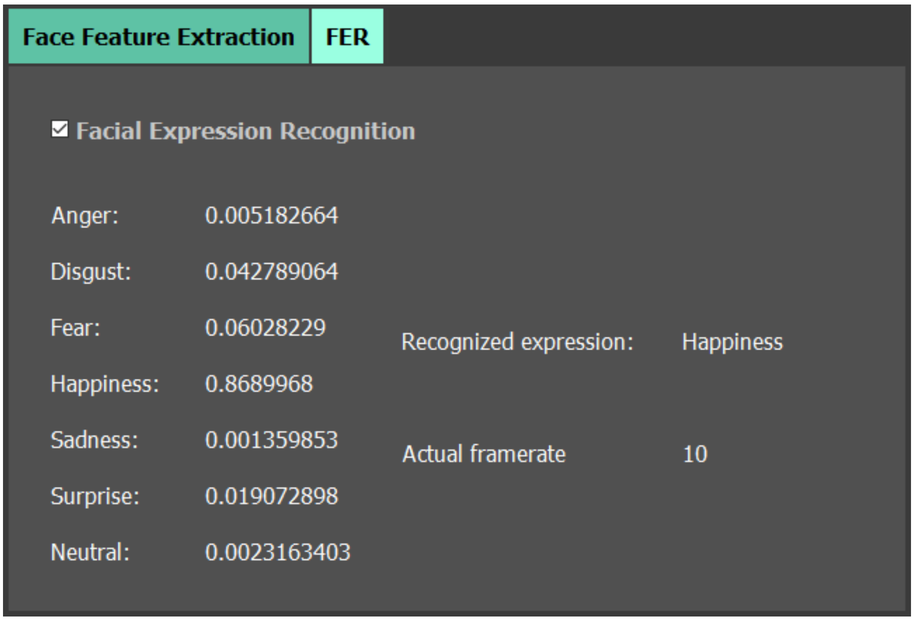
The best depth resolution (640 x 480) is always chosen to provide the best depth image possible, which provides the geometrical information, while the color resolution is typically set at 640 x 480 to favor the alignment process.

Frame rate, both for color and depth streams, is typically set at 30 FPS. Of course, the more is the value selected for the frame rate, the least is the resolution available. However, it is not necessary to pay too much attention to this parameter, because the application can adapt the frame rate according to the current load of the following data processing steps.

A label allows to understand if a device is connected. If it is connected, the information regarding the device model is displayed.

Finally, a checkbox permits to take an important decision: to align or not to align the color stream to the depth stream. If the streams are not aligned, all the information acquired by the color and the depth sensor is preserved; otherwise, there is surely a loss of information, but every single pixel is aligned, so that there is a one-to-one correspondence between pixels in color frames, representing texture information, and depth frames, representing the distance of that point from the camera.

The feature extraction panel on the bottom right of the screen contains two tabs: the first one is designed to provide the possibility of displaying gradients and descriptors in real-time; this functionality has been implemented to give the opportunity of better analyzing the geometrical features acquired by the depth sensor and, eventually, to be ready to feed the neural network with a channel containing this kind of information. The second tab is designed to activate the face expression recognition through a neural network provided as an .h5 model (Figure 0.3).



☑ Facial Expression Recognition

Anger:	0.005182664		
Disgust:	0.042789064		
Fear:	0.06028229	Recognized expression:	Happiness
Happiness:	0.8689968		
Sadness:	0.001359853	Actual framerate	10
Surprise:	0.019072898		
Neutral:	0.0023163403		

Figure 0.3 - rsGUI FER with CNN

The user can control the CNN activation through a checkbox. Once that the application is running, the numerical values, output of the neural network, identify the degree of membership for every emotion that the CNN aims to recognize, i.e. the six basic emotions (anger, disgust, fear, happiness, sadness, surprise) and the neutral face, namely the absence of expressed emotions.

On the right of the panel the recognized expression, evaluated as the emotion with the highest score is displayed, as well as the actual frame rate. This information is useful to understand how much the downgrade from the frame rate selected at the beginning of the session is in order to take countermeasure. This means to evaluate if the frame rate can be still considered real-time or if it is necessary to simplify one or more steps in data processing.

References

[1] L. Ulrich, E. Vezzetti, S. Moos and F. Marcolin, "Analysis of RGB-D camera technologies for supporting different facial usage scenarios," *Multimedia Tools and Applications,* vol. 79, no. 39-40, pp. 29375-29398, 2020.

[2] P. Henry, M. Krainin, E. Herbst, X. Ren and D. Fox, "RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments," *The International Journal of Robotics Research,* vol. 31, no. 5, pp. 647-663, 2012.

[3] X. Chen, H. Ma, J. Wan, B. Li and T. Xia, "Multi-View 3D Object Detection Network for Autonomous Driving," in *30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[4] S. Yarboro, P. H. Richter and D. M. Kahler, "The evolution of 3D imaging in orthopaedic trauma care," *Der Unfallchirurg,* vol. 120, no. 1, pp. 5-9, 2017.

[5] I. Valverde, G. Gomez, A. Gonzalez, C. A. A. Suarez-Mejias, J. F. Coserria, S. Uribe, T. Gomez-Cla and A. R. Hosseinpour, "Three-dimensional patient-specific cardiac model for surgical planning in Nikaidoh procedure," *Cardiology in the Young,* vol. 25, no. 4, pp. 698-704, 2015.

[6] R. H. Taylor, A. Menciassi, G. Fichtinger, P. Fiorini and P. Dario, "Medical robotics and computer-integrated surgery," in *Springer Handbook of Robotics*, Springer International Publishing, 2016, pp. 1657-1683.

[7] M. P. Chae, W. M. Rozen, P. G. McMenamin, M. W. Findlay, R. T. Spychal and D. J. Hunter-Smith, "Emerging applications of bedside 3D printing in plastic surgery," *Frontiers in surgery,* vol. 2, p. 1, 2015.

[8] M. I. Mohammed, J. Tatineni, B. Cadd, G. Peart and I. Gibson, "Applications of 3D topography scanning and multi-material additive manufacturing for facial prosthesis development and production," in *Solid Freeform Fabrication 2016: Proceedings of the 27th Annual*

International Solid Freeform Fabrication Symposium - An Additive Manufacturing Conference, 2016.

[9] A. Dawood, B. M. Marti, V. Sauret-Jackson and A. Darwood, "3D printing in dentistry," *British dental journal,* vol. 219, no. 11, pp. 521-529, 2015.

