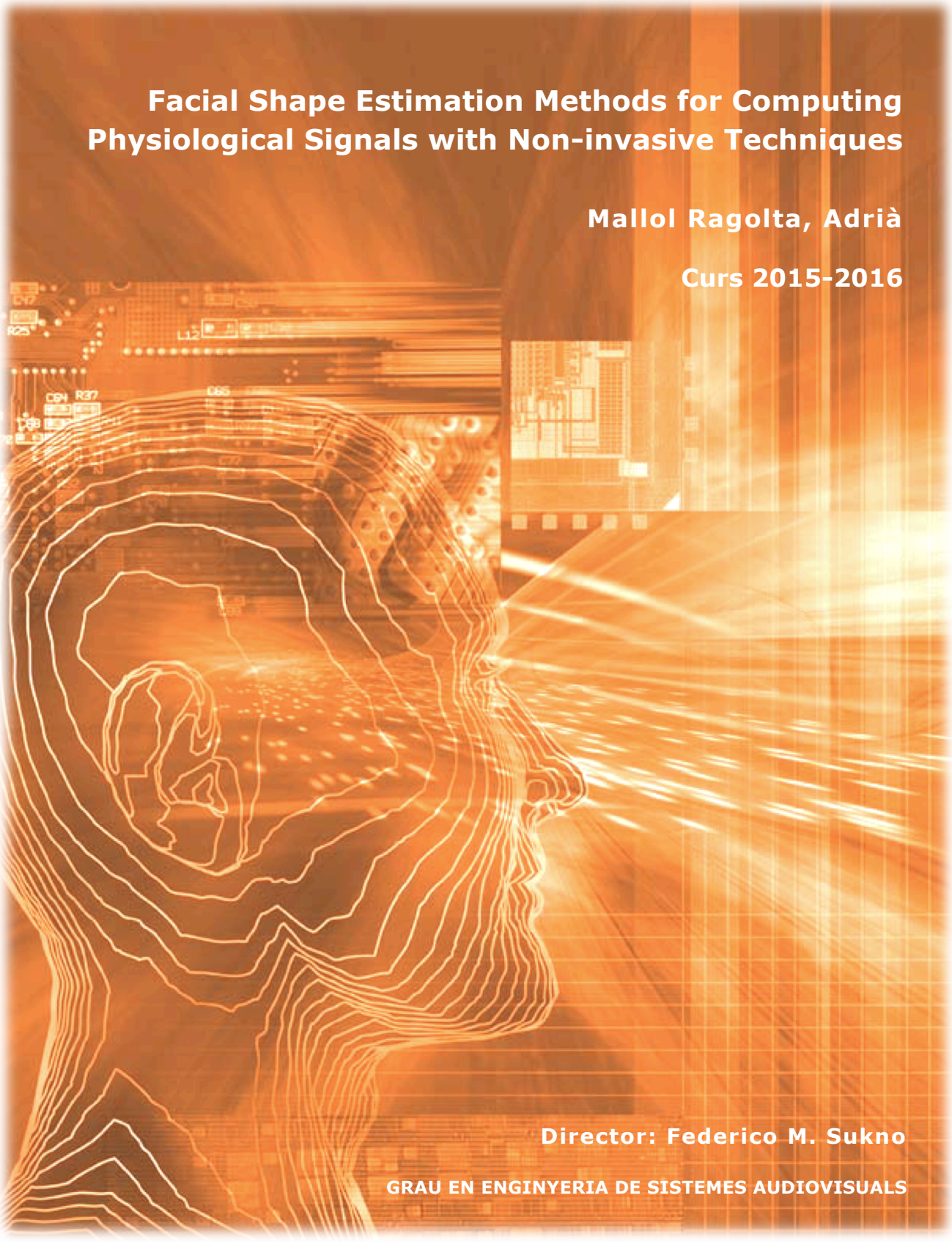

Facial Shape Estimation Methods for Computing Physiological Signals with Non-invasive Techniques

Mallol Ragolta, Adrià

Curs 2015-2016

Director: Federico M. Sukno

GRAU EN ENGINYERIA DE SISTEMES AUDIOVISUALS

Universitat Pompeu Fabra Barcelona
Escola Superior Politècnica

Treball de Fi de Grau

Facial Shape Estimation Methods for Computing Physiological Signals with Non-invasive Techniques

Adrià Mallol Ragolta

UNDERGRADUATE THESIS / 2016

THESIS ADVISOR
Dr. Federico M. Sukno
Department of Information and Communication Technologies

Als meus pares.
Pel vostre esforç incalculable, pel vostre suport incondicional i per la vostra estima inigualable.
A.

Acknowledgements

First of all, I would like to express my profound gratitude to my advisor, Dr. Federico Sukno, for his continuous teaching and guidance throughout this project, for his confidence and determination in overcoming challenging issues, and for his comments and corrections that have improved the quality of the presented work.

I would also like to thank CMTech Research Group for hosting me, especially Dr. Oriol Martínez for the shared knowledge, Professor Adrià Ruiz for introducing me on the image-based facial analysis field, as well as Dr. Pol Cirujeda and Dr. Brais Martínez for their expertise and technical support.

Thanks also to Alex Pereda and Miguel Barreda from Eurecat, who loaned us the specialized equipment to record ground truth information.

I am especially grateful to the finantial support provided by the *Ministerio de Educación, Cultura y Deporte* and the *Agència de Gestió d'Ajuts Universitaris i de Recerca* (AGAUR) through the COLAB scholarship program.

Finally, a special thanks to my parents, Anselm and Elena, my sister, Ariadna, and my spouse, Lídia, for making me feel tenacious, courageous, and self-confident.

Abstract

For the last few years, there has been an increasing interest in the estimation of physiological data in the context of human behavior understanding and human computer interaction with many different applications such as the automatic recognition of either the physical or the emotional state of a person, or telehealth, among others. Nowadays, however, an important drawback is that these signals can only be measured by means of invasive techniques such as electrodes or pulse oximeters.

In this project we address estimation of physiological data by means of indirect, non-invasive measurements. Specifically, we focus on the estimation of the heart rate by amplifying the subtle color variations that appear in the facial skin due to the blood stream pulse. Hence, we develop a fully automatic scheme based on landmark localization in order to segment the facial skin region where to extract the signals that are processed and analyzed with the purpose to estimate the heart rate.

Evaluation of the proposed scheme is provided in quantitative terms with respect to ground truth obtained by means of invasive measurements. To this end, we have gathered a small multimodal database comprising high-resolution facial videos and ECG recordings from specialized equipment.

The main conclusion of this project is the severe difficulty in the measurement of the heart rate from dynamic videos, in which head rotations in both 2D and 3D appear. However, the proposed scheme was able to accurately estimate the heart rate in static videos: in experiments on seventy-eight videos from thirteen subjects we obtained a median error of 2.64 beats per minute, which is comparable to the state of the art.

Resum

En els darrers anys, s'ha produït un interès creixent en l'estimació de senyals fisiològics sota el paradigma de la comprensió del comportament humà i de la interacció persona-ordinador amb múltiples aplicacions tals com el reconeixement automàtic de l'estat, tant físic com emocional, d'una persona, o la telemedicina, entre d'altres. Actualment, però, una de les limitacions més importants és que aquests senyals només es poden mesurar a través de tècniques invasives com són els electrodes o els pulsioxímetres.

En aquest projecte ens centrarem en l'estimació de senyals fisiològics a través de mesures indirectes i no-invasives. Concretament, ens centrarem en l'estimació del ritme cardíac a través de l'amplificació de les variacions subtils de color que es produeixen a la regió cutània de la cara. Amb aquesta finalitat, hem desenvolupat un paradigma totalment automàtic basat en la localització de punts característics per tal de segmentar la regió cutània de la cara de la qual n'extreurem els senyals que seran processats i analitzats amb l'objectiu d'estimar el ritme cardíac.

El paradigma proposat s'avalua quantitativament respecte dades reals obtingudes a través de mesures invasives. Amb aquest propòsit, hem compilat una petita base de dades multimodal, la qual conté videos facials a alta resolució i electrocardiogrames enregistrats amb equipament especialitzat.

La conclusió principal que es pot extreure d'aquest projecte és la dificultat de mesurar la freqüència cardíaca a partir de vídeos dinàmics, en els quals la posició del cap dels subjectes canvia tot descrivint rotacions en 2D i en 3D. Això no obstant, el paradigma proposat ha sigut capaç d'estimar amb precisió el ritme cardíac en vídeos estàtics: en experiments realitzats en setanta-vuit vídeos de tretze subjectes diferents, vam obtenir un error de 2,64 batecs per minut de mediana, el qual és comparable amb l'estat de l'art.

Contents

List of figures

List of tables

Chapter 1

INTRODUCTION

Computers are more than just passive machines. Whereas humans have eyes, ears and mouth, computers have a camera, as it is the webcam, a microphone and loudspeakers, respectively. Therefore, we cannot only think in human-computer interaction based on the usage of some hardware as the mouse or the keyboard, but also, in an interaction based on the information perceived by computers' *senses*, captured by their sensors.

For such purpose, research ought to be focused on the information from the users that could be gathered using these sensors. In this direction, recent studies have demonstrated that some humans' physiological signals could be measured using cameras, and among those signals there is the heart rate.

In this project we will focus on the measurement of the heart rate from facial videos; which implies a measurement with no contact between the user's skin and the sensor. This technology might well have many different applications. For instance, it could be used in emotion recognition to be combined with other facial clues in order to determine the real, not only the apparent, emotional state of a subject. Moreover, it could also be used in telemedicine and telehealth in order to monitor a patient avoiding the usage of the uncomfortable equipment currently used; in lie detection ap-

plications, since heart rate tends to increase when lying; or in heart attack prevention through constant monitorization using inexpensive equipment, as video cameras.

1.1 Motivation

Heart rate is a powerful signal, since it provides lots of valuable information related with our health. However, people's heart rate might be altered because of the discomfort caused by the current procedure used to measure the heart rate. Hence, the usage of non-invasive techniques, which mean that there is no contact with the skin of the patient, could be an interesting way to increase the comfort on this procedure and the accuracy on the measurement.

All things considered, I got attracted by the idea of contributing to the research in this field by studying the feasibility of a scheme able to measure the heart rate of a subject given a video of his or her face.

1.2 Project Objectives

The main goal of this project is the implementation of a scheme able to measure the heart rate of a subject given a facial video. At the same time, we have defined a list of objectives to accomplish, which have also guided the development of this project. The objectives of this project are listed below.

- Collection of the DARHR Database, which stands for *Data Acquisition Related with Heart Rate*. This database will contain facial videos from different subjects at different heart rates, measuring this physiological signal with conventional equipment based on electrodes to serve as ground truth.
- Implementation of a landmark localization software to identify the facial elements of the facial input videos.

- Determination of the best model to be used in the landmark localization stage of our scheme.
- Skin region segmentation based on landmark information to determine the region where to extract the heart rate.
- Measurement of the oscillation caused by the bloodstream on the skin region segmented based on color intensity changes.
- Analysis of the best scheme to determine the heart rate with the highest agreement possible among all available signals.
- Skin color amplification of the original video according to the estimated heart rate as the ouput of the scheme.
- Optimization of the code in order to achieve a performance as close as possible to real-time.
- Experimental evaluation to qualify the performance of the different stages of the proposed scheme.

1.3 Project Structure

This project is structured as follows. Chapter 1 provides an overview of the project with the motivation to carry out this research and its goals. In Chapter 2 we briefly describe some of the previous work that has been done in the field, while in Chapter 3 we explain the materials that have been used and those especially created for this project. In this chapter, the implementation of the scheme presented in this project is also detailed. The results obtained on the different stages of our implementation as well as a global evaluation of our system are presented and analyzed in Chapter 4, while Chapter 5 summarizes and concludes this report.

Chapter 2

PREVIOUS WORK

In this chapter we will go through the literature recently published on the field of physiological signals measurement using non-invasive techniques. Firstly, some general background will be provided in order to contextualize the project. Secondly, state-of-the-art methods will be presented and globally explained. Finally, novelties of our scheme against previous stated methods will be highlighted.

2.1 General Background

Heart rate measurements have been performed since long ago. The most common method used to this end is the electrocardiography (ECG), which is considered an invasive method[1] since electrodes need to be in contact with the subject's skin to perform the measurement. However, thermal imaging [3], Doppler phenomenon, both optical [4] and ultrasonic [5], piezoelectric measurements [6] and photoplethysmography [7] are other methods which might well be used in order to measure the cardiac pulse.

[1] Although invasive methods in the medical field usually involve entry into the body [2], in this project invasive methods will be considered those in which contact between the subject's skin and the equipment to perform the measurement is required. Consequently, non-invasive methods will be considered those in which this contact is not necessary.

Photoplethysmography is based on the measurement of light absortion changes that occur in the dermis. The theoretical foundation behind is that blood absorbs more light than surrounding tissue and, as a result, it correspondingly affects the amount of reflected light. Therefore, there is a relationship between the reflected light and the blood stream. Verkruysse *et al.* [8] demonstrated the feasibility of measuring the pulse from human faces using ambient light and a simple camera.

2.2 State-of-the-art

Poh *et al.* [9, 10] proposed a framework to measure the cardiac pulse from color videos of the human face based on automatic face tracking. Then, the pixel values inside the region of interest (ROI) were averaged in order to create a single signal from the overall video. Poh *et al.* realized that the physiological signal of interest usually fell in the same frequency band as the noise; therefore, in order to solve this issue, they tried to remove this noise using Independent Component Analysis [11]. Lewandowska *et al.* [12] estimated the heart rate from two different regions of interest: the first one was a rectangle containing the whole facial region, while the second one was a rectangular-shaped region of the forehead. The pixels inside were then averaged and multiple red, green and blue (RGB) channel combinations were processed using principal component analysis (PCA) [13] and independent component analysis (ICA) [14] in order to estimate the heart rate. They finally concluded that the forehead ROI was representative for the whole facial region and that the analysis done using PCA was faster than using ICA.

Wu *et al.* [15] proved the existence of color intensity variations in the facial region invisible to the naked eye that might well be caused by the blood stream, despite neither performing any kind of segmentation in the input image nor estimating the heart rate.

Soleymani *et al.* [16] gathered a multimodal database where subjects were recorded when affectively stimulated. Moreover, recordings were synchronized with supplementary information from the subjects such as the ECG, among others. Therefore, the provided database has been commonly used in the literature since its publication in order to test the different methods proposed in the field of affective computing because of the ECG recorded that could be used as ground truth.

Recently, Li *et al.* [17] proposed a method focused on reducing effects of both illumination variations and subjects' motion. The ROI they used was a region determined by a subset of detected landmarks and the green channel of all pixels inside were averaged and used as the raw signal. After some signal processing to perform illumination rectification and non-rigid motion elimination, the power spectral density of the signal was computed using the Welch's method [18] and the frequency with the greatest power was defined as the frequency corresponding to the heart rate.

Later on, Osman *et al.* [19] proposed a method which used machine learning techniques to detect heart beats from facial pixel intensity changes. Their idea was to learn a model able to predict whether given a windowed patch of a raw signal, it could be considered as an intensity peak or not. If predicted intensity peaks were greater than a certain threshold, they were considered as reliable peaks and the time difference between two consecutive reliable peaks was defined as the period of the fundamental frequency related to the heart rate.

Finally, Lam and Kuno [20] proposed a skin appearance model in order to model how illumination changes and cardiac activity affected the evolution of facial pixels over time. Using this model, they estimated the plethysmographic signal by computing FastICA from a pair of facial pixels over time. After extracting the signal, they computed its power spectral density through the Welch's method and the peak with the greatest power was considered as the fundamental frequency corresponding

to the heart rate. Finally, the heart rate of the overall video was computed through a majority voting scheme.

2.3 Novelties

In the scheme we propose, the region of interest (ROI) is adaptated to rotations or movements of the subject's face and each pixel inside is analyzed independently. This means that one signal is extracted from each pixel inside the ROI. Moreover, from each extracted signal an independent frequency analysis is performed in order to take into account the possible disparity between the frequencies estimated at different facial regions.

In addition, the heart rate is computed periodically; *i.e.*, its estimation is performed every certain number of frames. The reason for such performance is that heart rate is an unsteady signal since it changes over time for different reasons, which will be mentioned later on. Therefore, estimating a single frequency for a whole video is an unrealistic situation and, from our point of view, the accuracy in the measurement of heart rate fluctuations are both a more realistic and a more interesting scenario to be studied.

Chapter 3

METHODOLOGY

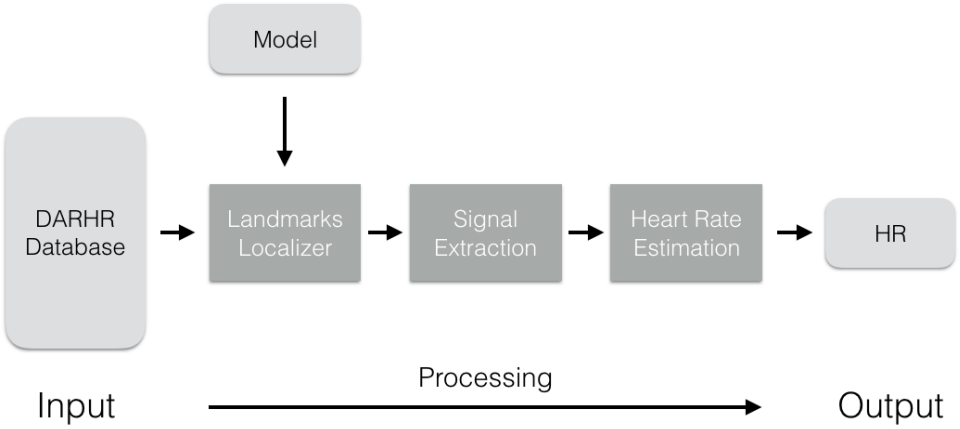

Figure 3.1: Workflow diagram of the overall scheme.

In this chapter the scheme proposed in our undergraduate thesis for estimating the heart rate of a facial video is detailed. Our method takes a facial video from DARHR database as input data. Right after, the facial region corresponding to the skin is determined using a landmark localizer software. Once determined, signals from pixels inside the region defined

by the landmarks are built and then processed in order to perform the appropriate signal analysis to estimate the heart rate, which is the output of our scheme. A workflow diagram of the overall scheme proposed can be seen in figure 3.1.

The contents of this chapter are divided in four different sections. In the first section, the materials created for this project are explained as well as the equipment used and the recording protocol followed. In the second section, the landmark localizer implementation is detailed, as well as its learning model procedure. In the third section, we explain the process of signal extraction from facial images, while in the fourth section the signal processing and the strategy used to estimate the heart rate from facial videos are stated.

3.1 DARHR Database

One of the goals of this project was the gathering of our own multimodal database containing both facial videos and electrocardiogram data measured with conventional equipment in order to serve as ground truth. The DARHR database, which stands for *Data Acquisition Related with Heart Rate*, consists of seventy-eight facial videos from thirteen different subjects at a ratio of six videos per subject. All the videos for each subject were recorded at three fixed time intervals after asking subjects to do some exercise. Moreover, at each interval the recording and the measurement were repeated.

The purpose of this database is the creation of materials to evaluate the performance of our method with fair comparisons between the heart rate estimated and the real one. For this reason, for each facial video the corresponding electrocardiogram measured using invasive equipment is also provided.

3.1.1 Settings

In this section the room conditioning and the equipment used to gather the DARHR database is described. The video equipment that was used to record facial videos is detailed as well as the specialized equipment used to invasively measure the heart rate, which will serve as ground truth.

Lighting Conditions

Different works have demonstrated the feasibility to measure the cardiac pulse from the evolution of pixel-color intensity changes over time [8], [9], [17]. Since we aim at measuring subtle intensity changes, the lighting of the room where the videos are recorded is a determining factor to correctly analyze plethysmographic signals. Therefore, we evaluated two scenarios where to record the videos: the first one was a room with natural illumination, while the second one was a room with artificial illumination.

After recording some data in both scenarios, a comparison of the dynamic range between the obtained videos is shown in figure 3.2; this comparison is done through averaging the green channel, the channel with the greatest plethysmographic component [8], of the ten first frames of both videos. Analyzing figure 3.2, it can be seen that the dynamic range of the video recorded with natural lighting is wider than the dynamic range of the video recorded with artificial lighting.