[10] M. Boffano, P. Pellegrino, N. Ratto, M. Giachino, U. Albertini, A. Aprato, E. Boux, G. Collo, A. Ferro, S. Marone, A. Massè and R. Piana, "Custom-made 3D-printed pelvic prosthesis: is it a safe option for the limb salvage in tumours and catastrophic total hip arthroplasty failures?," in *Orthopaedic Proceedings*, 2018.

[11] S. Siebert and J. Teizer, "Mobile 3D mapping for surveying earthwork projects using an Unmanned Aerial Vehicle (UAV) system," *Automation in Construction,* no. 41, pp. 1-14, 2014.

[12] M. Kedzierski and A. Fryskowska, "Terrestrial and aerial laser scanning data integration using wavelet analysis for the purpose of 3D building modeling," *Sensors,* vol. 14, no. 7, pp. 12070-12092, 2014.

[13] M. Forte, "3D archaeology: new perspectives and challenges - the example of Çatalhöyük," *Journal of Eastern Mediterranean Archaeology & Heritage Studies,* vol. 2, no. 1, pp. 1-29, 2014.

[14] V. Albino, U. Berardi and R. M. Dangelico, "Smart cities: Definitions, dimensions, performance, and initiatives," *Journal of urban technology,* vol. 22, no. 1, pp. 3-21, 2015.

[15] H. Wu and H. L. P. Xu, "Design and Implementation of Cloud Service System Based on Face Recognition," in *Conference on Complex, Intelligent, and Software Intensive Systems*, 2020.

[16] D. Robertson, D. G. Macfarlane, R. I. Hunter, S. L. Cassidy, N. Llombart, E. Gandini, T. Bryllert, M. Ferndahl, H. Lindstrom, J. Tenhunen, H. Vasama, J. Huopana, T. Selkala and A.-J. Vuotikka, "High resolution, wide field of view, real time 340GHz 3D imaging radar for security screening," in *Passive and Active Millimeter-Wave Imaging XX*, 2017.

[17] M. Zollhofer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Perez, M. Stamminger, M. Niessner and C. Theobalt, "State of the art on monocular 3D face reconstruction, tracking, and applications," *Computer Graphics Forum,* vol. 37, no. 2, pp. 523-550, 2018.

[18] S. Riaz, U. Park, J. Choi and P. Natarajan, "Age progression by gender-specific 3D aging model," *Machine Vision and Applications,* vol. 30, no. 1, pp. 91-109, 2019.

[19] Z. Geng, C. Cao and S. Tulyakov, "3d guided fine-grained face manipulation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[20] J. Deng, A. Roussos, G. Chrysos, E. Ververas, I. Kotsia, J. Shen and S. Zafeiriou, "The menpo benchmark for multi-pose 2D and 3D facial landmark localisation and tracking," *International Journal of Computer Vision,* vol. 127, no. 6-7, pp. 599-624, 2019.

[21] A. Abate, M. Nappi, D. Riccio and G. Sabatino, "2D and 3D Face Recognition: A Survey," *Pattern Recognition Letters,* vol. 28, no. 14, pp. 1885-1906, 2007.

[22] J. Cao, Y. Hu, B. Yu, R. He and Z. Sun, "3D aided duet GANs for multi-view face image synthesis," *IEEE Transactions on Information Forensics and Security,* vol. 14, no. 8, pp. 2028-2042, 2019.

[23] L. Tran and X. Liu, "On learning 3d face morphable model from in-the-wild images," in *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[24] S. Zhou and S. Xiao, " 3D face recognition: a survey," *Human-centric Computing and Information Sciences,* vol. 8, no. 1, p. 35, 2018.

[25] E. Vezzetti, F. Marcolin, S. Tornincasa, L. Ulrich and N. Dagnes, "3D geometry-based automatic landmark localization in presence of facial occlusions," *Multimedia Tools and Applications,* vol. 77, no. 11, pp. 14177-14205, 2018.

[26] N. Dagnes, E. Vezzetti, F. Marcolin and S. Tornincasa, "Occlusion detection and restoration techniques for 3D face recognition: a literature review," *Machine Vision and Applications,* vol. 29, no. 5, pp. 789-813, 2018.

[27] E. Vezzetti, S. Moos, F. Marcolin and V. Stola, "A pose-independent method for 3D face landmark formalization," *Computer Methods and Programs in Biomedicine,* vol. 108, no. 3, pp. 1078-1096, 2012.

[28] P. Verschuren, H. Doorewaard and M. J. Mellion, Designing a research project, The Hague: Eleven International publishing house, 2010.

[29] Y. Akao, "Development history of quality function deployment," *The Customer Driven Approach to Quality Planning and Deployment,* vol. 339, p. 90, 1994.

[30] A. W. Young, E. De Haan and R. Bauer, "Face perception: A very special issue," *Journal of neuropsychology,* vol. 2, no. 1, pp. 1-14, 2008.

[31] J. Morton and M. H. Johnson, "CONSPEC and CONLERN: a two-process theory of infant face recognition," *Psychological review,* vol. 98, no. 2, pp. 164-181, 1991.

[32] R. L. Fantz, "The origin of form perception," *Scientific American,* vol. 204, no. 5, pp. 66-73, 1961.

[33] J. C. Meadows, "The anatomical basis of prosopagnosia," *Journal of Neurology, Neurosurgery & Psychiatry,* vol. 37, no. 5, pp. 489-501, 1974.

[34] A. R. Damasio, H. Damasio and G. W. Van Hoesen, "Prosopagnosia Anatomic basis and behavioral mechanism," *Neurology,* vol. 32, no. 4, pp. 331-441, 1982.

[35] S. Mondal, I. Mukhopadhyay and S. Dutta, "Review and comparisons of face detection techniques," in *Proceedings of Interantional Ethical Hacking Conference*, Singapore, 2019.

[36] B. C. Becker and E. G. Ortiz, "Evaluation of face recognition techniques for application to facebook," in *8th IEEE International Conference on Automatic Face & Gesture Recognition*, 2008.

[37] F. Erden, A. Z. Alkar and A. E. Cetin, "A robust system for counting people using an infrared sensor and a camera," *Infrared Physics & Technology,* vol. 72, pp. 127-134, 2015.

[38] S. Lamba, N. Nain and H. Chahar, "A robust multi-model approach for face detection in crowd," in *12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, 2016.

[39] M. Loey, G. Manogaran, M. H. N. Taha and N. E. M. Khalifa, "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic,"

Measurement: Journal of the International Measurement Confederation, vol. 167, p. 108288, 2020.

[40] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition*, 2001.

[41] C. Zhang and Z. Zhang, "A Survey of Recent Advances in Face Detection," *Learning,* vol. June, p. 17, 2010.

[42] C. Liu, "A Bayesian discriminating features method for face detection," *IEEE transactions on pattern analysis and machine intelligence,* vol. 25, no. 6, pp. 741-754, 2003.

[43] R. Féraud, O. J. Bernier, J. E. Viallet and M. Collobert, "A fast and accurate face detector based on neural networks," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 23, no. 1, pp. 42-53, 2001.

[44] A. Colombo, C. Cusano and R. Schettini, "3D face detection using curvature analysis," *Pattern recognition,* vol. 39, no. 3, pp. 444-455, 2006.

[45] B. Heisele, T. Serre and T. Poggio, "A component-based framework for face detection and identification," *International Journal of Computer Vision,* vol. 74, no. 2, pp. 167-181, 2007.

[46] C. Maes, T. Fabry, J. Keustermans, D. Smeets, P. Suetens and D. Vandermeulen, "Feature detection on 3D face surfaces for pose normalisation and recognition," in *4th IEEE International Conference on Biometrics: Theory Applications and Systems*, 2010.

[47] A. Neethu, S. Athi Narayanan and B. Kamal, "People count estimation using hybrid face detection method," in *International Conference on Information Science (ICIS)*, 2016.

[48] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters,* vol. 23, no. 10, pp. 1499-1503, 2016.

[49] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *12th IEEE International Conference on Automatic Face & Gesture Recognition*, 2017.

[50] X. Sun, P. Wu and S. C. Hoi, "Face detection using deep learning: An improved faster RCNN approach," *Neurocomputing,* vol. 299, pp. 42-50, 2018.

[51] J. Ashbourn, Biometrics: Advanced identity verification: The complete guide, Springer, 2014.

[52] A. K. Jain, L. Hong, S. Pankanti and R. Bolle, "An identity-authentication system using fingerprints," *Proceedings of the IEEE,* vol. 85, no. 9, pp. 1365-1388, 1997.

[53] J. Galbally, S. Marcel and J. Fierrez, "Image quality assessment for fake biometric decision: Application to iris, fingerprint, and face recognition," *IEEE transactions on image processing,* vol. 23, no. 2, pp. 710-724, 2014.

[54] M. E. Fathy, V. M. Patel and R. Chellappa, "Face-based active authentication on mobile devices," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[55] K. Jonsson, J. Kittler, Y. P. Li and J. Matas, "Support vector machines for face authentication," *Image and Vision Computing,* vol. 20, no. 5-6, pp. 369-375, 2002.

[56] Q. Tao and R. N. Veldhuis, "Biometric authentication for a mobile personal device," in *3rd Annual International Conference on Mobile and Ubiquitous Systems-Workshops*, 2006.

[57] P. Samangouei, V. M. Patel and R. Chellappa, "Facial attributes for active authentication on mobile devices," *Image and Vision Computing,* vol. 58, pp. 181-192, 2017.

[58] A. Ross and A. Jain, "Information fusion in biometrics," *Pattern recognition letters,* vol. 24, no. 13, pp. 2115-2125, 2003.

[59] A. Seal and C. Panigrahy, "Human authentication based on fusion of thermal and visible face images," *Multimedia Tools and Applications,* vol. 78, no. 21, pp. 30373-30395, 2019.

[60] K. Matsuoka, M. Irvan, R. Kobayashi and R. Shigetomi Yamaguchi, "A Score Fusion Method by Neural Network in Multi-Factor Authentication," in *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, 2020.

[61] Y. Taigman, M. Yang, M. A. Ranzato and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[62] A. Rattani, R. Derakhshani and A. Ross, Selfie Biometrics: Advances and Challenges, Springer Nature, 2019.

[63] M. Sajjad, M. Nasir, K. Muhammad, S. Khan, Z. Jan, A. Sangaiah, M. Elhoseny and S. Baik, "Raspberry Pi assisted face recognition framework for enhanced law-enforcement services in smart cities," *Future Generation Computer Systems,* vol. 108, pp. 995-1007, 2020.

[64] L. A. Cament, F. J. Galdames, K. W. Bowyer and C. A. Perez, "Face recognition under pose variation with local Gabor features enhanced by active shape and statistical models," *Pattern recognition,* vol. 48, no. 11, pp. 3371-3384, 2015.

[65] M. S. Hossain and G. Muhammad, "Cloud-assisted speech and face recognition framework for health monitoring," *Mobile Networks and Applications,* vol. 20, no. 3, pp. 391-399, 2015.

[66] L. A. Elrefaei, A. Alharthi, H. Alamoudi, S. Almutairi and F. Al-rammah, "Real-time face detection and tracking on mobile phones for criminal detection," in *2nd International Conference on Anti-Cyber Crimes (ICACC)*, 2017.

[67] T. P. Driver, S. Sundaram, G. Khandelwal and M. Sahasrabudhe, "Systems And Methods For Patient Identification Using Mobile Face Recognition". U.S. Patent 11/945, 2009.

[68] R. Raghavendra, K. B. Raja, A. Pflug, B. Yang and C. Busch, "3d face reconstruction and multimodal person identification from video captured using smartphone camera," in *IEEE International Conference on Technologies for Homeland Security (HST)*, 2013.

[69] A. Sepas-Moghaddam, F. M. Pereira and P. L. Correia, "Face recognition: A novel multi-level taxonomy based survey," *IET Biometrics,* 2019.

[70] R. Jafri and H. Arabnia, "A survey of face recognition techniques," *Journal of Information Processing Systems,* vol. 5, no. 2, pp. 41-68, 2009.

[71] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *IEEE computer society conference on computer vision and pattern recognition*, 1991.

[72] X. He, S. Yan, Y. Hu, P. Niyogi and H. J. Zhang, "Face recognition using laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 27, no. 3, pp. 328-340, 2005.

[73] T. Ahonen, A. Hadid and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 28, no. 12, pp. 2037-2041, 2006.

[74] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry and Y. Ma, "Robust face recognition via sparse representation," *IEEE transactions on pattern analysis and machine intelligence,* vol. 31, no. 2, pp. 210-227, 2009.

[75] O. M. Parkhi, A. Vedaldi and A. Zisserman, "Deep face recognition," 2015.

[76] Y. Wen, K. Zhang, Z. Li and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*, 2016.

[77] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *30th IEEE conference on computer vision and pattern recognition*, 2017.