The intensity changes that we want to measure are extremely subtle; as a consequence, external noise that could be added to the signal might well have catastrophic consequences. Thus, in order to avoid induced noise caused by artificial illumination sources in our videos, we decided to record DARHR database in a room with natural lighting. In addition, in order to assure that the intensity ambient light was as constant as possible throughout the whole database, the videos were recorded at the same time slot of winter's days: from 2:30 p.m. to 5:30 p.m.

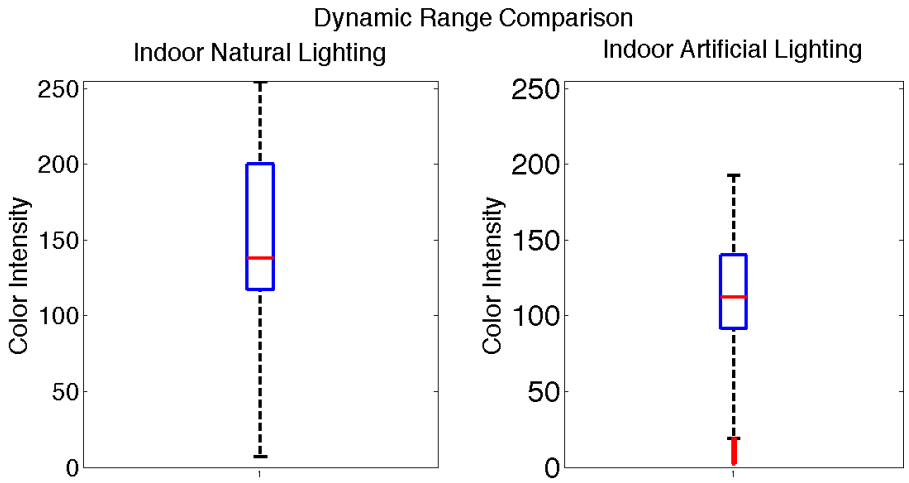

Figure 3.2: Dynamic range comparison obtained by averaging the green channel of the first ten frames of a video recorded with natural lighting (left) and a video recorded with artificial lighting (right). The red line represents the median of the data, while the lower and the upper sides of the blue rectangle corresponds to the first and the third quartile, respectively. The dashed lines represent the data that fit inside 1.5 times the interquartile range beyond the third quartile and below the first quartile, while red crosses correspond to the outliers.

Video Recording Materials

The facial videos were recorded using a Panasonic HPX 1711 camera at a progressive scanning of $720p$ and at a frame rate of 50 frames per second *(fps)* with a maximum constant gain of $12\ dB$. Moreover, the iris of the camera was set to its maximum possible level right before the appearance of the zebra effect inside the recording area, which helped us to ensure the correct exposure of the scene.

In addition, the camera was placed in the front of the subject at an approximate distance of 1.5 meters between them. This configuration was preserved for all recordings.

Physiological Signals Recording Materials

The heart rate of the subjects was recorded using BIOPAC MP150 hardware[1] and processed with AcqKnowledge software[2]. The most relevant specifications of the hardware are highlighted in the list below:

- Absolut Maximum Input: $\pm 15V$
- Operationa Input Voltage: $\pm 10V$
- Accuracy ($\%$ of the Full-Scale Range): ± 0.003

BIOPAC MP150 is a powerful and accurate tool to measure a wide range of physiological signals, which are measured from especific modules that can be connected to the central processing module. In our case, we only needed the ECG100C module, which is responsible of electrocardiogram measurements, with a sampling rate of $1000\ Hz$. Electrocardiogram waveform provides many information; however, we were only interested in the QRS complex since the heart rate can be computed as the inverse of the period between consecutive R peaks. As a consequence, the measured signal was automatically filtered by the hardware using an R-filter in order to only record the information we are interested in.

In order to measure the electrocardiogram, some electrodes must be plugged into the ECG100C module and ought to be placed on the skin of the subject. We decided to place electrodes according to a bipolar recording method. In this method, three electrodes are used and the voltage difference between two of them is measured with respect to the third, which is used as a reference. In our case, the voltage difference was measured between the inner right wrist and the inner left wrist with respect to the voltage at the inner right elbow.

[1] BIOPAC Systems, Inc.; Goleta, CA; *http://www.biopac.com/product/mp150-data-acquisition-systems/*

[2] Version *4.2*; BIOPAC Systems, Inc.; Goleta, CA; *http://www.biopac.com/manual/acqknowledge-tutorial/*

On a post-processing stage, the recorded signals were exported into a *.mat* file format in order to be readable from Matlab®. Additionally to the high resolution signal, this file also provides a downsampled version of the signal and an average of the whole measurement in order to facilitate the evaluation of the scheme proposed in this project.

3.1.2 Dataset

A total of thirteen subjects with ages from sixteen to fifty years old were recorded in the DARHR database: eight of them were females, while five of them were males. No specific physical condition was required in order to select the candidates for our database as we were interested in recording ordinary people regardless of their physical condition.

All subjects signed a Concern Form in which they explicitly agreed with the usage of the recorded information for reasearch purposes.

3.1.3 Physical Conditioning Prior to Recordings

Heart rate is a physiological signal that changes over time because of the body oxigen demand or the breathing, among other factors. We were interested in recording videos at different heart rates and an easy way to increase the cardiac pulse is doing some exercise. Research on the field of heart rate recovery, [21] and [22], showed how this signal changes in the time after a peak of exercise, which could be modeled as a decreasing exponential function. As a consequence, the highest heart rate can be measured in the first two minutes after the exercise, while its steady state can be reached after ten minutes of the exercise.

The exercise we proposed to our subjects was going up and down a series of stairs equivalent to the height of a three story building. For the reasons stated in the previous paragraph, we decided to do a first measurement right after the exercise and the final one ten minutes later. In addition, we were also interested in doing a measurement five minutes after the exer-

cise; this was a time interval in which the heart rate was expected to be neither maximum nor steady. Therefore, these are the three fixed time intervals at which the videos were recorded.

3.1.4 Data Acquisition Procedure

After welcoming the participants, they were told about the recordings that would be done and their usage in this project. Before exercising, they were asked to sit down in a chair in order to focus the image that would be recorded with the camera. Then, they were requested to do the proposed exercise in order to increase their heart rate.

When the exercise was completed, participants sat down in a chair and three electrodes were placed in the subjects' skin: two in their wrists and one in their right elbow. Righ after, the camera and the *Acqknowledge* software started recording simultaneously. This process was repeated at the three fixed time intervals and, at each interval, it was done twice. Each recording took a minute: during the first twenty seconds subjects were asked to be as static as possible, while from second twenty until the end of the experiment subjects moved and talked with total freedom.

In order to ensure that the procedure and parameters of both the hardware and the software were the same, a recording protocol was written and followed. The detailed protocol is provided in appendix A.

3.1.5 Database Contents

With all the recordings done, the database could be gathered. For each subject and each experiment there is a video file in *.mp4* format and a *.mat* file containing the heart rate, which can serve as ground truth.

The video file was obtained by transcoding the *.mxf* file obtained from the camera into an *.mp4* file using an H.264 codec with a bitrate of $3072\ kb/s$.

The *.mat* file contains the heart rate signal of the overall video sampled at $1000\ Hz$, a downsampled version of it and, finally, a global average of the heart rate signal for the whole video.

3.2 Landmark Localizer

Inside a face, we are able to identify several elements such as the eyebrows, the eyes, the nose or the mouth. In the image-based facial analysis field, the automatic localization of these elements is a fundamental task since it allows many different usages as face identification or emotion recognition, among others.

In the literature, one of the most used strategies is the localization of these elements by placing a set of points on the face. These points are named landmarks and they can be defined as reproducible anatomical points that could be used for biometric purposes. For instance, the corners of the eyes, or the corners of the mouth, are considered landmarks since they correspond to the same anatomical location in any face of any person around the world and, in addition, they are useful for modeling biometrical information from a person. In this project we will use a widespread template consisting of 68 landmarks to completely describe a face: 51 of them are placed inside the face and the others, in the edge of the face.

In our approach, we are interested in localizing these landmarks in order to define the area corresponding to the facial skin where to compute the heart rate. Therefore, in this section we will analyze the landmark localizer stage of our scheme.

3.2.1 Overview

Currently, there are different methods and strategies in order to localize facial landmarks in an image [23], [24]. For this project, a Supervised Descent Method [25] was used due to its conceptual simplicity and per-

formance.

The main idea of this method is the iterative adaptation of an initial set of landmarks until reaching the positions that best describe the real landmarks in a certain image; this iterative process is illustrated in figure 3.3. As any iterative process, its initialization is a crutial step; therefore, it should be determined carefully. In order to succeed in the localization, we need to learn the directions that point to the position of the actual landmarks where the previous ones should be displaced to. Consequently, we need to extract some information from the image in order to model these directions using pattern recognition techniques. Once we have this model computed, we will be able to localize, as accurately as possible, the landmarks on any facial image.

3.2.2 Supervised Descent Method Fundamentals

As it has previously said, to learn the directions where the landmarks ought to be displaced, some information needs to be extracted from the image. The information that Xiong *et al.* [25] extracted from the image was the SIFT descriptor [26] of each landmark since, with this information, a SIFT feature vector could be mapped into a point displacement. In other words, a relationship between the direction and the magnitude of landmarks displacement and the values of this descriptor would be established.

Due to the non-linearity of the SIFT function, a Non-linear Least Squares function could be minimized in order to compute the best approximation to the solution, which means that we will never reach the real solution of the function. This minimization is often computed using the Gauss-Newton algorithm, an iterative algorithm in which at each iteration of the process, the solution is closer and closer to the ideal solution until either fulfilling some convergence criterion or reaching a maximum number of

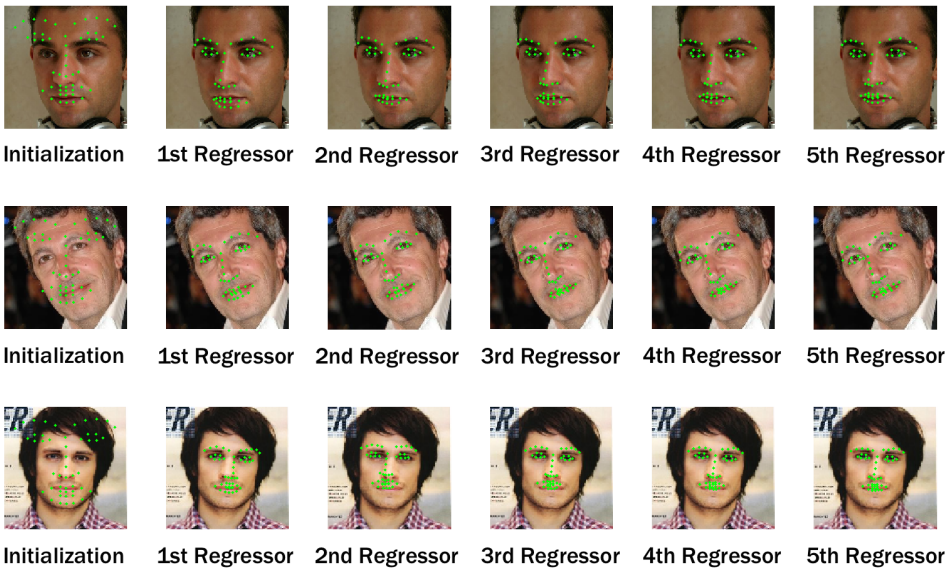

Figure 3.3: Illustration of the iterative process by which the landmarks' set of points is adaptated at each layer of the regressor until reaching a distribution close to their real position. Despite this figure being own developed, original images were extracted from ICCV2013 Challenge Database [1].

iterations. This kind of methods, however, are computationally expensive. Therefore, the intuition of Supervised Descent Method (SDM) [25] is to avoid these calculations by learning the descent directions of the data at each iteration until convergence using pattern recognition techniques.

Now that we have the intuition, let us move to the mathematic analysis of SDM. Firstly, we will start with the simplest case: when having an image as input data I^i of size 128×128 pixels, and p landmarks placed at coordinates $\mathbf{x_0}^i \in \Re^{1\times 2p}$, according to an initial landmarks configuration. Before proceeding further, a remark: subindex 0 denotes that landmarks are placed in some initial location, which is not their real position; their real position is called *Ground Truth* and has the nomenclature $\mathbf{x}^{*i} \in \Re^{1\times 2p}$.

At each landmark we compute its SIFT descriptor, which results in a row vector $\in \Re^{1\times 128}$. Then, we concatenate the SIFT descriptors of all landmarks in a single row vector $\in \Re^{1\times 128p}$, where *p* is the total number of landmarks placed in our image. For simplicity, let us express this vector as ϕ^i. At this point, the goal is to compute a linear regression between the landmarks displacement and their SIFT descriptors vector. The mathematical expression to model this problem is a linear combination of the SIFT descriptors plus a bias term as expressed in equation 3.1, where $\mathbf{x}^{*i}$ and $\mathbf{x_0}^i$ are defined such as

$$\mathbf{x}^{*i} = \begin{bmatrix} x_1^{*i} \\ \vdots \\ x_p^{*i} \\ y_1^{*i} \\ \vdots \\ y_p^{*i} \end{bmatrix}^T$$

and

$$\mathbf{x_0}^i = \begin{bmatrix} x_{0_1}^i \\ \vdots \\ x_{0_p}^i \\ y_{0_1}^i \\ \vdots \\ y_{0_p}^i \end{bmatrix}^T$$

with p corresponding to the total number of landmarks to analyze.

$$\mathbf{x}^{*i} - \mathbf{x_0}^i = \mathbf{\Delta}^i = \phi^i \mathbf{R} + \mathbf{b} \tag{3.1}$$

The system in equation 3.1 would be solved by computing $\mathbf{R}$ and $\mathbf{b}$. However, due to the non-linearity of this system, as previously stated, the way to solve it is by optimization, using an iterative method. Therefore, a sequence of descent directions $\mathbf{R_k}$ and bias terms $\mathbf{b_k}$ need to be computed and, as a result, equation 3.1 could be reformulated as defined

below, where N is the number of iterations required for the convergence of the algorithm.

$$\mathbf{x}^{*i} - \mathbf{x}^i_{k-1} = \phi^i_{k-1}\mathbf{R}_{k-1} + \mathbf{b}_{k-1}, \text{ for } k = 1, \ldots, N \tag{3.2}$$

In addition, at each iteration, after computing $\mathbf{R}_{k-1}$ and $\mathbf{b}_{k-1}$, we are required to update the points $\mathbf{x}^i_{k-1}$ using the equation:

$$\mathbf{x}^i_k = \mathbf{x}^i_{k-1} + \phi^i_{k-1}\mathbf{R}_{k-1} + \mathbf{b}_{k-1} \tag{3.3}$$

This update is necessary since SIFT descriptors ϕ^i should be recalculated according to the new landmark positions. Once all variables are updated, equation 3.2 can be recomputed.

To finish the iterative process, a convergence criterion has to be fulfilled. Ideally, the algorithm ought to finish when $\mathbf{x}^i_k = \mathbf{x}^{*i}$. However, in general, this equality is never reached. In those cases, the algorithm is assumed to converge when $\mathbf{x}^i_k - \mathbf{x}^i_{k-1} < \xi$, with ξ as an accepted threshold, or when a maximum number of iterations is reached.

The most interesting property of this method is that when $\mathbf{R}$ and $\mathbf{b}$ are computed, they can be stored and used in any new facial image as a model in order to perform landmark localization. Given any image outside the training data, a patch containing a single face from the image can be cropped and resized into 128×128 pixels. Right after, landmarks can be initialized and the SIFT descriptor at each landmark could be extracted. Finally, equation 3.3 could be iteratively computed until estimating the best possible landmarks location. A diagram of the global workflow of this method is illustrated in figure 3.4 for both the training and the testing stages.

3.2.3 Mean Shape Computation

Iterative algorithms are so sensitive to their initializations that, often, bad initializations imply not reaching the best possible solution of the functi-

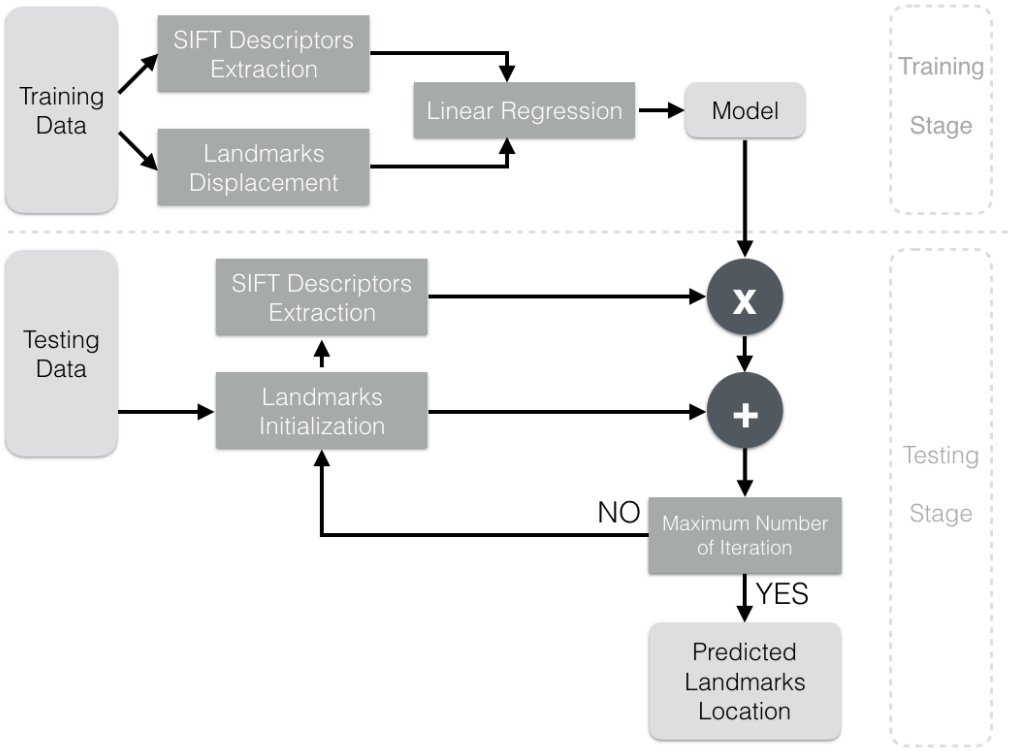

Figure 3.4: Workflow diagram of the Supervised Descent Method for both the training and the testing stage.

on to minimize and, as a result, the solution found can easily be wrong. Therefore, due to its importance, the data initialization is the first topic to focus on.

Our strategy was based on the initialization of our algorithm considering a single set of points reasonably representative of the high variability among facial shapes. Therefore, we are interested in computing a mean facial shape of as many labeled faces as available, which might be thought as the centroid of all facial shape locations.

In order to compute a mean facial shape, a large set of labeled faces is required first. Thus, we used a subset of the data provided in the ICCV2013

Challenge, which included *afw* database [27], *helen* database [28], *ibug* database [1] and *lpfw* database [29]. Every single one of these images was accompanied by a *.pts* file containing the ground truth of the facial shape to analyze. The same process was repeated for all images inside the mentioned subset:

1. Ground truth file reading. The *.pts* file includes the specific coordinates in the 2D plane of 68 annotated landmarks. Moreover, the coordinates are sorted in a way that each row of the file represents a specific landmark, which contributes to easily manage the data from this file.

2. Bounding box estimation from ground truth information. Using ground truth coordinates, a bounding box such that all landmarks fit inside can easily be determined. Its parameters might well be defined by computing minimum and maximum x-coordinates and y-coordinates among all ground truth landmarks.

3. Bounding box enlargement. The size of the bounding box estimated directly from the ground truth was increased in order to add a security region such that all landmarks and their immediate neighbors fit inside. In our case, the bounding box enlargement was about a 30% of its original width and height.