[78] J. Deng, J. Guo, N. Xue and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[79] A. Lagorio, M. Tistarelli, M. Cadoni, C. Fookes and S. Sridharan, "Liveness detection based on 3D face shape analysis," in *International Workshop on Biometrics and Forensics (IWBF)*, 2013.

[80] G. Albakri and S. Alghowinem, "The effectiveness of depth data in liveness face authentication using 3D sensor cameras," *Sensors,* vol. 19, no. 8, p. 1928, 2019.

[81] Y. Atoum, Y. Liu, A. Jourabloo and X. Liu, "Face anti-spoofing using patch and depth-based CNNs," in *IEEE International Joint Conference on Biometrics (IJCB)*, 2017.

[82] P. Ekman, R. W. Levenson and W. V. Friesen, "Autonomic nervous system activity distinguishes among emotions," *Science,* vol. 221, no. 4616, pp. 1208-1210, 1983.

[83] P. Ekman, "An argument for basic emotions," *Cognition & emotion,* vol. 6, no. 3-4, pp. 169-200, 1992.

[84] P. Ekman, W. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. LeCompte, T. Pitcairn, P. Ricci-Bitti and K. Scherer, "Universals and cultural differences in the judgments of facial expressions of emotion," *Journal of personality and social psychology,* vol. 53, no. 4, pp. 712-717, 1987.

[85] P. Ekman and W. V. Friesen, Unmasking the face: A guide to recognizing emotions from facial clues, Ishk, 2003.

[86] M. S. Bartlett, G. Littlewort, I. Fasel and J. R. Movellan, "Real time face detection and facial expression recognition: development and applications to human computer interaction," in *2003 Conference on computer vision and pattern recognition workshop*, 2003.

[87] C. Shan, S. Gong and P. W. McOwan, "Robust facial expression recognition using local binary patterns," in *IEEE International Conference on Image Processing*, 2005.

[88] C. Shan, S. Gong and McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and vision Computing,* vol. 27, no. 6, pp. 803-816, 2009.

[89] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE transactions on image processing,* vol. 16, no. 1, pp. 172-187, 2006.

[90] G. Guo and C. R. Dyer, "Learning from examples in the small sample case: face expression recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* vol. 35, no. 3, pp. 477-488, 2005.

[91] M. Matsugu, K. Mori, Y. Mitari and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Networks,* vol. 16, no. 5-6, pp. 555-559, 2003.

[92] B. Jingxin, L. Yinan and Z. Shuo, "3D Multi-poses Face Expression Recognition Based on Action Units," in *Proceedings of the 2019 International Conference on Information Technology and Computer Communications*, 2019.

[93] M. S. Bartlett, G. Littlewort, I. Fasel and J. R. Movellan, "Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction," in *Conference On Computer Vision and Pattern Recognition Workshop*, 2003.

[94] D. A. Small and N. M. Verrochi, "The face of need: Facial emotion expression on charity advertisement," *Journal of Marketing Research,* vol. 46, no. 6, pp. 777-787, 2009.

[95] J.-S. Lee and D.-H. Shin, "The relationship between human and smart TVs based on emotion recognition in HCI," in *International Conference on Computational Science and Its Applications*, 2014.

[96] D. McDuff, A. Mahmoud, M. T. J. Amr and R. E. Kaliouby, "AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2016.

[97] M. G. Calvo and L. Nummenmaa, "Perceptual and affective mechanisms in facial expression recognition," *Cognition and Emotion,* vol. 30, no. 6, pp. 1081-1106, 2016.

[98] E. C. Olivetti, M. G. Violante, E. Vezzetti, F. Marcolin and B. Eynard, "Engagement Evaluation in a Virtual Learning Environment via Facial Expression Recognition and Self-Reports: A Preliminary Approach," *Applied Sciences,* vol. 10, no. 1, p. 314, 2020.

[99] F. Nonis, E. C. Olivetti, F. Marcolin, M. G. Violante, E. Vezzetti and S. Moos, "Questionnaires or Inner Feelings: Who Measures the Engagement Better?," *Applied Sciences,* vol. 10, no. 2, p. 609, 2020.

[100] Z. Liu, M. Wu, W. Cao, L. Chen, J. Xu, R. Zhang, M. Zhou and J. Mao, "A facial expression emotion recognition based human-robot interaction system," *IEEE/CAA Journal of Automatica Sinica,* vol. 4, no. 4, pp. 668-676, 2017.

[101] G. R. Alexandre, J. M. Soares and G. A. Pereira Thé, "Systematic review of 3D facial expression recognition methods," *Pattern Recognition,* vol. 100, p. 107108, 2020.

[102] Y. Chen, R. Hu, J. Xiao and Z. Wang, "Multisource surveillance video coding by exploiting 3D and 2D knolwedge," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[103] Q. T. Luong and O. D. Faugeras, "The fundamental matrix: Theory, algorithms, and stability analysis," *International journal of computer vision,* vol. 17, no. 1, pp. 43-75, 1996.

[104] E. Vezzetti and F. Marcolin, "Geometry-based 3D face morphology analysis: soft-tissue landmark formalisation," *Multimedia Tools and Applications,* vol. 68, no. 3, pp. 895-929, 2014.

[105] "Stereolabs," [Online]. Available: https://www.stereolabs.com/zed/.

[106] "Carnegie Robotics," [Online]. Available: https://carnegierobotics.com/multisense-s7/.

[107] "e-con Systems," [Online]. Available: https://www.e-consystems.com/3D-USB-stereo-camera.asp.

[108] "Nerian," [Online]. Available: https://nerian.com/products/scenescan-stereo-vision/.

[109] "Roboception," [Online]. Available: https://roboception.com/en/rc_visard-en/.

[110] "Duo 3D," [Online]. Available: https://duo3d.com/product/duo-minilx-lv1#tab=specs.

[111] J. Salvi, J. Pages and J. Batlle, "Pattern codification strategies in structured light systems," *Pattern recognition,* vol. 37, no. 4, pp. 827-849, 2004.

[112] "Intel RealSense F200," [Online]. Available: https://communities.intel.com/docs/DOC-24012.

[113] "Intel RealSense SR300," [Online]. Available: https://www.intel.com/content/dam/support/us/en/documents/emerging-technologies/intel-realsense-technology/realsense-sr300-datasheet1-0.pdf.

[114] M. R. Andersen, T. Jensen, P. Lisouski, A. K. Mortensen, M. K. Hansen, T. Gregersen and P. Ahrendt, "Kinect depth sensor evaluation

for computer vision applications," *Electronics and Computer Engineering,* vol. 1, no. 6, p. 37, 2012.