4. Bounding box and ground truth resizing. The variability among face sizes in available images would have a distorting effect on the mean shape computation. Therefore, the facial patch determined by the enlarged bounding box should be resized to a common size, in our case 128×128 pixels, in order to fairly compute the mean shape. As a consequence, ground truth landmarks should also be mapped into their corresponding position in this new image space. This mapping could be easily done by following formula 3.4 for x-coordinates and formula 3.5 for y-coordinates, in which the bounding box is represented by variable BB.

$$GT_{(128\times128)_x} = (GT_{\text{original}_x} - BB_{\text{top left corner}_x}) * \frac{128}{BB_{\text{width}}} \tag{3.4}$$

$$GT_{(128\times128)_y} = (GT_{\text{original}_y} - BB_{\text{top left corner}_y}) * \frac{128}{BB_{\text{height}}} \tag{3.5}$$

Once all images were processed, the mean of all resized ground truth landmarks was computed. As a result, we got a matrix of size 68×2; *i.e.*, the 2D Cartesian coordinates for each landmark, corresponding to the mean shape of all processed images. This mean shape will be an appropriate landmark locations initialization to solve our minimization problem since it is a shape statistically meaningful due to its computational process.

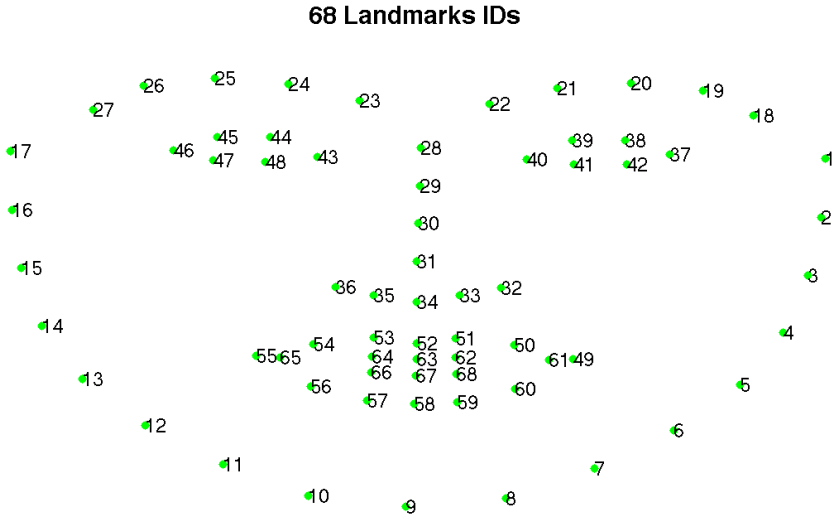

Figure 3.5: Mean shape with 68 landmarks. It includes the landmarks placed inside the facial region as well as the landmarks placed in the contour of the facial region.

Figure 3.5 illustrates the mean shape we computed. Since, unfortunately, landmarks in the contour of the face are extremely difficult to localize, in the literature it is common to use only the 49 landmarks inside the facial region. In this case, the mean shape must be modified accordingly as

illustrated in figure 3.6. Differences between figures 3.5 and 3.6 are that in the second one landmarks in the edge of the facial region are removed and, moreover, landmarks 61 and 65 are also eliminated due to their redundancy with landmarks 49 and 55.

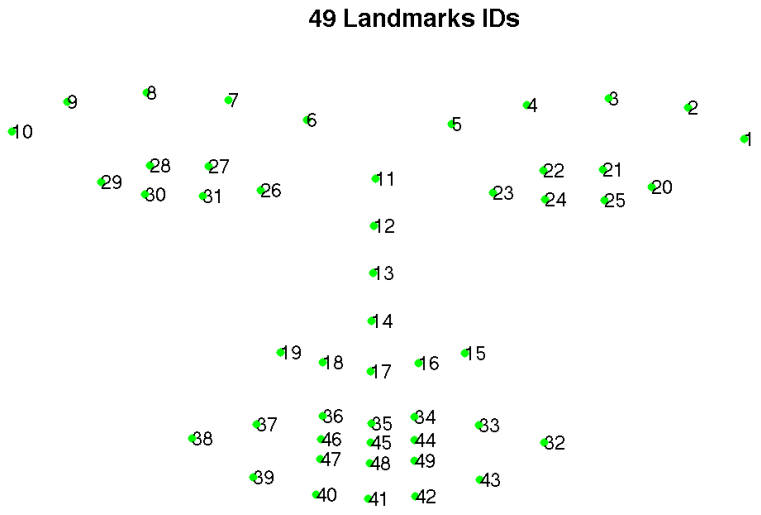

Figure 3.6: Mean shape with the 49 landmarks. This mean shape only includes the landmarks placed inside the facial region.

3.2.4 Training the Landmark Localizer Model

The parameter that affects to a great extent the accuracy of a landmark localizer software is the model used since it is in charge of determining directions where landmarks should be displaced. Due to its importance, in this section we are going to detail the procedure followed to compute the landmark localizer model.

Looking for the Ideal Training Set

A training set of images is the first element required in order to learn a model. The whole database available from ICCV2013 Challenge [1] could have been used; however, this is a challenging database in the wild with head poses difficult to detect in some cases. Therefore, we decided to build an ideal training set which contained those images such that the face computed by the face detector implemented corresponded with the face labeled in the ground truth.

In terms of the face detector, we first thought about using the one provided in the OpenCV library [30]. Nevertheless, we realized that in some cases it failed in the detection and, as a consequence, a more robust face detector was needed. To this end, a combination of the face detector from the OpenCV library and the face detector implemented in the Computer Vision Toolbox of Matlab® were used; both based on the Viola-Jones algorithm [31]. This detector was required to return a single bounding box of the face with the greatest possible area inside any image.

The ideal case would have been that the bounding box computed by our face detector perfectly matched the bounding box corresponding to the face labeled in the ground truth file; however, this situation did not occur in all images. In order to quantitatively analyze similarities between the bounding box estimated by our face detector and the bounding box computed directly from the ground truth information, a percentage of the correlation between both bounding boxes was computed. This factor can be computed, in the per-unit system, as the ratio between the area of the intersection between both bounding boxes and their total area.

$$OverlapFactor = \frac{Area_{\text{intersection}}}{Area_{\text{total}}} \tag{3.6}$$

The total area can be understood as the union of areas between two elements. According to sets theory, the union of two elements is defined as the first element plus the second one, minus the intersection between

both.

$$|A \cup B| = |A| + |B| - |A \cap B| \tag{3.7}$$

Analogously, the union of areas between the estimated bounding box and the computed one can be formalized as formulated below.

$$Area_{\text{total}} = Area_{\text{bbox computed}} + Area_{\text{bbox estimated}} - Area_{\text{intersection}} \tag{3.8}$$

In order to compute the intersection of both areas, the last unknown in equations 3.6 and 3.8, four points need to be calculated. These points are:

1. $x1$, which is the maximum x-coordinate of the top-left coordinates between both bounding boxes.
2. $y1$, which is the maximum y-coordinate of the top-left coordinates between both bounding boxes.
3. $x2$, which is the minimum x-coordinate of the bottom-right coordinates between both bounding boxes.
4. $y2$, which is the minimum y-coordinate of the bottom-right coordinates between both bounding boxes.

Finally, with these points defined, the area of the intersection can be easily computed using the following formula:

$$Area_{\text{intersection}} = (x2 - x1) * (y2 - y1) \tag{3.9}$$

This correlation factor, equation 3.6, was computed for all images in the database in order to determine the percentage of images inside ICCV2013 Challenge database such that their correlation factor was greater than a fixed threshold. The results of this experiment are shown in figure 3.7.

Results shown in figure 3.7 told us that more than 80% of images had a correlation factor of 0.1, while only around a 5% of images had a correlation factor of 0.7. These results make sense since the greater the threshold is, the less likely it is to have large amounts of images because of the remote likelihood that the bounding box computed from the ground truth was equal than the bounding box estimated from the face detector.

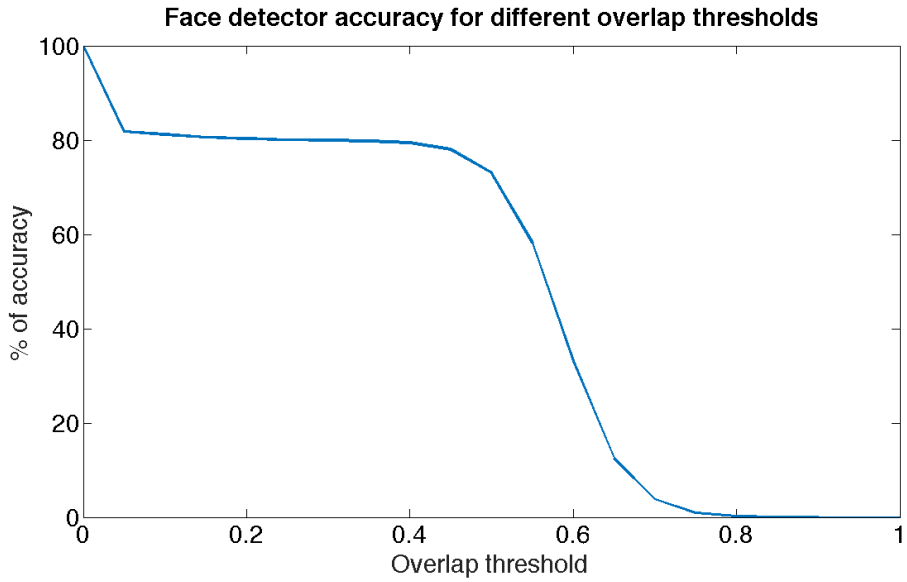

Figure 3.7: Percentage of images in the whole database such that the correlation factor between the bounding box estimated from the face detector and the bounding box computed directly from the ground truth is greater than a fixed threshold.

Analyzing results displayed in figure 3.7, the turning point of the curve can be determined, which is placed at threshold 0.5. Below this threshold, the overlapping between both bounding boxes is inadequate since it is too small, which means that the face detected is not very close to the face labeled in the ground truth, although more than 80% of the images satisfy the correlation condition. Above this threshold, despite the fact that the correlation factor is much better, which means that the estimated face is very close to the labeled one, we would have an insufficient amount of images to train our model, only around 30%. Therefore, a compromise between the correlation factor and the amount of training and testing images needs to be taken. As a result, the turning point with a correlation factor of 0.5 was chosen as the appropritate threshold since the overlap between the bounding boxes was acceptable, as it is illustrated in figure 3.8, and the amount of images available to train and test the model was

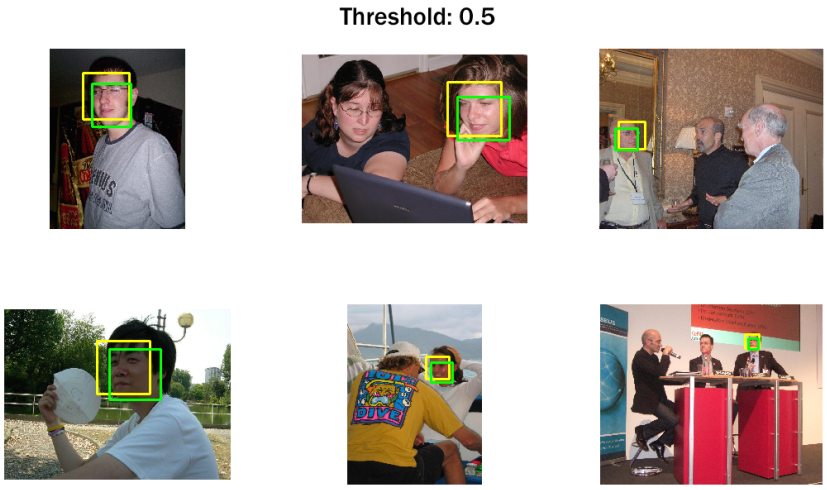

Figure 3.8: Examples of images in which the correlation factor between the bounding box estimated with the face detector and the bounding box computed directly from the ground truth is 0.5. Bounding boxes in green and yellow are the bounding box computed from the ground truth and the bounding box estimated from the face detector, respectively. Despite this figure being own developed, original images were extracted from ICCV2013 Challenge Database [1].

reasonable (75% and 25%, respectively).

Training Data Set Up

In pattern recognition, large amounts of data are required in order to train algorithms. However, training data is usually limited and, as a consequence, synthesis of new data is sometimes required. This has also been the case we had to deal with.

The model we are interested in learning from the training data, the regressor, is a matrix of size 6273×98. Despite the fact that its size will be examinated later on, the size of the regressor shows us that at least 6273 images are required in order to successfuly train the model. If less

images were used, the system of equations would be undetermined, since it would have fewer equations than unknowns and, as a result, the system would have infinite possible solutions. Preferably, the system should be overdetermined, which means that it has more equations than unknowns. In this case, the solution estimated will be the one that minimizes the error; geometrically speaking, the solution estimated will correspond with the hyperplane that minimizes the distances between the data distribution and the hyperplane itself.

In this section, the method used in order to build the training data for our regressor is detailed. For simplicity, the method used will be explained for a single image although it must be repeated for all images inside the training set.

At first, the bounding box where the face labeled in the ground truth file was placed needs to be computed. This bounding box must be estimated with the same face detector that will be implemented in the landmark localizer software, which uses the face detector provided in the Matlab® Vision Toolbox. Once determined, a security region needs to be added to the computed bounding box, as it is illustrated in figure 3.9. This augmentation is necessary for two main reasons:

1. To ensure that all landmarks fit inside the bounding box.

2. To ensure that around all landmarks there is enough information in case the neighborhood around them is useful to extract some further data.

Once the augmented bounding box was determined, it was time to synthesize new instances of the same facial region in order to increase the data available to train the regressor with the purpose to overdetermine the system. According to the available database, nine new instances per image

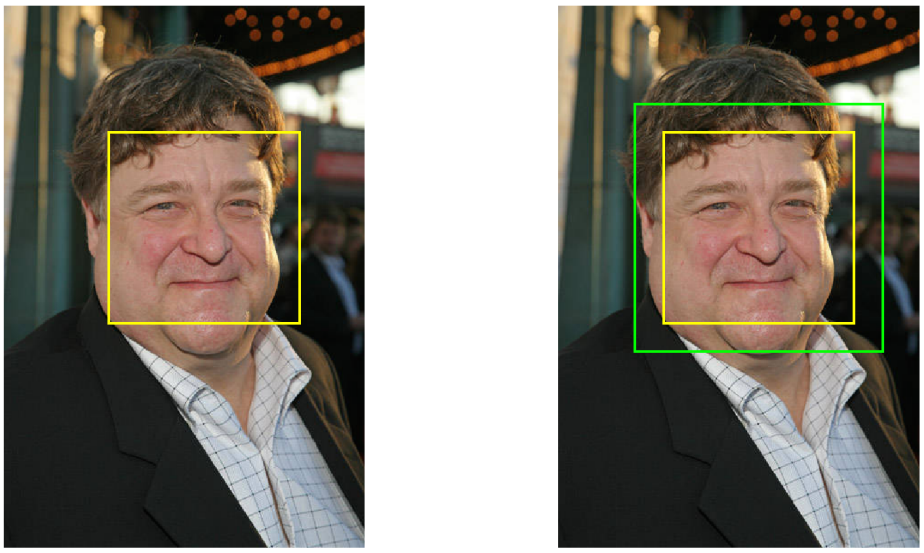

Figure 3.9: Illustration of the bounding box augmentation strategy by which a security region is added to the original one in order to ensure that all landmarks fit inside. In the left-side image, the bounding box from the face detector is plotted in yellow color. In the right-side image, the augmented bounding box is plotted in green color as well as the bounding box computed from the face detector in order to visualize the difference between them. Despite this figure being own developed, original images were extracted from ICCV2013 Challenge Database [1].

were generated. Therefore, ten instances would be obtained from a single image: one computed directly from the implemented face detector and the others synthesized, as it is illustrated in figure 3.10. The process of synthesis was done by randomly defining for each new instance the top-left coordinate of the bounding box, its width and its height. The top-left coordinate was shifted by a factor between -5% and 5% of the original image size, while its width and its height were randomly increased or decreased by a factor between -5% and 5% of the original bounding box width and height, respectively. Additionally, each new synthesized bounding box was checked in order to assure that all ground truth landmarks

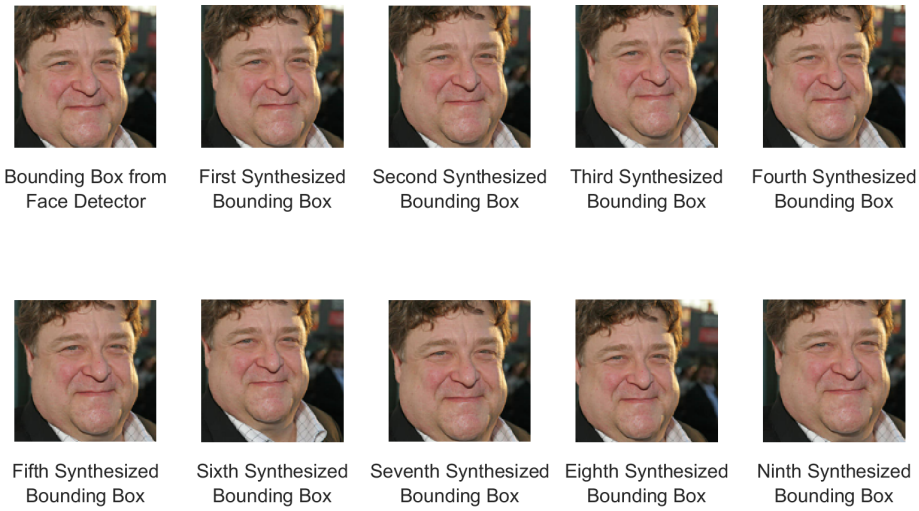

Figure 3.10: Illustration of synthetic patches generated to train the model. The first patch corresponds with the bounding box computed using the face detector, while other nine patches are extracted automatically from a random deformation, given certain restrictions, of the bounding box determined. Despite this figure being own developed, original images were extracted from ICCV2013 Challenge Database [1].

fit inside; otherwise, another random bounding box was initialized.

Finally, for each instance the following data was stored:

- The corresponding image patch resized to 128×128 pixels (since all patches to analyze must have the same size).
- Ground truth landmarks in the 128×128 pixels space. The conversion between the computed bounding box and the 128×128 bounding box could easily be done by following equation 3.4 for x-coordinates and equation 3.5 for y-coordinates.
- Landmarks initialized to the mean shape in the 128×128 pixels space. This process was done by normalizing the mean shape such that it was centered at $(0, 0)$ Cartesian coordinates and placed inside the range $[-1, 1]$. Then, its width and its height were multiplied by

half of the width and the height of the desired output bounding box, respectively, and shifted to its center.

- Additional information related with the position of the bounding box inside the original image in order to place localized landmarks at their original location.

All this data was then added to the training data pool and ready to be learnt by our regressor, which implementation is going to be detailed in the following section.

Learning the Regressor

The data pool containing all the set up information ready to learn the regressor can be defined as a set $I = \{I^1, I^2, ..., I^N\}$, where N is the total number of instances in the pool. For each instance I^i the information available is detailed below:

1. The image patch where to localize the landmarks.
2. The landmarks initialization, $landmarks_{\text{init}}$.
3. The ground truth, GT.