[115] "Asus," [Online]. Available: https://www.asus.com/us/3D-Sensor/Xtion_PRO_LIVE/specifications/.

[116] "Ensenso," [Online]. Available: https://www.ensenso.com/support/modellisting/?id=N35-606-16-BL.

[117] "Orbecc3d," [Online]. Available: https://orbbec3d.com/astra-mini/.

[118] "Photoneo," [Online]. Available: https://www.photoneo.com/phoxi-3d-scanner/.

[119] "Structure," Occipital, [Online]. Available: https://support.structure.io/article/157-what-are-the-structure-sensors-technical-specifications.

[120] F. S. D. Remondino, TOF Range-Imaging Cameras, Springer, 2013.

[121] A. Corti, S. Giancola, G. Mainetti and R. Sala, "A metrological characterization of the Kinect V2 time-of-flight camera," *Robotics and Autonomous Systems,* vol. 75, pp. 584-594, 2016.

[122] "IFM," [Online]. Available: https://www.ifm.com/us/en/product/O3D303.

[123] "Sick," [Online]. Available: https://www.sick.com/it/it/visione/visione-3d/visionary-t/c/g358152.

[124] "Basler," [Online]. Available: https://www.baslerweb.com/en/products/cameras/3d-cameras/time-of-flight-camera/.

[125] "PMD," [Online]. Available: https://cdn.pressebox.de/f/b25e5f1dca1f55f7/attachments/0406150.attachment.

[126] "Mesa 4000," [Online]. Available: http://www.adept.net.au/cameras/Mesa/SR4000.shtml.

[127] "Mesa 4500," [Online]. Available: http://www.adept.net.au/cameras/Mesa/SR4500.shtml.

[128] "Sony Depth Sensing," [Online]. Available: https://www.sony-depthsensing.com/Portals/0/Download/WEB_20120907_SK_DS325_ Datasheet_V2.1.pdf.

[129] R. D. Bock, "Low-cost 3D security camera," *Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything,* vol. 10643, p. 106430E, 2018.

[130] M. Carfagni, R. Furferi, L. Governi, C. Santarelli, M. Servi, F. Uccheddu and Y. Volpe, "Metrological and critical characterization of the Intel D415 stereo depth camera," *Sensors,* vol. 19, no. 3, p. 489, 2019.

[131] "Intel RealSense R200," [Online]. Available: https://www.intel.it/content/www/it/it/support/articles/000016214/emerging-technologies/intel-realsense-technology.html.

[132] "Intel," [Online]. Available: https://www.intel.com/content/dam/support/us/en/documents/emerging-technologies/intel-realsense-technology/Intel-RealSense-D400-Series-Datasheet.pdf.

[133] "Intel Euclid," [Online]. Available: https://click.intel.com/media/productid2100_10052017/335926-001_public.pdf.

[134] P. T. Chuang, "Combining the analytic hierarchy process and quality function deployment for a location decision from a requirement perspective," *The International Journal of Advanced Manufacturing Technology,* vol. 18, no. 11, pp. 842-849, 2001.

[135] C. Kahraman, T. Ertay and G. Büyüközkan, "A fuzzy optimization model for QFD planning process using analytic network approach," *European Journal of Operational Research,* vol. 171, no. 2, pp. 390-411, 2006.

[136] M. Aljohani and T. Alam, "Real time face detection in ad hoc network of android smart devices," *Advances in Computational Intelligence,* vol. 509, pp. 245-255, 2017.

[137] S. Chen, Y. Liu, X. Gao and Z. Han, "Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices," in *Chinese Conference on Biometric Recognition*, Cham, 2018.

[138] J. V. Patil and P. Bailke, "Real time facial expression recognition using RealSense camera and ANN," in *International Conference on Inventive Computation Technologies*, 2016.

[139] R. Horaud, M. Hansard, G. Evangelidis and C. Menier, "An overview of depth cameras and range scanners based on time-of-flight technologies," *Machine vision and applications,* vol. 27, no. 7, pp. 1005-1020, 2016.

[140] L. Streeter and Y. Kuang, "Metrological aspects of time-of-flight range imaging," *IEEE Instrumentation & Measurement Magazine,* vol. 22, no. 2, pp. 21-26, 2019.

[141] E. Kirsten, L. Inocencio, M. Veronez, L. Da Silveira, F. Bordin and F. Marson, "3D data acquisition using stereo camera," in *IEEE International Geoscience and Remote Sensing Symposium*, 2018.

[142] M. Chowdhury, J. Gao and R. Islam, "Human detection and localization in secure access control by analysing facial features," in *IEEE 11th Conference on Industrial Electronics and Applications*, 2016.

[143] T. Pribanic, T. Petkovic, M. Donlic, V. Angladon and S. Gasparni, "3D structured light scanner on the smartphone," in *International Conference on Image Analysis and Recognition*, Cham, 2016.

[144] Z. Wang, "Robust three-dimensional face reconstruction by one-shot structured light line pattern," *Optics and Lasers in Engineering,* vol. 124, p. 105768, 2020.

[145] S. Giancola, M. Valenti and R. Sala, "Metrological Qualification of the Intel D400™ Active Stereoscopy Cameras," in *A Survey on 3D Cameras: Metrological Comparison of Time-of-Flight, Structured-Light and Active Stereoscopy Technologies*, Springer, 2018, pp. 71-85.

[146] S. Fukuda, Emotional engineering: service development, Springer Science & Business Media., 2010.

[147] D. Cortés Sáenz, C. E. Díaz Domínguez, P. Llorach-Massana, A. Abella García and J. L. Hernández Arellano, "A Series of Recommendations for Industrial Design Conceptualizing Based on Emotional Design," in *Managing Innovation in Highly Restrictive Environments*, Cham, Springer, 2019, pp. 167-185.

[148] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology,* vol. 39, no. 6, pp. 1161-1178, 1980.

[149] J. A. Russell and B. Fehr, "Relativity in the perception of emotion in facial expressions," *Journal of Experimental Psychology: General,* vol. 116, no. 3, pp. 223-237, 1987.

[150] E. Vezzetti, F. Marcolin, S. Tornincasa, L. Ulrich and N. Dagnes, "3D geometry-based automatic landmark localization in presence of facial occlusions.," *Multimedia Tools and Applications,* vol. 77, no. 11, pp. 14177-14205, 2018.