Variables $landmarks_{\text{init}}$ and GT are mathematically defined in equations 3.10 and 3.11, respectively, for $i = 1, 2, \ldots, N$, where p is the total number of landmarks to be used.

$$landmarks^i_{\text{init}} = \begin{bmatrix} x^i_1 \\ x^i_2 \\ \vdots \\ x^i_p \\ y^i_1 \\ y^i_2 \\ \vdots \\ y^i_p \end{bmatrix}^T \tag{3.10}$$

$$GT^i = \begin{bmatrix} x_1^{*i} \\ x_2^{*i} \\ \vdots \\ x_p^{*i} \\ y_1^{*i} \\ y_2^{*i} \\ \vdots \\ y_p^{*i} \end{bmatrix}^T \tag{3.11}$$

The mathematical definition of the regressor computation is stated in equation 3.2. However, since Matlab® is a software optimized for matricial operations, equation 3.2 ought to be modified such that it has a matrix form, in which

$$\mathbf{X}^* = \begin{bmatrix} \mathbf{x}^{*1} \\ \mathbf{x}^{*2} \\ \vdots \\ \mathbf{x}^{*N} \end{bmatrix}$$

and

$$\mathbf{X}_{k-1} = \begin{bmatrix} \mathbf{x}_{k-1}^1 \\ \mathbf{x}_{k-1}^2 \\ \vdots \\ \mathbf{x}_{k-1}^N \end{bmatrix}$$

and

$$\varphi_{k-1} = \begin{bmatrix} \phi_{k-1}^1 \\ \phi_{k-1}^2 \\ \vdots \\ \phi_{k-1}^N \end{bmatrix}$$

Therefore, it can be defined as formalized in the following equation:

$$\begin{bmatrix} \mathbf{X}^* - \mathbf{X}_{k-1} \end{bmatrix} = \begin{bmatrix} \varphi_{k-1} & \mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{R}_{k-1} \\ \mathbf{b}_{k-1} \end{bmatrix} \tag{3.12}$$

and expressed in a more compact way in the equation below, where $\mathbf{\Delta}_{k-1} = \begin{bmatrix}\mathbf{X}^* - \mathbf{X}_{k-1}\end{bmatrix}$, $\mathbf{\Phi}_{k-1} = \begin{bmatrix}\varphi_{k-1} & \mathbf{1}\end{bmatrix}$ and $\mathbf{\Gamma}_{k-1} = \begin{bmatrix}\mathbf{R}_{k-1} & \mathbf{b}_{k-1}\end{bmatrix}^T$.

$$\mathbf{\Delta}_{k-1} = \mathbf{\Phi}_{k-1}\mathbf{\Gamma}_{k-1} \tag{3.13}$$

The relevant aspect at this point is the data structure of variables $\mathbf{\Delta}$ and $\mathbf{\Phi}$. Variable $\mathbf{\Delta}$ will be a matrix of the form $\begin{bmatrix}\mathbf{\Delta}^1 \\ \mathbf{\Delta}^2 \\ \vdots \\ \mathbf{\Delta}^N\end{bmatrix}$, where the structure of each $\mathbf{\Delta}^i$ is:

$$\mathbf{\Delta}^i = \begin{bmatrix} x_1^{*i} - x_1^i \\ x_2^{*i} - x_2^i \\ \vdots \\ x_p^{*i} - x_p^i \\ y_1^{*i} - y_1^i \\ y_2^{*i} - y_2^i \\ \vdots \\ y_p^{*i} - y_p^i \end{bmatrix}^T \tag{3.14}$$

Therefore, $\mathbf{\Delta}$ will have dimensions $N \times 2P$.

Similarly, $\mathbf{\Phi}$ will have the form $\begin{bmatrix}\mathbf{\Phi}^1 \\ \mathbf{\Phi}^2 \\ \vdots \\ \mathbf{\Phi}^N\end{bmatrix}$, where each $\mathbf{\Phi}^i$ has the structure defined by:

$$\mathbf{\Phi}^i = \begin{bmatrix} f(x_1^i, y_1^i) \\ f(x_2^i, y_2^i) \\ \vdots \\ f(x_p^i, y_p^i) \\ 1 \end{bmatrix}^T \tag{3.15}$$

Let us remark that the function f in equation 3.15 represents the function to extract the SIFT descriptor at each landmark. As a result, $\mathbf{\Phi}$ will have

dimensions $N \times (128P + 1)$, since function f returns 128 coefficients per landmark.

In order to solve equation 3.13, however, Γ needs to be isolated, which corresponds to the model we are interested in learning. To this end, equation 3.13 shoud be developed using matrix algebra as deducted below:

$$\begin{aligned} \mathbf{\Delta} &= \mathbf{\Phi\Gamma} \\ \mathbf{\Phi^T\Delta} &= \mathbf{\Phi^T\Phi\Gamma} \\ (\mathbf{\Phi^T\Phi})^{-1}\mathbf{\Phi^T\Delta} &= (\mathbf{\Phi^T\Phi})^{-1}\mathbf{\Phi^T\Phi\Gamma} = \mathbf{I\Gamma} \\ \mathbf{\Phi^+\Delta} &= \mathbf{\Gamma} \end{aligned} \tag{3.16}$$

Analysing the dimensionality of equation 3.16, dimensions of model $\mathbf{\Gamma}$ are easily determined. Since $\mathbf{\Phi}^+$, which is the pseudoinverse of $\mathbf{\Phi}$, has dimensions $(128P + 1) \times N$ and $\mathbf{\Delta}$, $N \times 2P$, $\mathbf{\Gamma}$ is a matrix of dimensions $(128P + 1) \times 2P$. In case that 49 landmarks were used in order to learn the model, $\mathbf{\Gamma}$ dimensions would be 6273×98. In contrast, if 68 landmarks were used, $\mathbf{\Gamma}$ dimensions would be 8705×136. These results reveal the reason why using 49 landmarks at least 6273 images are required to learn the model, as it is the number of equation that the system has; while using 68 landmarks the minimum amount of images required rise up to 8705.

Finally, it has been previously said that the stage of learning the regressor was an iterative process. Hence, equation 3.16 needs to be repeated as many times as defined by a maximum number of iterations. Nevertheless, before recomputing the model at the new iteration, landmarks positions should be updated using equation 3.3. Moreover, SIFT descriptors vectors must be also recomputed according to new landmarks positions updated at each iteration. All things considered, equation 3.3 could be written in a more compact way such as:

$$\mathbf{X}_k = \mathbf{X}_{k-1} + \mathbf{\Phi}_{k-1}\mathbf{\Gamma}_{k-1} \tag{3.17}$$

Dimensionality reduction through Principal Component Analysis

In computer vision, it is common to deal with high dimensional problems due to the huge amount of data that needs to be processed. However, in some cases, data contains noisy elements that could affect the performance of our algorithms. As a consequence, data dimensionality reduction is a widespread practise. The main advantages provided by dimensionality reduction strategies are the following:

1. Noise removal from the data.

2. Fast computation due to the low dimensional resulting problem.

A common tool to perform dimensionality reduction is the usage of PCA [13]. This method estimates high variance directions of data, through eigenvectors, and their corresponding magnitudes, through eigenvalues. Computed eigenvectors and eigenvalues determine an orthogonal basis such that the projection error of the original data into the new basis is minimized.

Finally, in order to perform dimensionality reduction, a subspace of the orthogonal basis might be selected and the original data should be projected into this new subspace, which was set up with the most relevant eigenvectors corresponding to the greatest eigenvalues. The subspace dimension might be tunned according to the application and the data dispersion.

In our regressor problem, dimensionality reduction could be performed on the SIFT descriptors. If so, the minimum number of images required to learn the regressor may well be decreased and, accordingly, the computation time necessary to set up the training data as well as to compute the linear regression would clearly be reduced.

3.2.5 Landmark Localizer Implementation

Once the regressor is learnt, landmarks can be localized on any facial image despite being out of any database that has been used until the moment, which corresponds to the real use of a landmark localizer software. The code written to this end is the only block of code needed to be used in the proposed scheme in order to localize landmarks since all the code written for the training stage had only the purpose to compute the model.

The landmark localizer was implemented as follows. At first, faces in the images should have been detected. To this end, the face detector implemented in the Computer Vision Toolbox of Matlab® was used. Right after, all bounding boxes estimated were processed independently, since there could have been multiple faces in a single image. At first, a security region was added to each bounding box and the facial region obtained was cropped from the whole image and resized into an image of 128×128 pixels size. In this patch, landmark locations were initialized using the mean shape.

After the initialization, each layer of the regressor was applied to the corresponding landmarks position at each stage of the regressor, according to equation 3.17, until estimating the best possible landmarks location. Finally, overlap between estimated output landmarks were checked in order to reduce as many false-positive locations as possible.

The testing stage, which corresponds to the landmark localizer implementation, is the right time to analyze quantitatively, which results will be presented in section 4.1, and qualitatively the performance of the model learnt. Qualitative evaluation can be done by visualizing figure 3.11, where landmarks are localized in nine frames of nine different videos from DARHR database using the regressor computed. As can be seen from figure 3.11, despite the variability of poses among subjects, the implemented software estimates landmarks locations with a high accuracy;

Figure 3.11: Illustration of output landmarks placed on a subset of videos from DARHR database using the implemented landmark localizer with the model learnt.

which proves the generalization of the model learnt and the feasibility of the implemented localizer.

3.3 Signal Extraction

This project is based on measuring intensity color oscillations produced on the facial skin due to the bloodstream in order to estimate the frequency corresponding to the heart rate. After segmenting the facial skin region by means of the implemented landmark localizer software, information from frames needs to be processed to obtain useful signals that could describe this phenomenon. Therefore, this section is focused on the procedure followed to extract signals from videos, where the heart rate would be estimated.

3.3.1 Overview

Heart rate estimation can be done by means of signal processing. However, until this point, the available information is only represented as images. In order to extract signals from video sequences, some steps need to be done. At first, each frame needs to be processed in order to condition the information in such a way that physiological signals appear; this processing includes the selection of the most suitable color space to work with. Finally, appropriate skin pixels where to extract signals need to be defined.

Information presented in sections 3.3 and 3.4 is highly related so as to achieve the purpose of estimating the heart rate from facial videos. As a result, the workflow of both sections can be synthesized in a single diagram shown in figure 3.12.

3.3.2 Intraframe Analysis

The signal that could be extracted by selecting a pixel of an original resolution video over time is quantized due to the effect of the internal camera sensor, figure 3.13.

Because of the quantization effect, the obtained signal, which has slope shapes, is far from the smooth oscillation we are interested in measuring. Therefore, the video should be downsampled in order to smooth the extracted signal, since by downsampling the video each new pixel would be computed as an average of the neighborhood around each pixel. However, in order to avoid aliasing in the spatial domain, which might be caused by the downsampling process, the orignal video should be first filtered. In our case, a convolution between the original video and a 2D Gaussian kernel with standard deviation 2σ, where σ was the number of levels to downsample, was choosen as an antialiasing strategy.

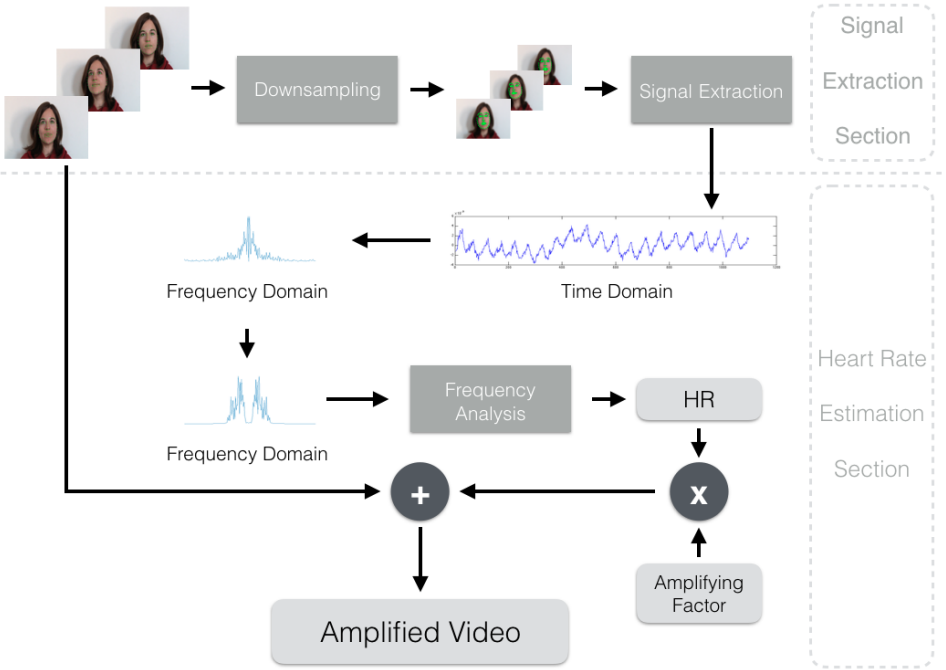

Figure 3.12: Workflow diagram of the signal extraction and the heart rate estimation processes from skin region of facial videos.

The defined convolution implies that the value of each new pixel would be a weighted average of the pixels inside a neighborhood, which size would be a multiple of a 5×5 pixels patch depending on the number of downsampling levels according to equation 3.18. As a consequence, the value of a certain pixel at a low resolution frame would depend on many different high resolution pixel values, which relationship is illustrated in figure 3.14. After filtering each frame, the downsampling can be performed just by taking one every 2^n pixels, where n defines the number of levels to downsample the original video.

$$neighborhood_{\text{size}} = 5 * 2^{\text{number of downsampling levels}-1} \tag{3.18}$$

Finally, the signal that could be extracted from a manually defined pixel in the low resolution video was completely smooth and the shape presented

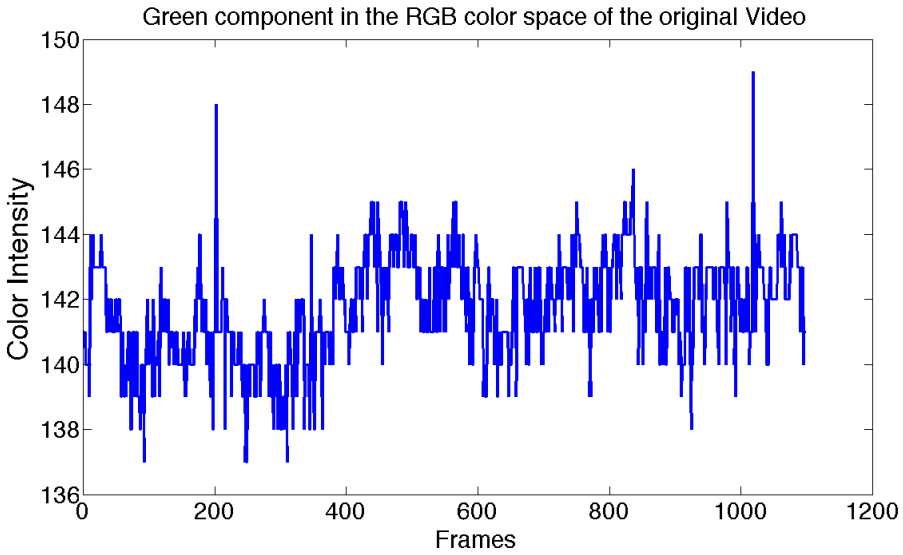

Figure 3.13: Forehead pixel intensity color evolution over time at the original resolution of the recorded video.

clearly remembered the bloodstream waveform. The described phenomenon is shown in figure 3.15, where signals extracted from the same pixel at high and low resolutions, respectively, are compared. Figure 3.15 demonstrated that while the signal in high resolution might not be easy to interpret, the signal in low resolution clearly characterizes the physiological phenomenon we are interested in measuring.

3.3.3 Color Space Analysis

Photo-plethysmography states that the green channel of a color image features the strongest plethysmography signal [8]. As a result, analyzing the green channel from an RGB video could be enough to measure the bloodstream pulse. However, red and blue channels also contain plethysmographic information. As a consequence, if information inside these channels would be taken into account, the luma component, Y, of the

Figure 3.14: Illustration of the neighborhood required at high resolution, defined with a yellow rectangle, to compute the value of a single pixel at low resolution, represented with a yellow star. The video at low resolution was obtained using four downsampling levels.

video converted into the YCbCr color space could be analyzed since it is defined by a linear combination between *R*, *G* and *B* channels, where constants K_R and K_B depend on the standard conversion used:

$$Y = K_R R + (1 - K_R - K_B)G + K_B B \tag{3.19}$$

Besides RGB and YCbCr color spaces, HSV is also a very common and useful color space. Therefore, in order to define the most suitable color space to work with in order to extract signals from which heart rate would be estimated, a comparison of signals automatically extracted from the same pixel of a DARHR Database video is shown if figure 3.16. Signals visualized in figure 3.16 were extracted from *G*, *Y*, *Cb*, *Cr*, *H*, *S* and *V* channels of RGB, YCbCr and HSV color spaces, respectively.

Analyzing figure 3.16 it can be seen that only *G* and *Y* channels clearly have a shape close to the bloodstream waveform. *V* channel has a similar behavior as *G* channel but with awful correlation. Therefore, we may conclude that *G* and *Y* channels are the most appropriate ones to work with when extracting signals related with the bloodstream.

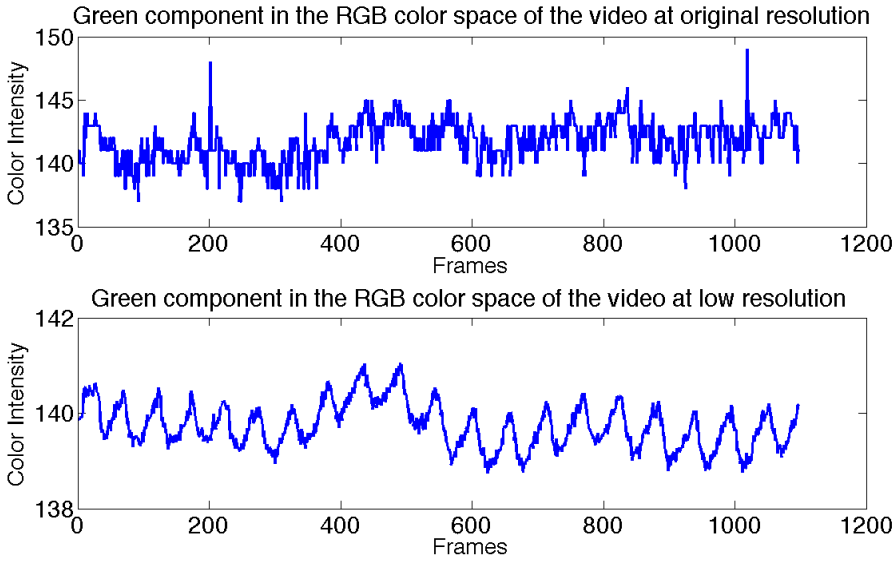

Figure 3.15: Comparison of the signals obtained from a forehead pixel all over the same video in high resolution (top) and in low resolution (bottom), which was computed by downsampling four levels the original resolution video.

In conclusion, in our scheme, input videos were converted to the YCbCr color space and signals were extracted from the luma channel since it takes into account all plethysmographic information contained in R, G and B channels.

3.3.4 Pixels to Analyze Definition

After segmenting the facial skin region, downsampling the video and selecting the most appropriate color space to process input videos, the pixels where to extract signals ought to be defined. These pixels should be invariant to any movement of the head and even rotations in order to assure the meaningfulness of the signals we are trying to measure. To solve this issue, a referential landmarks point location, as the mean shape stated in section 3.2.3, was defined and pixels inside were identified.

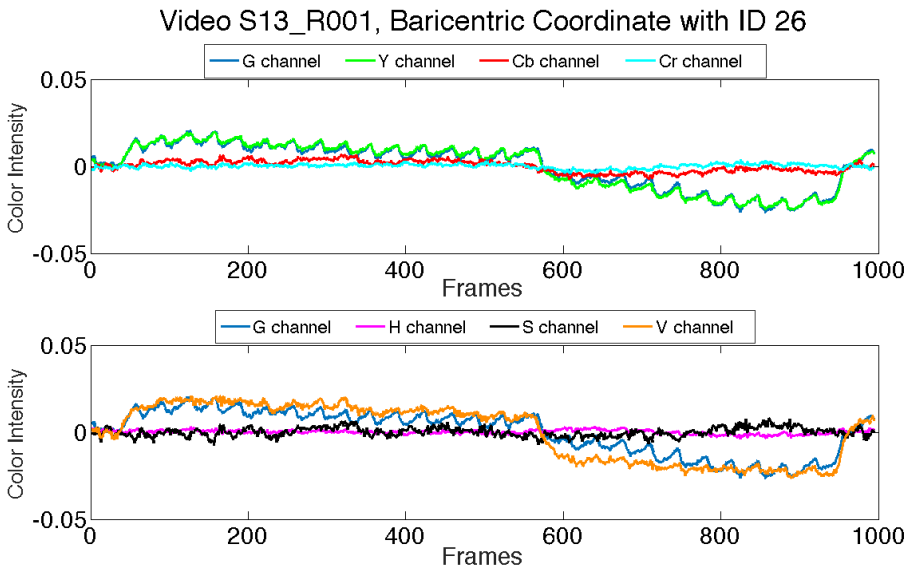

Figure 3.16: Signals comparison extracted from different channels of different color spaces at the same automatically defined pixel. The DC component of all signals was removed for a better interpretation of the obtained results.