[151] F. Marcolin and E. Vezzetti, "Novel descriptors for geometrical 3D face analysis," *Multimedia Tools and Applications,* vol. 76, no. 12, pp. 13805-13834, 2017.

[152] G. R. Swennen, F. A. Schutyser and J. E. Hausamen, Three-dimensional cephalometry: a color atlas and manual, Springer Science & Business Media, 2005.

[153] W. Wundt, Outlines of psychology, vol. 1, 1896, p. 14.

[154] R. S. Woodworth, B. Barber and H. Schlosberg, Experimental psychology, Oxford and IBH Publishing, 1954.

[155] P. Ekman, "A methodological discussion of nonverbal behavior," *The Journal of psychology,* vol. 43, no. 1, pp. 141-149, 1957.

[156] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci and J. F. Cohn, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[157] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition & emotion,* vol. 9, no. 1, pp. 87-108, 1995.

[158] M. Soleymani, M. Pantic and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE transactions on affective computing,* vol. 3, no. 2, pp. 211-223, 2011.

[159] G. Chanel, C. Rebetez, M. Bétrancourt and T. Pun, "Emotion assessment from physiological signals for adaptation of game difficulty," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans,* vol. 41, no. 6, pp. 1052-1063, 2011.

[160] C. Taskiran, S. Karaismailoglu, H. T. Cak Esen, Z. Tuzun, A. Erdem, Z. D. Balkanci, A. Dolgun and S. Cengel Kultur, "Clinical features and subjective/physiological responses to emotional stimuli in the presence of emotion dysregulation in attention-deficit hyperactivity disorder," *Journal of Clinical and Experimental Neuropsychology,* vol. 40, no. 4, pp. 389-404, 2018.

[161] S. Migliore, G. Curcio, C. Porcaro, C. Cottone, I. Simonelli, G. D'aurizio and M. M. Filippi, "Emotional processing in RRMS patients: Dissociation between behavioural and neurophysiological response," *Multiple sclerosis and related disorders,* vol. 27, pp. 344-349, 2019.

[162] C. Moret-Tatay, P. M. Rueda, G. Bernabé-Valero and D. Gamermann, "Emotional Recognition in Schizophrenia: An Analysis of Response Components in Middle-Aged Adults," *Psychiatric Quarterly,* vol. 90, no. 3, pp. 543-552, 2019.

[163] E. Bekele, D. Bian, Z. Zheng, J. Peterman, S. Park and N. Sarkar, "Responses during facial emotional expression recognition tasks using virtual reality and static iaps pictures for adults with schizophrenia," in *International Conference on Virtual, Augmented and Mixed Reality*, 2014.

[164] M. A. A. Peter, T. A. Klimstra, M. Faulborn and A. J. J. M. Vingerhoets, "Subjective emotional responses to IAPS pictures in patients with borderline personality disorder, cluster-C personality disorders, and non-patients," *Psychiatry research,* vol. 273, pp. 712-718, 2019.

[165] F. Pistoia, M. Conson, A. Carolei, M. G. Dema, A. Splendiani, G. Curcio and S. Sacco, "Post-earthquake distress and development of emotional expertise in young adults," *Frontiers in behavioral neuroscience,* vol. 12, p. 91, 2018.

[166] A. Navarro Martinez, "Medición de las Respuestas Emocionales en la Violencia contra las Mujeres: Una Revisión Sistemática," 2019.

[167] I. Dominguez-Centeno, R. Jurado-Barba, A. Sion, A. Martínez-Maldonado, G. Castillo-Parra, F. López-Muñoz, I. Rubio and I. Martínez-Gras, "Psychophysiological Correlates of Emotional-and Alcohol-Related Cues Processing in Offspring of Alcohol-Dependent Patients," *Alcohol and Alcoholism,* vol. 55, no. 4, pp. 374-381, 2020.

[168] E. S. Dan-Glauser and K. R. Scherer, "The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and

normative significance," *Behavior research methods,* vol. 43, no. 2, pp. 468-477, 2011.

[169] A. Betella and P. F. Verschure, "The affective slider: A digital self-assessment scale for the measurement of human emotions," *PloS one,* vol. 11, no. 2, p. e148037, 2016.

[170] W. Huang and S. Chiang, "The international affective picture system (IAPS)," *Advances in Psychology,* vol. 4, no. 2, pp. 202-209, 2014.

[171] C. L. Mul, S. D. Stagg, B. Herbelin and J. E. Aspell, "The feeling of me feeling for you: Interoception, alexithymia and empathy in autism," *Journal of Autism and Developmental Disorders,* vol. 48, no. 9, pp. 2953-2967, 2018.

[172] Y. Moriguchi, J. Decety, T. Ohnishi, M. Maeda, T. Mori, K. Nemoto, H. Matsuda and G. Komaki, "Empathy and judging other's pain: an fMRI study of alexithymia," *Cerebral Cortex,* vol. 17, no. 9, pp. 2223-2234, 2007.

[173] A. M. Meneghini, R. Sartori and L. Cunico, "The Italian adaptation of the Balanced Emotional Empathy Scale (BEES) by Albert Mehrabian," Giunti Organizzazioni Speciali, Firenze, 2012.

[174] R. M. Bagby, J. D. Parker and G. J. Taylor, "The twenty-item Toronto Alexithymia Scale-I. Item selection and cross-validation of the factor structure," *Journal of Psychosomatic Research,* vol. 38, no. 1, pp. 23-32, 1994.

[175] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of emotion*, 1980, pp. 3-33.

[176] J. A. Coan and J. J. Allen, Handbook of emotion elicitation and assessment, New York: Oxford University Press, 2007.

[177] M. M. Bradley, M. Codispoti, D. Sabatinelli and P. J. Lang, "Emotion and motivation II: sex differences in picture processing," *Emotion,* vol. 1, no. 3, pp. 300-319, 2001.

[178] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations*, 2015.

[179] S. Edwards and P. M. Salkovskis, "An experimental demonstration that fear, but not disgust, is associated with return of fear in phobias," *Journal of Anxiety Disorders,* vol. 20, no. 1, pp. 58-71, 2006.

[180] P. Michel and R. El Kaliouby, "Real time facial expression recognition in video using support vector machines," in *Proceedings of the 5th international conference on Multimodal interfaces*, 2003.

[181] C. Cao, Q. Hou and K. Zhou, "Displaced dynamic expression regression for real-time facial tracking and animation," *ACM Transactions on graphics (TOG),* vol. 33, no. 4, pp. 1-10, 2014.