The goal at this point was to determine correspondences between pixels inside the referential shape and pixels inside the skin region segmented using the landmark localizer software. To this end, a Delaunay triangulation on both distributions was applied and using barycentric coordinates [32], pixels inside the referential mesh could be estimated to their corresponding pixels inside the skin region.

From Delaunay triangulations obtained, we applied slight changes to it in such a way that all triangulations were symmetric with respect to the vertical axis, as it is shown in figure 3.17. Moreover, triangles corresponding to the eyes and the mouth were removed since they correspond to non-skin regions and, as a consequence, signals extracted from these regions, which may be affected differently by the bloodstream, might distort measurements to be done.

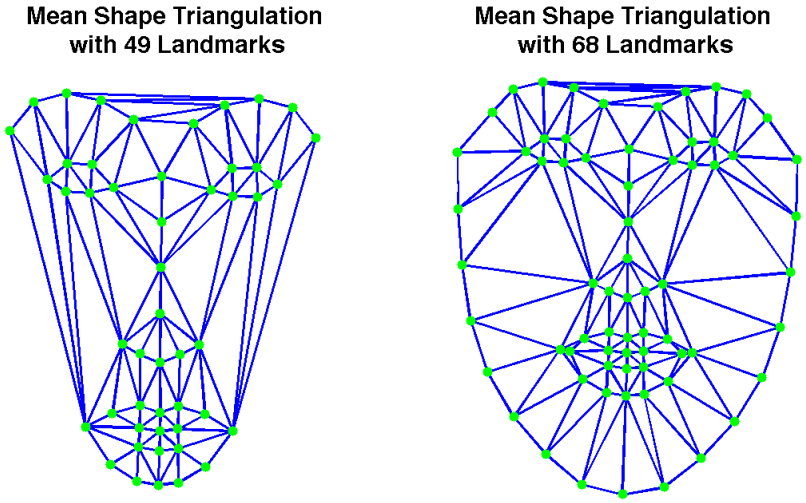

Figure 3.17: Illustration of the triangulations defined using 49 landmark points (left), and 68 landmarks points (right).

Finally, pixel coordinates of the downsampled video that correspond to pixels inside the referential shape could easily be computed through barycentric coordinates, as it is shown in figure 3.18. In addition, the intensity color value of these pixels all over the video were extracted using bilinear interpolation in order to build the signals where the heart rate will be estimated. With this strategy, facial points all over the video where signals would be extracted are assured to be invariant to movements or even 2D rotations of the subject's head.

3.4 Heart Rate Estimation

The main purpose of this project is to estimate the heart rate of a person given a facial video. Up to now, the process of data acquisition, the skin segmentation through landmarks localization and the extraction of signals from the skin region were explained. This section will be focused on the

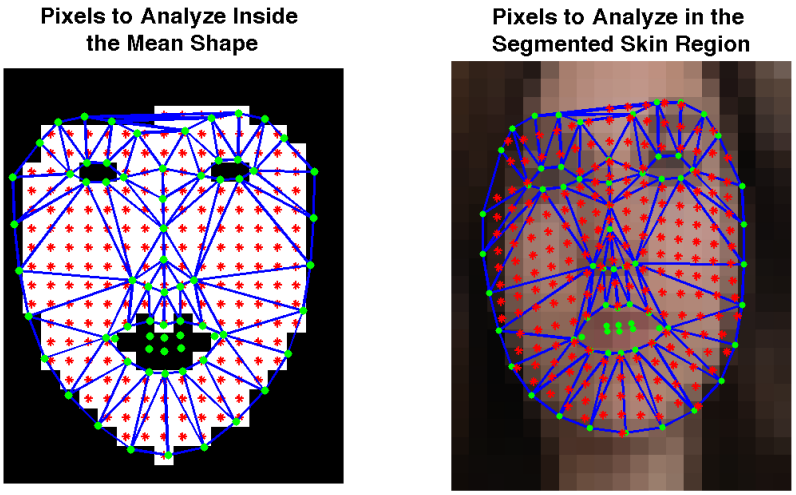

Figure 3.18: Mapping of pixels inside the mean shape points distribution to the down-sampled image with the landmarks distribution estimated by our landmark localizer.

remaining step: the estimation of this physiological signal.

Some previous work [12] proved that the temperature of the face was constant all over the facial region. As a consequence, the information provided by the skin facial region would be hypothetically homogeneous. However, we were interested on the homogeneity of the frequency information rather than the temperature information. Hence, given a facial video, the fundamental frequency of each pixell over all video was estimated. The obtained result was visualized in a map where the disparity among fundamental frequencies estimated could be analyzed, figure 3.19. To this end, fundamental frequencies were naively computed by selecting those frequencies with the greatest peak in the spectrum within a meaningful frequency band in terms of the heart rate, *i.e.* between $40\ bpm$ and $220\ bpm$, as fundamental frequencies of signals.

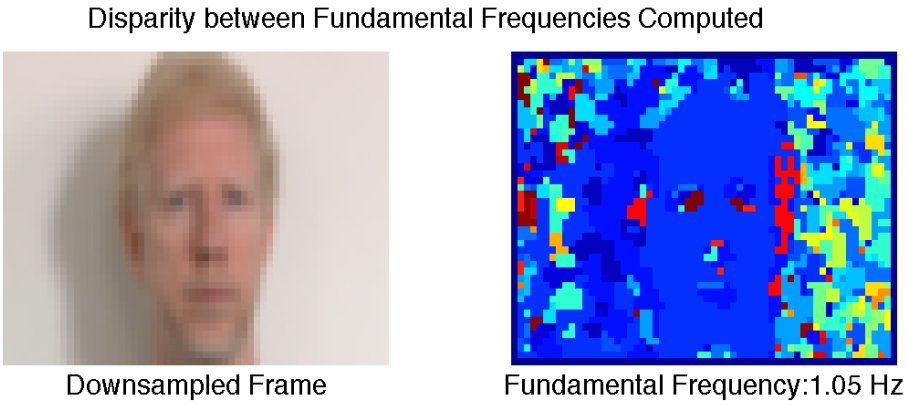

Figure 3.19: Illustration of fundamental frequencies disparity estimated at every single pixel of the facial video. Frequencies inside the facial region are homogeneous, while frequencies in the background are completely random.

Analyzing results shown in figure 3.19, two main conclusion can be extracted:

1. Homogeneity of fundamental frequencies inside the facial region seems to hold, which could be understood as if the blood pulse was constant all over the region.

2. Fundamental frequencies in the background seems to be random, contrary to what happens on the facial region.

Moreover, analyzing a facial video provided by the authors of *Eulerian video magnification for revealing subtle changes in the world* paper [15], we realized about the fact that oscillations over time extracted from multiple pixels inside the facial region were in phase, as it can be seen in figure 3.20; which also supports the hypothesis that the bloosdtream fluctuations are homogeneous all over the facial region.

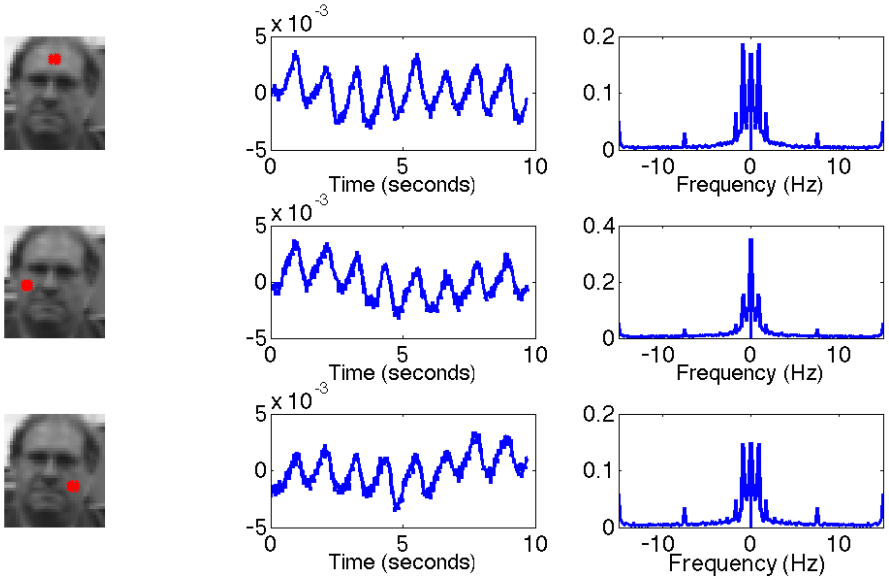

Figure 3.20: Signals comparison extracted at five different pixels in a facial video. In the left column, the neighborhoods of pixels analyzed are defined. In the central column, average signals extracted from patches over time are displayed, while in the right column, spectra of each signal are plotted.

Analyzing the spectra of these signals, some clear peaks that could correspond with the fundamental frequency and the second harmonic of a bloodstream signal can be identified, although other works reported the identification up to the fourth harmonic [8]. In order to verify this hypothesis, a synthetic spectrum only containing those peaks identified in their corresponding frequencies were created. Moreover, the second peak was rounded to the nearest integer multiple of the fundamental frequency in order to easily analyze their relationship. Finally, the time domain signal was reconstructed from this synthesized spectrum and both signals, either in the time domain or in the frequency domain, are illustrated in figure 3.21. The reconstructed signal can be computed as the summation of as many sinusoids as frequencies are in the spectrum. Since the reconstruction was done using the fundamental frequency and the second

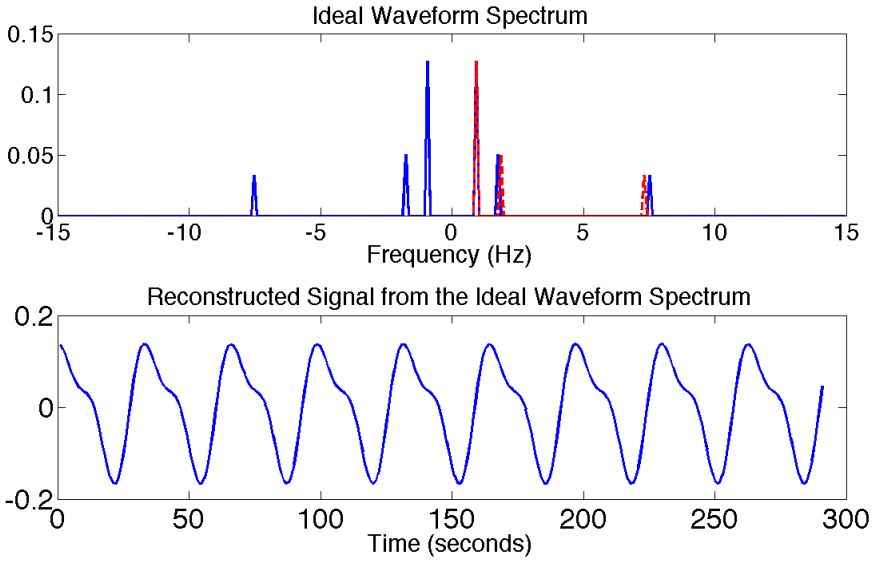

Figure 3.21: Intensity color variation signal reconstruction from an ideal spectrum, using the fundamental frequency and the second harmonic determined from the color intensity evolution of a facial pixel over time. In blue, frequencies computed from a signal extracted from a skin pixel. In red, previous frequencies rounded in order to be integer multiples of the fundamental frequency.

harmonic, the corresponding signal was defined by equation 3.20, where $A_2 = 0.3944$ and $\phi_2 = 2.1254\ rad$. These values were computed maintaining the magnitude ratio between the fundamental frequency and the second harmonic, assuming that the fundamental frequency had amplitude one, and also maintining the phase difference between the fundamental frequency and the second harmonic, assuming that, in this case, the phase of the fundamental frequency was $0\ rad$.

$$S = cos(2\pi f_0 t) + A_2 cos(2\pi 2 f_0 t + \phi_2) \tag{3.20}$$

Analyzing the signal obtained in figure 3.21, the feasibility of defining the bloodstream pulse with only two frequencies is verified. In addition, the feasibility of computing the fundamental frequency associated with the heart rate from a facial video is also proven.

3.4.1 Overview

Heart rate is an unsteady signal which changes over time due to many different factors such as the breathing or the oxygen consumption of our body. As a consequence, signals extracted from a set of pixels inside the skin region of a facial video cannot be processed as a whole since heart rate variability would be missed. Therefore, windowing signals and processing them with certain overlap would be preferable as heart rate would be estimated continuously at fixed time intervals. This would be the most appropriate scheme to implement since it would take into account the heart rate variability over time.

After windowing, signals should be converted to the frequency domain using the Fourier Transform. Since the frequency band where heart rate frequencies can be found is perfectly known, signals could be filtered. By filtering, side information would be removed and the relevant information to contribute towards the estimation of the heart rate would be reinforced. Finally, the most likely frequency related to the heart rate could be estimated and color amplification could be applied to the original input video. Figure 3.12 illustrates the workflow explained in a diagram form, where the signal extraction section is also displayed.

3.4.2 Signal Windowing

In order to estimate the heart rate, which is measured in beats per minute (bpm), with certain periodicity, signals should be processed at defined time intervals. These signals might be noisy; therefore, it is preferable to process them in such a way that only the information we are interested in remains intact. Selecting a portion of signals without any processing is equivalent to window signals by a rectangular window. However, rectangular windows have many drawbacks since their Fourier transform corresponds to a sync function, which has undesired properties due to its low-pass filter behavior and the cancellation of those frequencies corres-

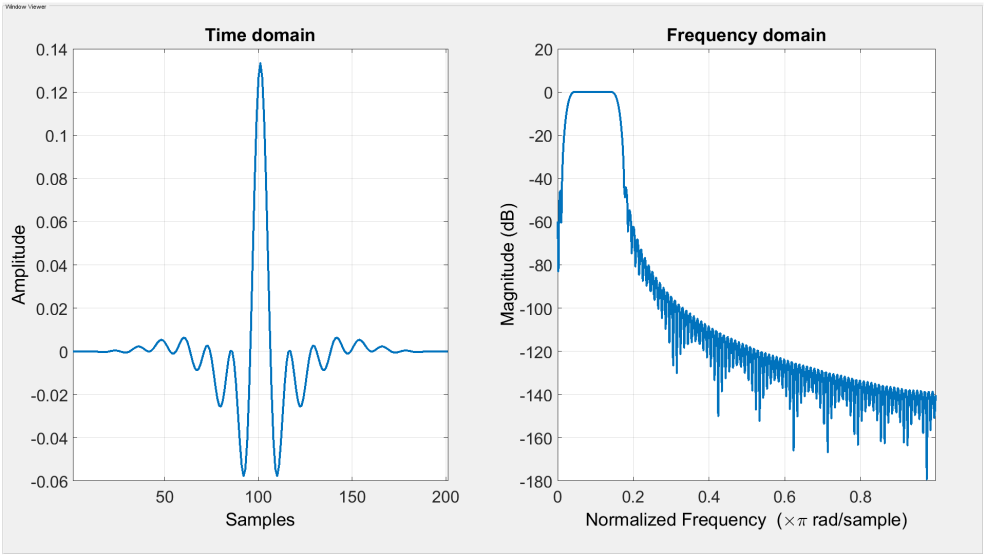

Figure 3.22: Visualization of the window-based filter designed using a Hanning window and cut-off frequencies such that the main lobe of the filter in the frequency domain is centered in the range between $40\ bpm$ and $240\ bpm$. The visualization on the left corresponds to the filter in the time domain, while on the right, in the frequency domain. This filter was designed using 201 coefficients.

ponding to its zeros [33]. Therefore, a more appropriate window selection relies on the usage of a Hanning window [34] with its main lobe centered at frequencies between $[40\ bpm, 240\ bpm]$, which are the heart rate range of ground truth data from DARHR Database. The window-based filter designed and implemented can be seen in figure 3.22.

At this point, windowed signals in the time domain should be filtered by the designed filter. After filtering, signals were ready to be converted to the frequency domain in order to perform the frequency analysis required to estimate the heart rate.

3.4.3 Frequency Analysis

After windowing signals and converting them into the frequency domain using the Fast Fourier Transform, the spectrum was obtained and their frequency components were analyzed. The spectrum contains information related with the magnitude and the phase of a signal. However, we were only interested in the magnitude since it is far more informative than the phase.

At this point of our scheme, our goal was to estimate the fundamental frequency of each one of the spectra related to the signals extracted from the facial skin region. At first, all peaks of the spectrum were identified. Then, due to the symmetry of the spectrum, only the peaks belonging to the positive frequencies in the spectrum were selected. Among these frequencies there should be a fundamental frequency corresponding to the heart rate.

The heart rate should correspond to the peak with the greatest magnitude in the spectrum. Nevertheless, due to the weakness of signals extracted, some of the computed spectra might be considered as noise and useless in terms of heart rate estimation.

These useless spectra were identified through the following procedure: the peak with the greatest magnitude inside the frequency range between $[40\ bpm, 240\ bpm]$ was considered as the fundamental frequency. Right after, the Signal to Noise Ratio (SNR) was computed between the magnitude of the fundamental frequency and the summation of all magnitudes inside the same range with the exception of the magnitude associated with the fundamental frequency. If the SNR was lower than a certain threshold, $-10\ dB$,the spectrum was considered noisy and the fundamental frequency would be discarded; otherwise, if the SNR was greater, the fundamental frequency was considered as the heart rate frequency associated with the corresponding signal.

3.4.4 Heart Rate Estimation through Voting Scheme

After analyzing the spectra computed from each pixel inside the mesh related to the facial skin region, an array of fundamental frequencies was obtained. In the ideal case, all frequencies would be exactly the same due to the frequency information homogeneity inside the skin facial region. However, this array usually contained multiple frequencies due to artifacts associated with landmark localization errors or the acquisition of movements produced by non-rigid elements such as the eyebrows, among others.

Despite this variability, a single fundamental frequency was extracted by consesus. The consesus applied in the proposed scheme was based on a voting scheme. At first, the range of possible fundamental frequencies was divided into 180 bins. Subsequently, the number of fundamental frequencies estimated that fit inside each bin were counted. Finally, the frequency associated with the most voted bin was defined as the heart rate of the time interval analyzed.

3.4.5 Color Amplification

After the frequency corresponding to the heart rate was estimated, our aim was to process the original input video in such a way that the color oscillations caused by the blood stream could be seen by the naked eye.

To this end, the estimated frequency, or a patch around it, inside spectra computed from facial skin region signals were selected by filtering. Then, these spectra were transformed back to the time domain and only their real parts were amplified by a factor of seventy, obtaining a sequence of images related to the desired frequencial information of the original video. These resulting images were added to the original ones in order to obtain the output video of the proposed scheme.

Chapter 4

RESULTS AND EVALUATION

In this chapter the most revelant experimental results obtained during the implementation of our scheme are presented.

Firstly, the performance of the landmark localizer software implemented is evaluated. Multiple models were learnt and tested to determine the model with the greatest generalization capability. Such model would be the most appropriate one to be used in the implemented software.

Secondly, the time consumption of the proposed scheme is analyzed since one of the goals of this project was the heart rate measurement from facial videos with an execution time as close as possible to the real time.

Thirdly, the nature of both the signals extracted from the segmented skin region and the heart rate signals are studied in order to understand the type of signals to process and to estimate, respectively.

Finally, the accuracy of the system is tested by means of comparing the heart rate estimated by the proposed scheme with the ground truth data measured with invasive and especialized equipment.

4.1 Landmark Localizer Evaluation

The evaluation of a landmark localizer software can be done by comparing the estimated locations of the landmarks with their real positions. This information is not available for all images; therefore, labeled data should be used to this end.