[182] J. Cockburn, M. Bartlett, J. Tanaka, J. Movellan, M. Pierce and R. Schultz, "SmileMaze: A tutoring system in real-time facial expression perception and production in children with autism spectrum disorder," in *ECAG 2008 workshop facial and bodily expressions for control and adaptation of games*, Amsterdam, 2008.

[183] M. Gao, J. Jiang, G. Zou, V. John and Z. Liu, "RGB-D-based object recognition using multimodal convolutional neural networks: A survey," *IEEE Access,* vol. 7, pp. 43110-43136, 2019.

[184] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *International Conference on Intelligent Robots and Systems*, 2015.

[185] H. Li, J. Sun, Z. Xu and L. Chen, "Multimodal 2D+ 3D facial expression recognition with deep fusion convolutional neural network," *IEEE Transactions on Multimedia,* vol. 19, no. 12, pp. 2816-2831, 2017.

[186] J. Li, Y. Mi, G. Li and Z. Ju, "CNN-Based Facial Expression Recognition from Annotated RGB-D Images for Human–Robot Interaction," *International Journal of Humanoid Robotics,* vol. 16, no. 4, p. 1941002, 2019.

[187] "Intel RealSense SR300 datasheet," [Online]. Available: https://www.intelrealsense.com/wp-content/uploads/2019/07/RealSense_SR30x_Product_Datasheet_Rev_002.pdf.

[188] A. Dantcheva and J. L. Dugelay, "Female facial aesthetics based on soft biometrics and photo-quality," *IEEE Conference on Institute for Computational and Mathematical Engineering,* vol. 5, pp. 1-6, 2011.

[189] R. Min, J. Choi, G. Medioni and J. L. Dugelay, "Real-Time 3D Face Identification from a Depth Camera," in *21st International Conference on Pattern Recognition*, 2012.

[190] [Online]. Available: https://www.3dlab.polito.it/.

[191] L. Ulrich, J. L. Dugelay, E. Vezzetti, S. Moos and F. Marcolin, "Perspective Morphometric Criteria for Facial Beauty and Proportion Assessment," *Applied Sciences,* vol. 10, no. 1, p. 8, 2020.

[192] S. M. Nelson, "Diversity of the Upper Paleolithic "Venus Figurines and Archeological Mythology"," *Archeological Papers of the American Anthropological Association,* vol. 2, no. 1, pp. 11-22, 1990.

[193] F. B. Naini, M. T. Cobourne, F. McDonald and A. N. A. Donaldson, "The influence of craniofacial to standing height proportion on perceived attractiveness," *International journal of oral and maxillofacial surgery,* vol. 37, no. 10, pp. 877-885, 2008.

[194] D. De Campos, T. Malysz, J. A. Bonatto-Costa, G. Pereira Jotz, L. Pinto de Oliveira Junior and A. Oxley da Rocha, "More than a neuroanatomical representation in The Creation of Adam by Michelangelo Buonarroti, a representation of the Golden Ratio," *Clinical Anatomy,* vol. 28, no. 6, pp. 702-705, 2015.

[195] Y. Jefferson, "Facial beauty--establishing a universal standard," *International journal of orthodontics,* vol. 15, no. 1, pp. 9-22, 2004.

[196] K. Schmid, D. Marx and A. Samal, "Computation of a face attractiveness index based on neoclassical canons, symmetry, and golden ratios," *Pattern Recognition,* vol. 41, no. 8, pp. 2710-2717, 2008.

[197] B. W. Baker and M. G. Woods, "The role of the divine proportion in the esthetic improvement of patients undergoing combined orthodontic/orthognathic surgical treatment," *The International journal of adult orthodontics and orthognathic surgery,* vol. 16, no. 2, pp. 108-120, 2001.

[198] E. Holland, "Marquardt's phi mask: pitfalls of relying on fashion models and the golden ratio to describe a beautiful face," *Aesthetic plastic surgery,* vol. 32, no. 2, pp. 200-208, 2008.

[199] P. M. Pallett, S. Link and K. Lee, " New "golden" ratios for facial beauty," *Vision research,* vol. 50, no. 2, pp. 149-154, 2010.

[200] D. Zhang, Q. Zhao and F. Chen, "Quantitative analysis of human facial beauty using geometric features," *Pattern recognition,* vol. 44, no. 4, pp. 940-950, 2011.

[201] R. J. Edler, "Background considerations to facial aesthetics," *Journal of orthodontics,* vol. 28, no. 2, pp. 159-168, 2001.

[202] D. R. Valenzano, A. Mennucci, G. Tartarelli and A. Cellerino, "Shape analysis of female facial attractiveness," *Vision research,* vol. 46, no. 8-9, pp. 1282-1291, 2006.

[203] G. Dimitriadis, "From Palaeolithic "Venus" up to the anthropomorphic statue-menhir: The ideological evolution of the human body in prehistoric art," *International Congress Series,* vol. 1286, pp. 7-12, 2006.

[204] H. W. Janson and A. Janson, History of Art, New York: Harry N. Abrams, 1991.

[205] C. Bax, The beauty of women, London: Frederick Muller Ltd., 1946.

[206] S. Romm, "The changing face of beauty," *Aesthetic plastic surgery,* vol. 13, no. 2, pp. 91-98, 1989.

[207] A. M. Cheney, ""Most Girls Want to be Skinny" Body (Dis) Satisfaction Among Ethnically Diverse Women," *Qualitative Health Research,* vol. 21, no. 10, pp. 1347-1359, 2011.

[208] A. H. Iliffe, "A study of preferences in feminine beauty," *British Journal of Psychology,,* vol. 51, no. 3, pp. 267-273, 1960.

[209] J. R. Udry, "Structural correlates of feminine beauty preferences in Britain and the United States: A comparison," *Sociology & Social Research,* vol. 49, no. 3, pp. 330-342, 1965.

[210] M. R. Cunningham, "Measuring the physical in physical attractiveness: quasi-experiments on the sociobiology of female facial beauty," *Journal of personality and social psychology,* vol. 50, no. 5, pp. 925-935, 1986.

[211] C. Sforza, A. Laino, R. D'Alessio, C. Dellavia, G. Grandi and V. F. Ferrario, "Three-dimensional facial morphometry of attractive children

and normal children in the deciduous and early mixed dentition," *The Angle Orthodontist,* vol. 77, no. 6, pp. 1025-1033, 2007.

[212] C. Sforza, A. Laino, R. D'Alessio, G. Grandi, G. M. Tartaglia and V. F. Ferrario, "Soft-tissue facial characteristics of attractive and normal adolescent boys and girls," *The Angle Orthodontist,* vol. 78, no. 5, pp. 799-807, 2008.