To perform the evaluation of the landmark localizer software implemented, labeled images from both the Ideal Training Set and the ICCV2013 Challenge Database were used. This evaluation was based on computing the error between the labeled positions of the landmarks and the estimated ones. This error was computed as the average of the Euclidean distances between the estimated landmarks and the ground truth, where P is the total number of landmarks to be localized:

$$d = \frac{1}{P} \sum_{i=1}^{P} \sqrt{\left(x_{\text{landmarks}_i} - x_{\text{GT}_i}\right)^2 + \left(y_{\text{landmarks}_i} - y_{\text{GT}_i}\right)^2} \tag{4.1}$$

The error determined from equation 4.1, however, may well not offer a fair comparison of the estimation accuracy due to the high dependency of this measurement on the image size; *i.e.*, bad localizations would have a greater error on high resolution facial images rather than on low resolution ones. Hence, in order to downplay the effect of this variability, the normalization of the error computed with equation 4.1 by the interocular distance is commonly used in the literature.

The interocular distance is the Euclidean distance between the central pixel of the right eye and the central pixel of the left eye, which can be defined as C_{right} and C_{left}, respectively. Therefore, the interocular distance can be mathematically defined as:

$$intDistance = \sqrt{\left(C_{\text{left}_x} - C_{\text{right}_x}\right)^2 + \left(C_{\text{left}_y} - C_{\text{right}_y}\right)^2} \tag{4.2}$$

All things considered, the error normalized by the interocular distance ξ can be computed as formalized in the following equation:

$$\xi = \frac{d}{intDistance} \tag{4.3}$$

Finally, before presenting the results obtained, we would like to highlight that all errors computed in this section were normalized by the interocular distance.

4.1.1 Implementation Verifications

In this section, results obtained when evaluating the implementation of the landmark localizer software are presented. To this end, three different subsets from the Ideal Training Set were created:

- Dataset A: contained 2000 images of the Ideal Training Set with their corresponding initializations.
- Dataset B: contained the same 2000 images as Dataset A but, in this case, data was newly initialized.
- Dataset C: contained the remaining images of the Ideal Training Set that were not used neither in Dataset A nor Dataset B.

After splitting the data, two models were learnt from Dataset A: the first one for 49 landmarks localization and the second one for 68. In order to verify the implemented software, both models were tested in all three datasets.

Although testing a model with the same data used to train could be thought as useless, results obtained in this case had a special interest for two main reasons:

1. To assure the correct implementation of the algorithm since the error between the landmarks estimated during the training stage and their corresponding ground truth should decay exponentially over iterations.

2. To determine the number of iterations required for the algorithm to converge.

Testing learnt models with Dataset B and Dataset C was a suitable procedure to analyze how these models would behave towards unknown sets of data, in an attempt to analyze their generalization capability.

Results obtained, expressed as an average of the errors between estimated landmarks and their corresponding ground truth and normalized by the interocular distance, are presented below.

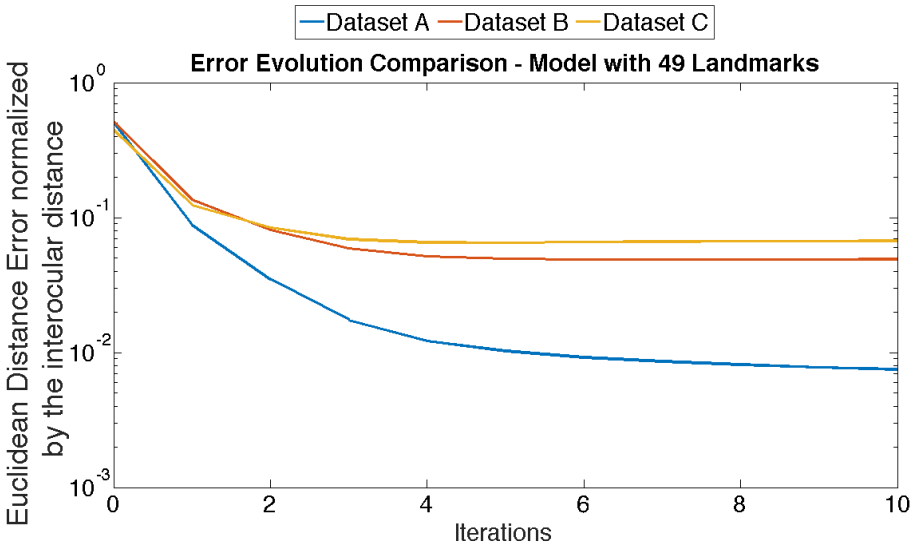

Figure 4.1: Error evolution of three different datasets computed by testing a 49 landmarks model trained with 2000 images of the Ideal Training Set. Errors had been normalized by the interocular distance.

Model with 49 Landmarks

The error evolution over iterations for the three datasets previously defined is illustrated in figure 4.1. The model used in this case was trained

in order to localize 49 landmarks inside the facial region and, as a result, only these landmarks were localized. For a better analysis of the learnt model, numerical results from figure 4.1 are extracted and synthesized in table 4.1.

	Initialization	It. 1	It. 2	It. 3	It. 4	It. 5	It. 6	It. 7	It. 8	It. 9	It. 10
Dataset A	0.5034	0.0882	0.0349	0.0174	0.0122	0.0102	0.0092	0.0086	0.0081	0.0078	0.0075
Dataset B	0.5116	0.1356	0.0808	0.0589	0.0516	0.0495	0.0488	0.0487	0.0488	0.0489	0.0490
Dataset C	0.4437	0.1241	0.0842	0.0690	0.0653	0.0651	0.0656	0.0662	0.0666	0.0669	0.0670

Table 4.1: Average of the interocular error evolution over iterations for the different datasets expressed on a logarithmic scale. Results were obtained from a 49 landmarks model learnt by using 2000 images of the Ideal Training Set.

Model with 68 Landmarks

The error evolution over iterations of the model trained with 68 landmarks and tested with the three previously defined datasets is illustrated in figure 4.2. In this case, since 68 landmarks were used to train the model, the software was able to localize all 68 landmarks: including those landmarks placed inside the facial region and those localized on the edge of the face. As done before, numerical results from figure 4.2 are extracted and displayed in table 4.2.

	Initialization	It. 1	It. 2	It. 3	It. 4	It. 5	It. 6	It. 7	It. 8	It. 9	It. 10
Dataset A	0.4861	0.0729	0.0245	0.0127	0.0093	0.0077	0.0067	0.0061	0.0056	0.0053	0.0050
Dataset B	0.4888	0.1257	0.0778	0.0617	0.0577	0.0568	0.0569	0.0572	0.0575	0.0577	0.0579
Dataset C	0.6083	0.3074	0.2727	0.2619	0.2607	0.2614	0.2622	0.2631	0.2635	0.2637	0.2639

Table 4.2: Average of the interocular error evolution over iterations for the different datasets expressed on a logarithmic scale. Results were obtained from a 68 landmarks model learnt by using 2000 images of the Ideal Training Set.

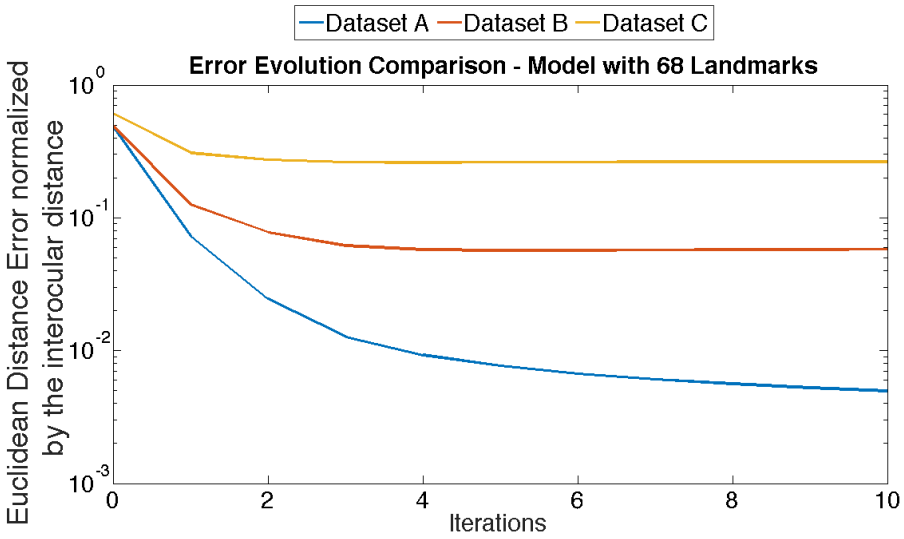

Figure 4.2: Error evolution of three different datasets computed by testing a 68 landmarks model trained with 2000 images of the Ideal Training Set. Errors had been normalized by the interocular distance.

Conclusions

According to results obtained by testing the three defined datasets with models learnt using 49 and 68 landmarks respectively, several conclusions can be drawn.

1. Analyzing Dataset A, the error using both models decays exponentially, as expected. Moreover, the last error computed is very close to zero, which validates the learning process since at each iteration a closer solution to the ground truth is obtained.

2. Analyzing both Dataset B and C, the error evolution also decays exponentially over iterations, which certifies the correct generalization of the models learnt.

3. Taking a closer look at the numercial results obtained from both Dataset B and C, the average error tends to increase after the fifth

iteration, which is caused by the overfitting of the model. This overfitting appears when the model is unable to improve the estimated landmarks location and starts shifting them away from the ground truth. This overfitting appears in both models, as it can be seen in table 4.1 as well as in table 4.2, respectively. Since the overfitting has undesired effects, which might well reduce the accuracy of the software, the maximum number of iterations for the estimated regressor to converge can be fixed at five iterations. Five iterations of the regressor would be enough in order to obtain accurate results since it is the iteration at which the average error normalized by the interocular distance is minimized.

4. Analyzing errors computed when analyzing Dataset C in both models, it can be seen that the average error is greater with the 68 landmarks model than with the 49 landmarks model. This fact proves our hypothesis by which the localization of 68 landmarks is by far a more challenging task than just localizing 49 landmarks.

4.1.2 Testing on the Ideal Training Set

A high generalization capability might well be the most important purpose when learning a model, which means the ability to obtain accurate results on unknown data. A single splitting of the data as done in the previous section, however, is not suitable for analyzing the generalization capabilities of the computed models. As a consequence, an appropriate statistical analysis to this end is the k-fold cross-validation. This model validation technique consists of splitting the available data into k sets $S_1, S_2, S_3, \ldots, S_k$. Then, $k-1$ sets are used in order to train the model and the remaining set is used to test it. This process was repeated k times so as to test all the data available. With this statistical analysis, the performance of models learnt to unknow data is better analyzed.

Hence, 4-fold cross-validation was performed on the Ideal Training Set in order to statistically verify the convergence of the iterative process de-

ducted in the previous section. Moreover, this technique was applied twice: the first time with a model learnt from 49 landmarks, and the second one with a model learnt from 68 landmarks; which results are going to be contrasted.

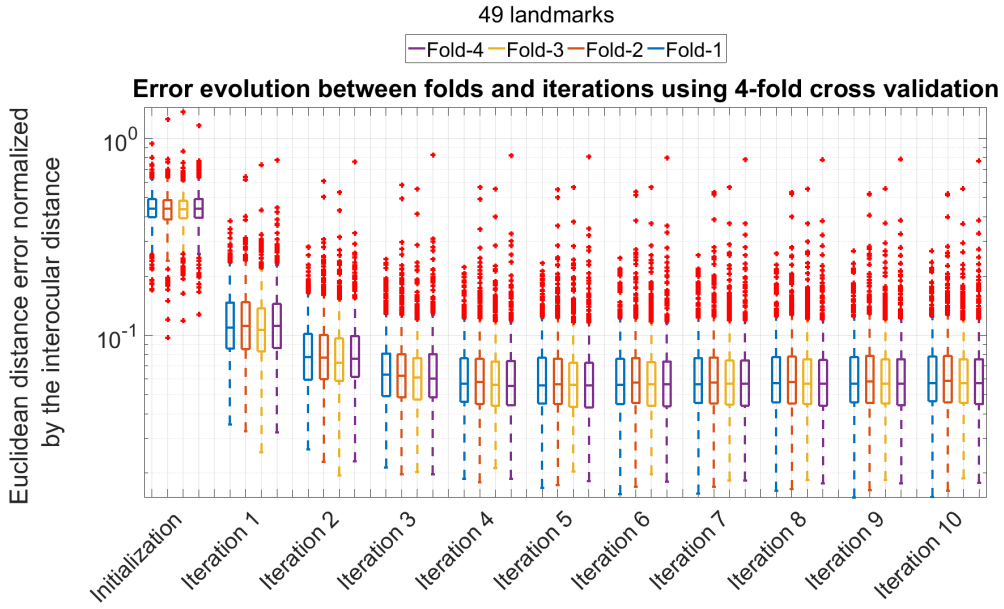

Figure 4.3: Error distribution over folds and iterations of the Ideal Training Set when performing a 4-fold cross-validation of the models learnt from 49 landmarks. Errors are displayed using a boxplot representation, which allows the visualization of data distributions through quartiles. Lines below, in the middle and at the top of the boxes are related with the first quartile, the median and the third quartile, respectively. The dashed lines represent the data that fit inside 1.5 times the interquartile range beyond the third quartile and below the first quartile, while red crosses are related to outliers.

49 Landmarks Model

Obtained results when performing a 4-fold cross-validation on the Ideal Training Set with 49 landmark models can be seen in figure 4.3. 4-fold cross-validation implies that four different models were learnt: each one

of them was learnt from three different folds and tested on the remaining one.

For a better analysis of figure 4.3, numerical results are extracted and displayed in table 4.3 where the mean and the standard deviation (SD) for each fold and iteration are detailed.

		Fold-1	Fold-2	Fold-3	Fold-4
Initialization	Mean	0.4472	0.4403	0.4399	0.4464
	SD	0.0865	0.0923	0.0889	0.0901
Iteration 1	Mean	0.1218	0.1230	0.1169	0.1231
	SD	0.0511	0.0578	0.0527	0.0589
Iteration 2	Mean	0.0861	0.0862	0.0831	0.0867
	SD	0.0360	0.0453	0.0417	0.0454
Iteration 3	Mean	0.0699	0.0714	0.0693	0.0696
	SD	0.0309	0.0424	0.0389	0.0437
Iteration 4	Mean	0.0654	0.0679	0.0650	0.0647
	SD	0.0304	0.0427	0.0383	0.0434
Iteration 5	Mean	0.0647	0.0675	0.0651	0.0644
	SD	0.0310	0.0434	0.0389	0.0438
Iteration 6	Mean	0.0652	0.0681	0.0657	0.0647
	SD	0.0317	0.0441	0.0392	0.0439
Iteration 7	Mean	0.0657	0.0685	0.0664	0.0655
	SD	0.0320	0.0447	0.0394	0.0437
Iteration 8	Mean	0.0661	0.0689	0.0668	0.0658
	SD	0.0322	0.0450	0.0396	0.0438
Iteration 9	Mean	0.0664	0.0694	0.0671	0.0662
	SD	0.0326	0.0450	0.0396	0.0441
Iteration 10	Mean	0.0666	0.0695	0.0673	0.0664
	SD	0.0326	0.0449	0.0398	0.0437

Table 4.3: Mean and standard deviation over folds and iterations of the Ideal Training Set when performing a 4-fold cross-validation of models learnt from 49 landmarks.

68 Landmarks Model

Results obtained when performing 4-fold cross-validation on the Ideal Training Set but in this case using a model learnt from 68 landmarks are illustrated in figure 4.4. As before, four different models were learnt and tested with their corresponding folds.

For a better analysis of figure 4.4, numerical results are extracted and displayed in table 4.4, where the mean and the standard deviation (SD) for each fold and iteration are detailed.

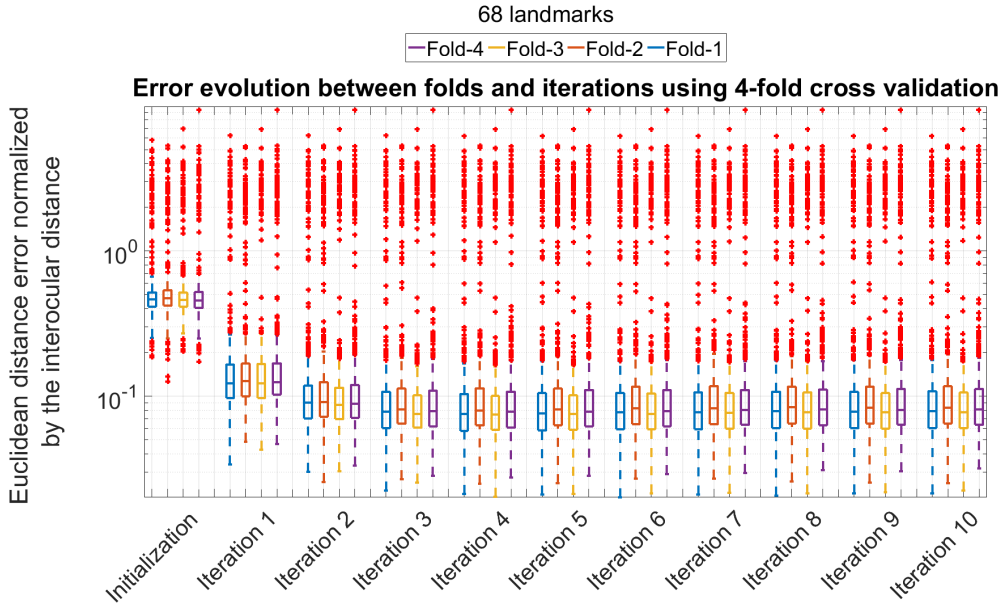

Figure 4.4: Error distribution over folds and iterations of the Ideal Training Set when performing a 4-fold cross-validation of the models learnt from 68 landmarks. Errors are displayed using a boxplot representation, which allows the visualization of data distributions through quartiles. Lines below, in the middle and at the top of the boxes are related with the first quartile, the median and the third quartile, respectively. The dashed lines represent the data that fit inside 1.5 times the interquartile range beyond the third quartile and below the first quartile, while red crosses are related to outliers.

Conclusions

Analyzing results obtained by the evaluation of both models, a pair of observations can be formalized. The first one can be observed in both models: after the fifth iteration the mean error at each fold starts rising instead of carrying on decreasing; which means that at the sixth iteration the model starts overfitting. Therefore, five stages of the iterative process are enough in order to obtain the best possible landmarks location and, as a result, only five regressors need to be learnt.

The second observation is that results obtained are homogeneous throughout the four folds. This implies that all models properly generalize since

		Fold-1	Fold-2	Fold-3	Fold-4
Initialization	Mean	0.6124	0.6200	0.6146	0.6248
	SD	0.6333	0.5943	0.6559	0.7294
Iteration 1	Mean	0.3058	0.3137	0.3056	0.3272
	SD	0.7175	0.6823	0.7339	0.8015
Iteration 2	Mean	0.2734	0.2814	0.2721	0.2932
	SD	0.7243	0.6896	0.7417	0.8085
Iteration 3	Mean	0.2634	0.2722	0.2615	0.2839
	SD	0.7269	0.6924	0.7436	0.8099
Iteration 4	Mean	0.2618	0.2709	0.2601	0.2831
	SD	0.7269	0.6922	0.7437	0.8098
Iteration 5	Mean	0.2626	0.2720	0.2606	0.2839
	SD	0.7268	0.6921	0.7437	0.8096
Iteration 6	Mean	0.2633	0.2729	0.2614	0.2848
	SD	0.7266	0.6919	0.7434	0.8093
Iteration 7	Mean	0.2640	0.2735	0.2621	0.2856
	SD	0.7265	0.6918	0.7430	0.8092
Iteration 8	Mean	0.2645	0.2742	0.2625	0.2859
	SD	0.7265	0.6916	0.7428	0.8092
Iteration 9	Mean	0.2647	0.2743	0.2629	0.2862
	SD	0.7264	0.6916	0.7426	0.8092
Iteration 10	Mean	0.2649	0.2745	0.2632	0.2863
	SD	0.7263	0.6915	0.7427	0.8090

Table 4.4: Mean and standard deviation over folds and iterations of the Ideal Training Set when performing a 4-fold cross-validation of the models learnt from 68 landmarks.

there is no fold such that the localization of the landmarks was extremely difficult in comparison with others.

4.1.3 Testing on the ICCV2013 Challenge Database

In pattern recognition, the performance of learnt models depends on the amount of data used to train it. Therefore, the usage of the greatest amount of available images is recommended when training models. In our case, we used the ICCV2013 Challenge Database and, as a result, nine different models were computed from this dataset. We performed cross-validation tests to measure the generalization ability of the models. We performed different training and test splittings to consider also the possibility that the learnt models would overfit to acquisition conditions of the training images; *i.e.*, overfit to the database.

The nine different models compared are described below. Their differences rely on three factors: the usage of 49 or 68 landmarks to train the

model, the usage of data augmentation in order to fullfil the condition to overdetermine the system, and the usage of PCA over SIFT descriptors in order to reduce the dimensionality of the problem.

- **M1**
  Model provided by Supervised Descent Method's authors [25]. Results obtained from this model are used as a baseline in order to analyze and compare the performance of the other models learnt.

- **M2**
  Model computed using one database leave out strategy. This means that every model was learnt from three datasets and tested on the remaining one. This model was computed using only the 49 landmarks inside the facial region, with data augmentation and without using PCA.

- **M3**
  Models computed using 4-fold cross-validation strategy in the Ideal Training Set. Therefore, each image in each database was tested with the model such that the affected image was excluded from the training folds. Images out of the Ideal Training Set were tested with one of the four models selected randomly. In this model 49 landmarks inside the facial region were learnt with data augmentation and without using PCA.

- **M4**
  Model computed using all images inside the ICCV2013 Challenge Dataset. Despite the fact that tested images were used to train the model, we were interested in computing it in order to discard the usage of the whole database by Supervised Descent Method's authors in a way to explain the good results that they obtained. In this model 49 landmarks inside the facial region were learnt with data augmentaion and without using PCA.

- **M5**
  Model computed using one database leave out strategy as M2. Howe-

ver, in this case, models of 68 landmarks were learnt using data augmentation and withouth using PCA.

- **M6**
  Model computed using 4-fold cross-validation strategy as M3. In this case, however, models of 68 landmarks were learnt with data augmentation and without using PCA.

- **M7**
  Models computed using all images inside the ICCV2013 Challenge Dataset as M4. However, in this case, models of 68 landmarks were learnt using data augmentation and without using PCA.

- **M8**
  Model computed using all images available from the ICCV2013 Challenge Dataset. The particularity is that in this case a model of 68 landmarks was learnt without using data augmentation and using PCA in order to reduce the dimensionality of the problem.

- **M9**
  Model computed using all images available from the ICCV2013 Challenge Dataset. In this model, despite the fact that a model of 68 landmarks was learnt, data augmentation was used as well as PCA in order to reduce the dimensionality of the problem.

Results obtained by each model are illustrated in figure 4.5. In this figure, errors distribution for each model and dataset are illustrated using boxplots. Moreover, in order to numerically detail the results, the mean and the standard deviation (SD) are computed from each dataset and each model and displayed in table 4.5.

Analyzing results presented in figure 4.5 and table 4.5, several conclusions can be extracted.

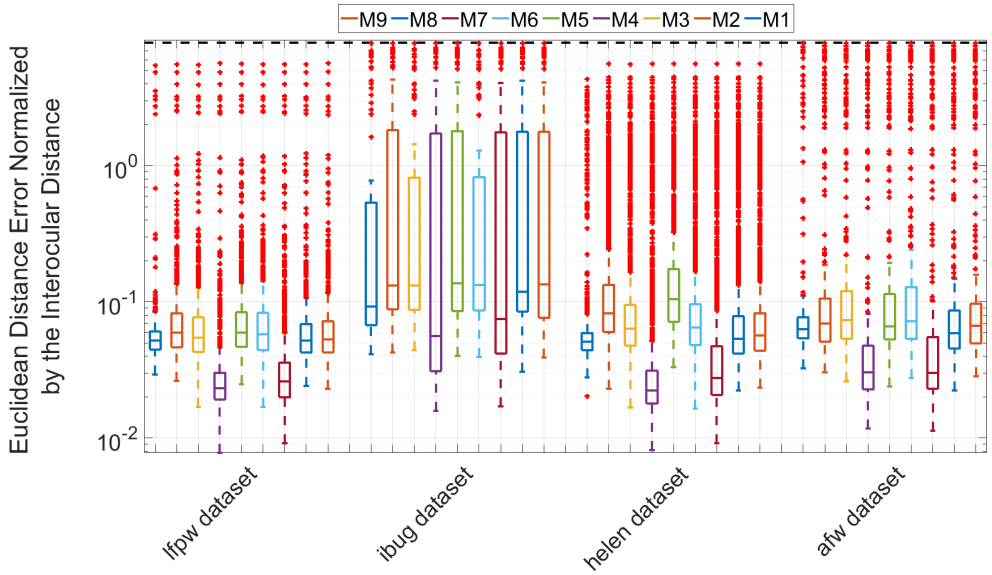

Figure 4.5: Error comparison of nine different models tested on the whole ICCV2013 Challenge Database. Errors are displayed using a boxplot representation, which allows the visualization of data distributions through quartiles. Lines below, in the middle and at the top of the boxes are related with the first quartile, the median and the third quartile, respectively. The dashed lines represent the data that fit inside 1.5 times the interquartile range beyond the third quartile and below the first quartile, while red crosses are related to outliers.

- *Ibug* is the most challenging dataset where to localize landmarks due to the high median error obtained using whichever model. This result can be explained by the fact that *ibug* dataset is a very difficult and tricky dataset due to head rotations and uncommon head poses of the subjects, which makes more difficult the face localization task.

- Results obtained from models M4 and M7 are biased since tested images were also used, identically, during the training stage of both models. Therefore, despite the good results presented, neither of them can be considered as the best model with the greatest generalization capability.

		M1	M2	M3	M4	M5	M6	M7	M8	M9
lpfw database	Mean	0.0808	0.1076	0.1024	0.0630	0.1106	0.1066	0.0707	0.0985	0.0983
	SD	0.3122	0.3350	0.3351	0.3363	0.3377	0.3363	0.3369	0.3376	0.3360
ibug database	Mean	2.0025	1.5083	1.8442	1.4399	1.5104	1.8499	1.4518	1.5099	1.4924
	SD	4.3713	2.7374	4.0917	2.7551	2.7299	4.0908	2.7424	2.7413	2.7232
helen database	Mean	0.1008	0.1938	0.1652	0.1259	0.2175	0.1683	0.1354	0.1573	0.1602
	SD	0.3338	0.4202	0.4203	0.4282	0.4189	0.4204	0.4280	0.4242	0.4231
afw database	Mean	0.9704	0.7096	0.7230	0.6607	0.7106	0.7253	0.6655	0.7060	0.7042
	SD	3.0538	2.0717	2.0875	2.0696	2.0671	2.0855	2.0675	2.0785	2.0654

Table 4.5: Numerical comparison of nine different models tested on the whole ICCV2013 Challenge Database.

- From the evaluation of models M2, M3, M5 and M6, the appropriateness of learning information throughout the whole available datasets can be highlighted since information provided in one dataset might be different from others and the generalization capability of the model might be affected. This fact can be explained by the reason that images from each subset inside the ICCV2013 Challenge Dataset are homogeneous in terms of head poses and rotations.

- Finally, M8 could be selected as the best model since it obtained the lowest median in all datasets. Model M8 was computed without using data aumentation and using PCA, while model M9 was computed using data augmentation and PCA. Hence, this result may prove the uselessness of data aumentation through data synthesis in order to obtain better results. This situation can be explained by the fact that synthetic data intializations are never computed by the algorithm itself and, as a result, directions learnt in those cases would have a noisy and distorting effect on the overall model behavior.

To sum up this section, we can conclude that the model with the best generalization capability is the model M8 and, as result, it is going to be the model used in the landmark localizer sofware implemented. Before closing this section it should be highlighted that ICCV2013 Challenge Dataset was captured in the wild, which means that images were taken without any constrain in neither the background nor the foreground. As a result, since DARHR Database was recorded in a controlled environment

with an easy background, landmark localization with the choosen model is expected to be reasonably accurate.

4.2 System Computational Time Analysis

In this section, the processing time of the implemented scheme to process a one minute-length video is analyzed. Nevertheless, the video may not be processed as a whole due to heart rate variations over time. Thus, in these cases, the strategy followed consisted of estimating a heart rate every certain amount of frames with a certain overlap in order to assure the continuity of the measurement.

At first, the number of frames to analyze simultaneously have to be defined. This is an important parameter since the frequency resolution of the spectrum mainly depends on the length of the input signal; *i.e.*, the number of frames analyzed at the same time. The frequency resolution is inversely proportional to the number of frames per window to be analyzed. Therefore, large fragments would have a finer frequency resolution. However, the greater the amount of frames per window are, the lower the resolution in the time domain is. Hence, a compromise needed to be taken. In our case, a window length of 501 frames was defined since it corresponded with ten video seconds, as used in other works [8], and, since videos were recorded at 50 frames per second, a frequency resolution of $6\ bpm$ per bin would be achieved with this window length, which is a reasonable resolution to obtain reliable results.

The second parameter that still needs to be defined is the overlap factor, which defines the periodicity of heart rate estimations. Two reasonable overlap factors would be 75% and 100%, since using lower overlapping factors the heart rate variability would be extremely difficult to measure. However, this choice would affect the processing time required for the system since the greatest the overlapping factor is, the more frequent heart rate estimations would be and, as a consequence, more computati-

ons would be done. Therefore, we decided to perform a time processing comparison between both overlapping factors in order to have quantitative results when analyzing the best overlapping factor to be used in our scheme.

Additionally, processing time was split into the different stages of the proposed scheme in order to evaluate the computational time of each stage.

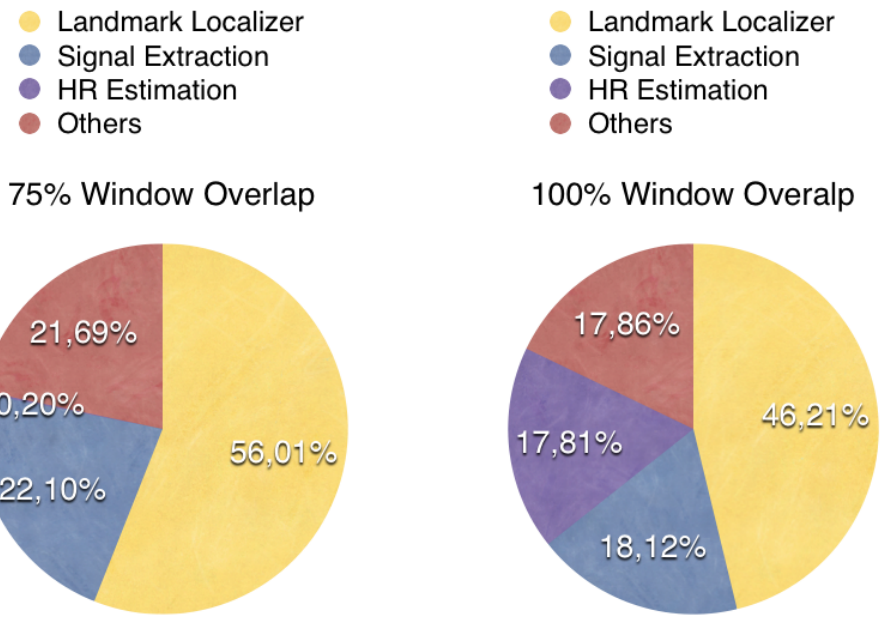

Figure 4.6: Percentual comparison of the processing time required for each stage of the proposed scheme according to the overlapping factor used, 75% (left) or 100% (right).

4.2.1 Overlap Factor of 75%

Firstly, the processing time with an overlapping factor of 75% was analyzed. Time consumption for each stage of the proposed scheme is detailed

in table 4.6, and graphically illustrated in the left side of figure 4.6. Time measurements were expressed in seconds and as a percentage of the total processing time.

	Time (Seconds)	Time (Percentage)
Landmark Localizer	353.818 s	56.01 %
Signal Extraction	139.572 s	22.10 %
HR Estimation	1.285 s	0.20 %
Others	136.980 s	21.69 %
Total Time	631.655 s	100 %

Table 4.6: Time consumption analysis evaluated at each stage of the proposed scheme using an overlapping factor of 75% between consecutive windows.

The selected video to perform this analysis contained 3020 frames and lasted a total time of 631.66 seconds. Therefore, we can conclude that, using an overlap factor of 75%, the system was able to process nearly five frames per second.

Moreover, more than half of the processing time was dedicated to the landmark localizer stage, while an insignificant percentage of time was dedicated to the heart rate estimation stage. This fact can be explained since, in this case, one heart rate estimation was done every hundred and twenty-six frames, nearly two and a half seconds.

Finally, another relevant result is that more than 75% of the processing time was dedicated to the main stages of the scheme; while less than 25% was dedicated to the load of pre-learnt models and data verifications, among other tasks.

4.2.2 Overlap Factor of 100%

Secondly, the processing time required for processing a video of one minute length with an overlapping factor of 100% was analyzed. Time consumptions for each stage of the proposed scheme are detailed in table 4.7

and graphically illustrated in the right side of figure 4.6. In this case, time measurements were also expressed in seconds and as a percentage of the total processing time.

	Time (Seconds)	Time (Percentage)
Landmark Localizer	350.788 s	46.21 %
Signal Extraction	137.566 s	18.12 %
HR Estimation	135.178 s	17.81 %
Others	135.610 s	17.86 %
Total Time	759.142 s	100 %

Table 4.7: Time consumption analysis evaluated at each stage of the proposed scheme using an overlapping factor of 100% between consecutive windows.

As the video analyzed was the same as for the previous evaluation, it contained 3020 frames; however, in this case the total processing time was 759.14 seconds. Therefore, we can conclude that, using an overlapping factor of 100%, the system was able to process nearly four frames per second.

Despite the landmark localizer stage being the most time-consuming stage of the scheme, the processing time consumed in this case for the heart rate estimation stage needed also to be highlighted. This fact was reasonable since using an overlapping factor of 100%, one heart rate estimation was performed at every frame, nearly one over fifty seconds.

Finally, although the increment of the time consumption of the heart rate estimation stage, the processing time required for the other stages were proportional as those obtained when evaluating the time consumption with an overlapping factor of 75%.

4.2.3 Conclusions

Analyzing the presented results, several conclusions can be extracted, which are listed below.

- The processing time required to perform *Landmark Localization* and *Signal Extraction* stages were proportionally the same regardless of the overlapping factor used. This was expected since their computations were not dependent on the overlapping factor defined.

- A vast majority of the processing time was dedicated to required stages to estimate the heart rate from facial videos rather than being dedicated to loading or checking data, among others.

- Computations done for the 100% overlapping factor took more time than using the 75% overlapping factor, since heart rate estimations in the first case were done nearly one hundred and fifty times more than in the second case.

- The processing time using 75% overlapping factor was only one frame per second faster than using 100% overlapping factor.

To sum up, the total processing time in both cases differs only for two minutes, approximately. Therefore, in order to get as many heart rate estimations as possible to reflect heart rate variability and due to the small difference in the processing time required, an overlapping factor of 100% was decided to be used in the proposed scheme.

4.3 Heart Rate Information Analysis

In this section, the most relevant aspects related with heart rate information are analyzed. Firstly, the frequency content of one of the extracted signals from one video belonging to the DARHR Database is analyzed. Secondly, the nature of the ground truth signals recorded in the DARHR Database are evaluated.

4.3.1 Frequency Analysis through Spectrogram

At this point, we were interested in taking a first look at the frequency information provided by signals extracted from the skin facial region. To

this end, one of the signals extracted from a video belonging to DARHR Database was selected, figure 4.7, and its corresponding spectrogram was computed, which can be seen in figure 4.8.

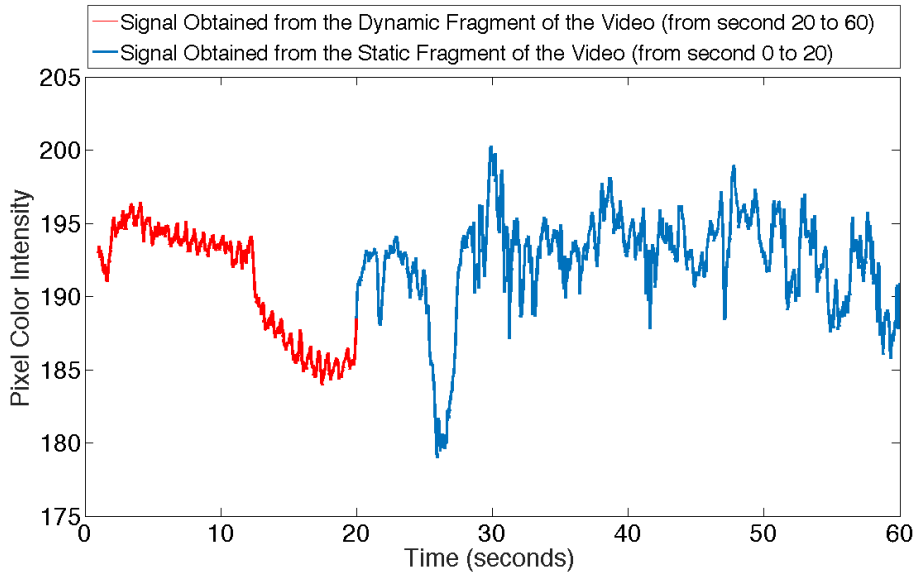

Figure 4.7: Signal automatically extracted from one barycentric coordinate inside the facial skin region representing the intensity color variations over time. The red section of the signal corresponds with the static fragment of the video, while the blue section of the signal corresponds with the dynamic fragment.

Analyzing the signal illustrated in figure 4.7, a clear difference in terms of color intensities can be seen in the static fragment of the video with respect to the dynamic one. Therefore, it would be reasonable to have a similar difference on the spectrogram.

Taking a look at the first twenty seconds of the spectrogram, figure 4.8, three main frequency bands can be easily identified. The lowest frequency band can provide information related with the respiration rate, since the

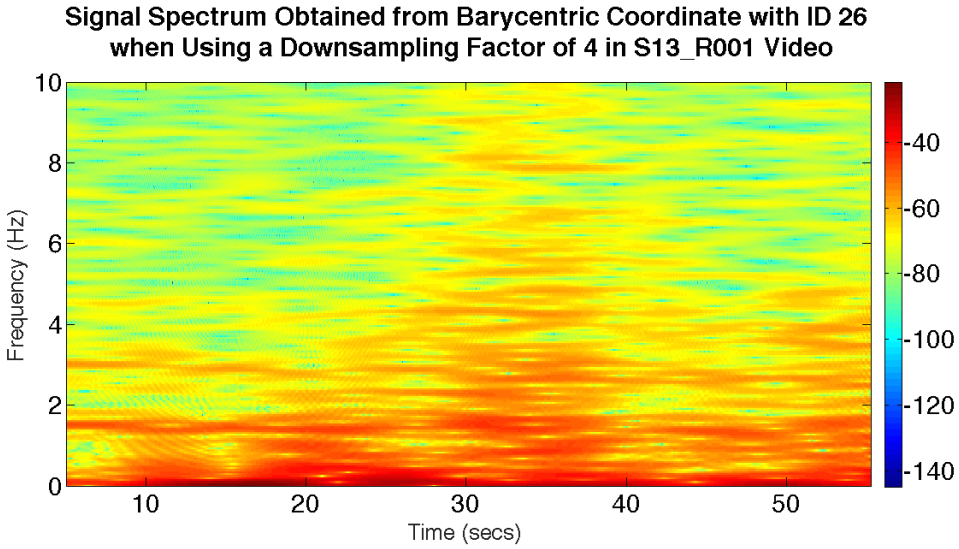

Figure 4.8: Spectrogram of a signal extracted from one barycentric coordinate inside the facial skin region of a video belonging to DARHR Database.

normal range of this physiological signal goes from twelve breaths per minute, $0.2\ Hz$, to eighteen breaths per minute, $0.3\ Hz$, on adults [35]. Analyzing the spectrogram, there were other frequencies closer to zero that had a high power. However, signals of frequencies under $0.1\ Hz$ were usually not considered to provide plethysmographic information due to the complex calibration of the devices used to measure them [7]. The second band of frequencies with the greatest power, around $1.5\ Hz$, corresponded to the frequencies related with the heart rate due to their similarity with the ground truth frequencies of the same video. Lastly, the third band of frequencies, around $3\ Hz$, which had the lowest power of the three analyzed bands, might correspond to the second harmonic of the fundamental frequency corresponding to the heart rate.