[213] Y. S. Jayaratne, C. K. Deutsch, C. P. McGrath and R. A. Zwahlen, "Are neoclassical canons valid for southern Chinese faces?," *PloS one,* vol. 7, no. 12, p. e52593, 2012.

[214] V. F. Ferrario, C. Sforza, C. E. Poggio and G. Tartaglia, "Facial morphometry of television actresses compared with normal women," *Journal of oral and maxillofacial surgery,* vol. 53, no. 9, pp. 1008-1014, 1995.

[215] C. Sforza, A. Laino, R. D'Alessio, G. Grandi, M. Binelli and V. F. Ferrario, "Soft-tissue facial characteristics of attractive Italian women as compared to normal women," *Angle Orthodontist,* vol. 79, no. 1, pp. 17-23, 2009.

[216] L. Galantucci, R. Deli, A. Laino, E. Di Gioia, R. D’Alessio, F. Lavecchia, G. Percoco and C. Savastano, "Three-dimensional anthropometric database of attractive Caucasian women: standards and comparisons," *The Journal of craniofacial surgery,* vol. 27, no. 7, pp. 1884-1895, 2016.

[217] E. C. Olivetti, S. Nicotera, F. Marcolin, E. Vezzetti, J. Sotong, E. Zavattero and G. Ramieri, "3D Soft-tissue prediction methodologies for orthognathic surgery—A literature review," *Applied Sciences,* vol. 9, no. 21, p. 4550, 2019.

[218] H. Peck and S. Peck, "A concept of facial esthetics," *The Angle Orthodontist,* vol. 40, no. 4, pp. 284-317, 1970.

[219] O. H. Karatas and E. Toy, "Three-dimensional imaging techniques: A literature review," *European journal of dentistry,* vol. 8, no. 1, pp. 132-140, 2014.

[220] J. Plooij, G. Swennen, F. Rangel, T. Maal, F. Schutyser and E. Bronkhorst, "Evaluation of reproducibility and reliability of 3D soft tissue analysis using 3D stereophotogrammetry," *International journal of oral and maxillofacial surgery,* vol. 38, no. 3, pp. 267-273, 2009.

[221] R. Deli, L. Galantucci, A. Laino, R. D'Alessio, E. Di Gioia, C. Savastano, F. Lavecchia and G. Percoco, "Three-dimensional methodology for photogrammetric acquisition of the soft tissues of the face: a new clinical-instrumental protocol," *Progress in orthodontics,* vol. 14, no. 1, p. 32, 2013.

[222] C. H. Kau, "Creation of the virtual patient for the study of facial morphology," *Facial Plastic Surgery Clinics of North America,* vol. 19, no. 4, pp. 615-622, 2011.

[223] N. F. Berlin, P. Berssenbrügge, C. Runte, K. Wermker, S. Jung, J. Kleinheinz and D. Dirksen, "Quantification of facial asymmetry by 2D analysis–A comparison of recent approaches," *Journal of Cranio-Maxillofacial Surgery,* vol. 42, no. 3, pp. 265-271, 2014.

[224] C. Borelli and M. Berneburg, "“Beauty lies in the eye of the beholder”? Aspects of beauty and attractiveness," *Journal of the German Society of Dermatology,* vol. 8, no. 5, pp. 326-330, 2010.

[225] E. Vezzetti and F. Marcolin, "Geometrical descriptors for human face morphological analysis and recognition," *Robotics and Autonomous Systems,* vol. 60, no. 6, pp. 928-939, 2012.

[226] C. Sforza and V. F. Ferrario, "Soft-tissue facial anthropometry in three dimensions: from anatomical landmarks to digital morphology in research, clinics and forensic anthropology," *Journal of Anthropological Sciences,* vol. 84, pp. 97-124, 2006.

[227] P. Hammond, T. Hutton, J. Allanson, L. Campbell, R. Hennekam, S. Holden and M. Patton, "3D analysis of facial morphology," *American journal of medical genetics,* vol. 126 A, no. 4, pp. 339-348, 2004.

[228] V. Nanda, B. Gutman, E. Bar, S. Alghamdi, S. Tetradis, A. Lusis, E. Eskin and W. Moon, "Quantitative analysis of 3-dimensional facial soft tissue photographic images: technical methods and clinical application," *Progress in orthodontics,* vol. 16, no. 1, pp. 1-9, 2015.

[229] W. R. Proffit, R. P. White and D. M. Sarver, Contemporary treatment of dentofacial deformity, St. Louis: Mosby, 2003.

[230] J. Fan, K. Chau, X. Wan, L. Zhai and E. Lau, "Prediction of facial attractiveness from facial proportions," *Pattern Recognition,* vol. 45, no. 6, pp. 2326-2334, 2012.

[231] N. Alyuz, B. Gökberk and L. Akarun, "Regional registration for expression resistant 3-D face recognition," *IEEE Transactions on Information Forensics and Security,* vol. 5, no. 3, pp. 425-440, 2010.

[232] N. Dagnes, F. Marcolin, F. Nonis, S. Tornincasa and E. Vezzetti, "3D geometry-based face recognition in presence of eye and mouth occlusion," *International Journal on Interactive Design and Manufacturing (IJIDeM),* vol. 13, no. 4, pp. 1617-1635, 2019.

[233] L. Yin, X. Wei, Y. Sun, J. Wang and M. J. Rosato, "A 3D Facial Expression Database For Facial Behavior Research," in *7th International Conference on Automatic Face and Gesture Recognition*, 2006.

[234] A. Savran, N. Alyüz, H. Dibeklioˇ glu, O. Çeliktutan, B. Gökberk, B. Sankur and L. Akarun, "Bosphorus database for 3D face analysis," in *European workshop on biometrics and identity management* , 2008.

[235] L. G. Farkas, Anthropometry of the Head and Face, vol. 107, 1994, pp. 112-112.

[236] G. R. Swennen, F. A. Schutyser and J. E. Hausamen, Three-dimensional cephalometry: a color atlas and manual, Springer Science & Business Media, 2005.

[237] E. Vezzetti and M. F., "Geometrical descriptors for human face morphological analysis and recognition," *Robotics and Autonomous Systems,* vol. 60, no. 6, pp. 928-939, 2012.

[238] J. J. Koenderink and A. J. Van Doorn, "Surface shape and curvature scales," *Image and vision computing,* vol. 10, no. 8, pp. 557-564, 1992.