After twenty seconds, the frequency bands with the highest power were not as well defined as in the first twenty seconds of the video. Therefore, the estimation of the fundamental frequency corresponding to the heart rate would be much more tricky in this period of the video.

4.3.2 Ground Truth Data Verification

Analyzing ground truth data from DARHR Database, we realized that there was a high variability among heart rates measured inside the same video. Despite the fact that heart rate was an unsteady signal, the variability measured in DARHR Database seemed excessive. Therefore, in order to contrast this fact, ground truth heart rate from MAHNOB-HCI Database [16] was also computed. This database provided ECG recordings; thus, corresponding heart rates can be computed as the inverse of the period between two consecutive R-peaks. A comparison of ground truth heart rates obtained from both databases is illustrated in figure 4.9.

As it can be seen from figure 4.9, ground truth heart rate from MAHNOB-HCI Database oscillated within an interval of $10\ bpm$, while ground truth heart rate oscillations from DARHR Database fluctuated nearly $20\ bpm$. This fact suggests that ground truth data from DARHR Database might not be as accurate as it should due to the strong ground truth variability. This variability might be produced by some interferences inducted to the signal because of slightly electrode displacements caused by subject's movements during the recordings.

To sum up, the accuracy of the ground truth data recorded in DARHR Database is questionable. It is our hypothesis that the reason for this could be attributed to noise in the measurements. Thus, the evaluation of results in this project will be focused on the average heart rate over long time periods (*e.g.* 20 or 60 seconds), preferably using robust statistics such as the median error.

4.4 Scheme Accuracy Evaluation

In this section, the accuracy of the scheme proposed in this project is evaluated. This evaluation would be done by comparing the heart rate

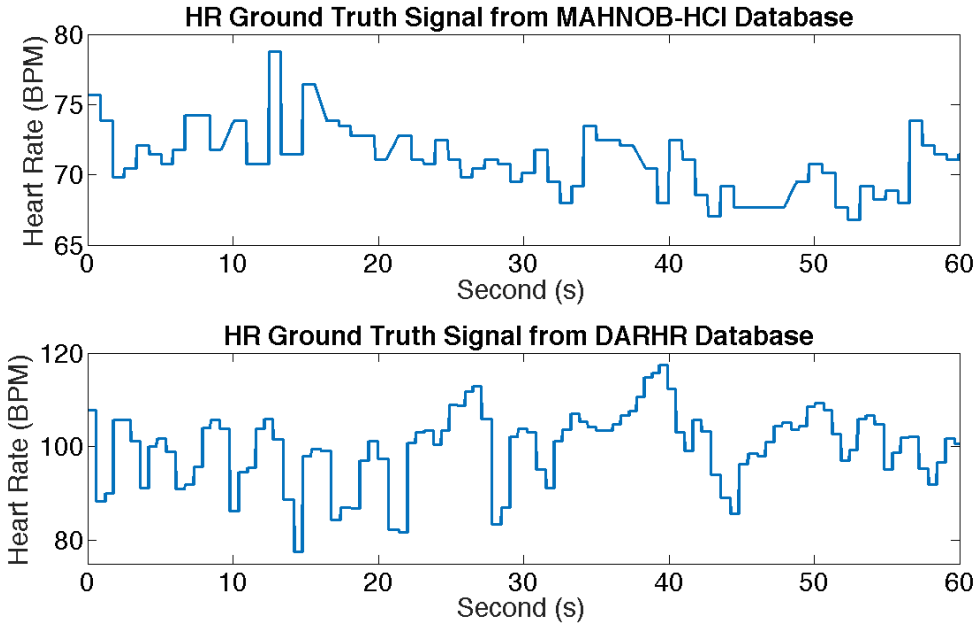

Figure 4.9: Nature signals comparison of ground truth heart rates from MAHNOB-HCI Database (top), and DARHR Database (bottom).

estimated with the proposed scheme with the ground truth heart rate measured using conventional equipment. As stated previously, the evaluations would be done in three steps: firstly, analyzing the static fragments of the videos; secondly, analyzing the dynamic fragments of the videos; and thirdly, analyzing the videos as a whole.

Analyzing the static fragments of the videos we got a global median error in the heart rate estimation of $2.64\ bpm$. However, the global median error obtained for dynamic fragments was $30.63\ bpm$, while analyzing the overall videos, the global median error obtained was $26.02\ bpm$. As errors obtained in the last two cases were greater than $6\ bpm$, the maximum heart rate error acceptable, since it corresponded with the frequency resolution that we had, detailed results for these cases will not be reported due to their clear inaccuracy.

In order to correctly analyze the results, the nomenclature used should be explained: code numbers with the prefix *S* express the identification code of the subjects recorded in the DARHR Database, code numbers with the prefix *R* express the identification code of the recordings. *S* codes are from one to thirteen, one for each subject, while *R* codes are from one to six, one for each subject's recordings.

4.4.1 Static Videos Evaluation

Static videos were considered to be the first twenty seconds recorded from DARHR Database facial videos since, during this period of time, subjects were asked to remain as quiet as possible. The median of heart rate values estimated by the proposed scheme during this period of time for the whole DARHR Database as well as the median of the corresponding ground truths recorded during the same period of time are shown in table 4.8.

	R001		R002		R003		R004		R005		R006	
	HR	GT	HR	GT	HR	GT	HR	GT	HR	GT	HR	GT
S01	71.86	97.88	89.82	88.11	83.83	84.27	77.84	80.97	71.86	82.99	71.86	83.92
S02	65.87	89.15	71.86	72.55	62.87	81.97	77.84	74.91	77.84	78.12	77.84	74.63
S03	47.90	120.24	59.88	107.14	95.81	96.31	89.82	89.02	89.82	91.19	83.83	84.75
S04	59.88	61.48	65.87	67.42	71.86	71.69	65.87	68.18	65.87	66.15	65.87	64.90
S05	59.88	104.53	71.86	85.96	65.87	78.33	77.84	78.02	62.87	77.62	77.84	75.28
S06	89.82	90.91	65.87	86.71	65.87	83.92	65.87	89.42	71.86	84.27	71.86	82.76
S07	62.87	96.66	77.84	80.21	77.84	80.00	83.83	82.08	77.84	79.16	77.84	77.12
S08	47.90	64.52	65.87	66.01	59.88	73.44	71.86	68.26	65.87	63.69	59.88	57.25
S09	65.87	64.79	71.86	71.86	71.86	72.99	65.87	67.42	65.87	69.28	71.86	69.12
S10	65.87	84.27	77.84	83.57	71.86	74.35	77.84	80.86	71.86	73.62	77.84	82.19
S11	53.89	50.63	65.87	71.26	77.84	78.02	77.84	79.58	71.86	78.84	59.88	66.37
S12	101.80	104.53	80.84	103.99	95.81	96.00	65.87	95.39	89.82	89.69	95.81	93.17
S13	83.83	89.69	65.87	69.77	71.86	71.26	71.86	71.34	71.86	71.01	71.86	72.99

Table 4.8: Heart rate evaluation during the first twenty seconds of DARHR Database videos. For each subject and each measurement, the median of the estimated heart rates and the median of the ground truth heart rates are compared. All measures are expressed in beats per minute (bpm).

In order to consider a heart rate estimation correct, we will accept that it was $\pm 6\ bpm$, which corresponds with the frequency resolution for windows of length 501 frames when using a frame rate of $50\ fps$. The abso-

lute value of the errors obtained are displayed in table 4.9.

Analyzing heart rate errors presented in table 4.9, it can be computed that 69.23% of the measurements fullfil the condition to differ only for $\pm 6\ bpm$. In addition, it can be seen that the most innacurate recordings were the first ones: those videos recorded after the subject's exercise. A similar situation happened with the second recordings since despite their median being better, it was quite near to the threshold. Therefore, we can conclude that heart rates estimated on static fragments of videos are reasonable good despite the fact that high heart rates are problematic.

	R001	R002	R003	R004	R005	R006
S01	26.02	1.71	0.44	3.13	11.13	12.06
S02	23.28	0.69	19.10	2.93	0.28	3.21
S03	72.34	47.26	0.50	0.80	1.37	0.92
S04	1.60	1.55	0.18	2.31	0.28	0.97
S05	44.65	14.10	12.46	0.18	14.75	2.56
S06	1.09	20.84	18.05	23.55	12.41	10.90
S07	33.79	2.37	2.16	1.75	1.32	0.72
S08	16.62	0.14	13.56	3.60	2.18	2.63
S09	1.08	0.00	1.13	1.55	3.41	2.74
S10	18.40	5.73	2.49	3.02	1.76	4.35
S11	3.26	5.39	0.18	1.74	6.98	6.49
S12	2.73	23.15	0.19	29.52	0.13	2.64
S13	5.86	3.90	0.60	0.52	0.85	1.13
Median	16.62	3.90	1.13	2.31	1.76	2.64

Table 4.9: Errors in measured heart rates compared with ground truth for static fragments of DARHR Database videos. All measures are expressed in beats per minutes (bpm).

4.4.2 Conclusions

As a final conclusion, it has been demonstrated that heart rate estimations from dynamic videos were completely inaccurate. The most likely reason to explain this result might be an inappropriate definition of the pixels where to extract signals in dynamic fragments.

The strategy followed in the proposed scheme consisted on the selection of a set of pixels inside the segmented skin region, which works for

static videos as it has been proved. However, during dynamic fragments of the videos, subjects rotated their heads freely; which means that rotations were produced not only in 2D but also in 3D. Therefore, it is not difficult to think that there are disagreements between the 2D barycentric coordinates estimated and the appropriate pixels when 3D rotations were performed by the subject.

This mismatch implies a wrong determination of the pixel to select and, as a result, a wrong extraction of the intensity value to build the signal to analyze. Hence, signals constructed in these cases might easily be noisy and useless in terms of heart rate estimation.

Chapter 5

CONCLUSIONS

In this last section, an evaluation of the overall project will be performed. Firstly, the compliance of the objectives defined at the beginning of this project will be assessed. Secondly, conclusions extracted from the development of this project are formalized and, finally, some further work and improvements are suggested.

5.1 Achievements

The main objective of this project, the measurement of physiological signals such as the heart rate using non-invasive techniques as facial videos, is fulfilled achieving a median error of $2.64\ bpm$ on static fragments of seventy-eight facial videos from thirteen different subjects.

Reviewing objectives stated in section 1.2, we can conclude that around 90% of them are successfully achieved. The only objective failed was the achievement of a scheme performance close to the real time since we were only able to get a processing time around four frames per second. However, this processing time was measured from a Matlab® prototypical version of the proposed scheme; therefore, if it had been implemented in a faster language as C++, this processing time might have been considerably improved.

5.2 Project Development Conclusions

From the development of this project, several conclusions were extracted; which are formalized in the list below.

- We realized that the ground truth data from DARHR Database was inaccurate in some cases. These inaccuracies may be caused by interferences inducted to measurements, which might be produced by an excessive shifting of the electrodes with subjects' movements.

- We identified the necessity to look for other strategies to define facial pixels, such that they are invariant to 2D and 3D rotations. In this manner, we would ensure the analysis of the same facial pixel over time despite subjects' rotations.

- We reinforced the relevancy of scale-space theory in order to find the desired information from the unknown at the appropriate scale. The signal recorded directly at high resolution did not contain any plethysmographic information. However, when downsampling original frames to the suitable low resolution, equivalent to look at the appropriate scale in the scale-space, plethysmographic signals appeared.

- We also highlighted the importance of working with the most suitable color space according to the desired application. In our case, for instance, working with the hue value of the HSV color space would have not provided any kind of plethysmographic information since the signal obtained had not got any periodicity.

- We made the hypothesis that information related with the respiration rate appeared in signals extracted from facial skin pixels over time since we got high peaks at frequencies in the spectrum that coincided with their theoretical results. Nevertheless, this plethysmographic signal was not deeply studied in this project.

- We finally concluded with the high difficulties we faced in order to estimate the heart rate from dynamic videos due to the noisy signals that were extracted with our approach.

5.3 Future Work and Improvements

This project would be carried on according to the following guidelines:

- Usage of 3D facial landmarks in order to ensure that the defined pixels where to extract signals are the same all over the video. In our case, rotations and head movements deliberately penalize the location of the defined pixels. This might be one of the reasons to understand why the proposed scheme fails when analyzing dynamic videos.

- Definition of a wider *pass-plethysmography* information window before computing the spectrum of extracted signals. In this way, frequency information around the heart rate as well as the respiration rate frequency bands of interest will remain intact. In this case, two physiological signals, heart rate and respiration rate, would be measured just by using a facial video.

- Scheme evaluation using the MAHNOB-HCI Database, in order to analyze the performance of the proposed scheme using proper ground truth data in the 100 % of the cases and allowing a more direct comparison to reported results from other researchers.

Bibliography

[1] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 397–403, 2013.

[2] Merriam-Webster, *Merriam-Webster's collegiate dictionary*. Merriam-Webster, 2004.

[3] M. Garbey, N. Sun, A. Merla, and I. Pavlidis, "Contact-free measurement of cardiac pulse based on the analysis of thermal imagery," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 8, pp. 1418–1426, 2007.

[4] S. S. Ulyanov and V. V. Tuchin, "Pulse-wave monitoring by means of focused laser beams scattered by skin surface and membranes," in *OE/LASE'93: Optics, Electro-Optics, & Laser Applications in Science& Engineering*, pp. 160–167, International Society for Optics and Photonics, 1993.

[5] D. Holdsworth, C. Norley, R. Frayne, D. Steinman, and B. Rutt, "Characterization of common carotid artery blood-flow waveforms in normal human subjects," *Physiological measurement*, vol. 20, no. 3, p. 219, 1999.

[6] K. Aminian, X. Thouvenin, P. Robert, J. Seydoux, and L. Girardier, "A piezoelectric belt for cardiac pulse and respiration measurements

on small mammals," in *Engineering in Medicine and Biology Society, 1992 14th Annual International Conference of the IEEE*, vol. 6, pp. 2663–2664, IEEE, 1992.

[7] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological measurement*, vol. 28, no. 3, p. R1, 2007.

[8] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," *Optics express*, vol. 16, no. 26, pp. 21434–21445, 2008.

[9] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation.," *Optics express*, vol. 18, no. 10, pp. 10762–10774, 2010.

[10] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, 2011.

[11] P. Comon, "Independent component analysis, a new concept?," *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[12] M. Lewandowska, J. Rumiński, T. Kocejko, and J. Nowak, "Measuring pulse rate with a webcam; a non-contact method for evaluating cardiac activity," in *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on*, pp. 405–410, IEEE, 2011.

[13] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[14] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4, pp. 411–430, 2000.

[15] H.-Y. Wu, M. Rubinstein, E. Shih, J. V. Guttag, F. Durand, and W. T. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Transactions on Graphics*, vol. 31, no. 4, pp. 1–8, 2012.

[16] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[17] X. Li, J. Chen, G. Zhao, and M. Pietikainen, "Remote heart rate measurement from face videos under realistic situations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4264–4271, 2014.

[18] P. D. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on audio and electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[19] A. Osman, J. Turcot, and R. El Kaliouby, "Supervised learning approach to remote heart rate estimation from facial videos," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 1, pp. 1–6, IEEE, 2015.

[20] A. Lam and Y. Kuno, "Robust heart rate measurement from video using select random patches," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3640–3648, 2015.

[21] W. M. Savin, D. M. Davidson, and W. L. Haskell, "Autonomic contribution to heart rate recovery from exercise in humans," *Journal of Applied Physiology*, vol. 53, no. 6, pp. 1572–1575, 1982.

[22] K. C. Darr, D. R. Bassett, B. J. Morgan, and D. P. Thomas, "Effects of age and training status on heart rate recovery after peak exercise," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 254, no. 2, pp. H340–H343, 1988.

[23] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[24] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3444–3451, 2013.

[25] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 532–539, 2013.

[26] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2, pp. 1150–1157, Ieee, 1999.

[27] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013.

[28] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *European Conference on Computer Vision*, pp. 679–692, Springer, 2012.

[29] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2879–2886, IEEE, 2012.

[30] Itseez, "Open source computer vision library." `https://github.com/itseez/opencv`, 2015.

[31] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[32] A. F. Möbius, *Der barycentrische calcul*. 1827.

[33] J. H. McClellan, R. W. Schafer, and M. A. Yoder, *Signal processing first*. Pearson/Prentice Hall, 2003.

[34] A. Oppenheim, R. Schafer, and J. Buck, *Discrete-time Signal Processing*. Prentice Hall international editions, Prentice Hall, 1999.

[35] K. Barrett, S. Barman, S. Boitano, and H. Brooks, *Ganong's Review of Medical Physiology, 24th Edition*. LANGE Basic Science, Mcgraw-hill, 2012.

Appendices

A Recordings Protocol

At the Beginning of Recordings Session

1. Mount the video camera and the BIOPAC software
2. Collocate a table close to the subject's chair in order to avoid electrodes tensions
3. Initialize *Acqknowledge* and verify the connection between the software and the hardware
4. Define the parameters for the heart rate graphical analysis such that:
    - minimum heart rate: 40 bpm
    - maximum heart rate: 220 bpm
    - midpoint graph: 80 bpm
5. Set the camera to *squeeze* mode at the second option inside the menu
6. Set the gain commutator to *HIGH*
7. Perform white balance:
    (a) focus on the white background
    (b) set *WHITE BAL* commutator to *A* position or *B* position
    (c) press *AWB* button until message *AWB ACH OK* appears
8. Set ND filters commutator to *OFF*

Pre Recordings

1. Verify the recording format to a progressive scanning of 720p and a frame rate of 50 frames per second

2. Ask the subject to sit in the chair and try to minimize as many shadows as possible

3. Set the zoom button to manual

4. Focus the image on the subject:

   (a) set the *RING* commutator to focus option

   (b) set the focus to manual

   (c) close the zoom

   (d) adjust the focus until obtaining a clear image

5. Press *IRIS* button until the appearance of the message: *MANUAL IRIS*

6. Adjust the iris to capture as much light as possible without the appearance of the zebra effect

Subject's Exercising

1. Ask the subject to perform some exercise: going up and down a series of stairs equivalent to the height of a three story building

2. Meanwhile, prepare three adhesive tapes in order to hold electrodes

Recordings

1. Put a conductive gel drop on the electrodes

2. Collocate the electrodes to the subject, such that

   (a) **VN-** pin is placed at the inner left wrist

   (b) **VN+** pin is placed at the inner right wrist

   (c) **GND** pin is placed at the inner right elbow

3. Start video recording and heart rate measurement whithin the first minute after the exercise
    (a) Press *START* to *Acqknowledge* software
    (b) Verify the correct signal reception
    (c) Start video recording and place a bookmark in the software at the same time
    (d) Record a single video of 60 seconds length: the first 20 seconds ought to be a static video, while the remaining 40 seconds ought to be a dynamic video interviewing the subject
4. Repeat the recording and the measurement
5. Verify the correct storage of the recorded materials
6. Start video recording and heart rate measurement five minutes after the exercise
    (a) Press *START* to *Acqknowledge* software
    (b) Verify the correct signal reception
    (c) Start video recording and place a bookmark in the software at the same time
    (d) Record a single video of 60 seconds length: the first 20 seconds ought to be a static video, while the remaining 40 seconds ought to be a dynamic video interviewing the subject
7. Repeat the recording and the measurement
8. Verify the correct storage of the recorded materials
9. Start video recording and heart rate measurement ten minutes after the exercise
    (a) Press *START* to *Acqknowledge* software
    (b) Verify the correct signal reception

(c) Start video recording and place a bookmark in the software at the same time

(d) Record a single video of 60 seconds length: the first 20 seconds ought to be a static video, while the remaining 40 seconds ought to be a dynamic video interviewing the subject

10. Repeat the recording and the measurement

11. Verify the correct storage of the materials recorded in the overall session

Post Recordings

1. Take out the adhesive tapes carefully

2. Clean the remaining conductive gel from electrodes