Camera-Based Heart Rate Extraction in Noisy Environments

Amin Rezaei

Master's thesis

University of Turku
Institute of Biomedicine

February 21, 2023

Master's degree in Biomedical Imaging

Specialization Theme:

Image Processing in Digital Health

Credits:

40 ECTS

Supervisors:

Iman Azimi, PhD

Adjunct Professor, Department of Computing

Antti Airola, PhD

Assistant Professor, Department of Computing

UNIVERSITY OF TURKU
Faculty of Medicine, Institute of Biomedicine

AMIN REZAEI: Camera-Based Heart Rate Extraction in Noisy Environments

Master's thesis, 104 pp.
Master's Degree Programme in Biomedical Sciences: Biomedical Imaging
January 2023

Abstract:

Remote photoplethysmography (rPPG) is a non-invasive technique that benefits from video to measure vital signs such as the heart rate (HR). In rPPG estimation, noise can introduce artifacts that distort rPPG signal and jeopardize accurate HR measurement. Considering that most rPPG studies occurred in lab-controlled environments, the issue of noise in realistic conditions remains open.

This thesis aims to examine the challenges of noise in rPPG estimation in realistic scenarios, specifically investigating the effect of noise arising from illumination variation and motion artifacts on the predicted rPPG HR. To mitigate the impact of noise, a modular rPPG measurement framework, comprising data preprocessing, region of interest, signal extraction, preparation, processing, and HR extraction is developed. The proposed pipeline is tested on the LGI-PPGI-Face-Video-Database public dataset, hosting four different candidates and real-life scenarios. In the RoI module, raw rPPG signals were extracted from the dataset using three machine learning-based face detectors, namely Haarcascade, Dlib, and MediaPipe, in parallel. Subsequently, the collected signals underwent preprocessing, independent component analysis, denoising, and frequency domain conversion for peak detection.

Overall, the Dlib face detector leads to the most successful HR for the majority of scenarios. In 50% of all scenarios and candidates, the average predicted HR for Dlib is either in line or very close to the average reference HR. The extracted HRs from the Haarcascade and MediaPipe architectures make up 31.25% and 18.75% of plausible results, respectively. The analysis highlighted the importance of fixated facial landmarks in collecting quality raw data and reducing noise.

Keywords: Remote photoplethysmography, noise, heart rate, machine learning, independent component analysis, fast Fourier transform, facial landmarks.

Contents

Acronyms

AM Arithmetic Mean.

BFP Band-pass filter.

BGR Blue Green Red.

BP Blood Pressure.

BPM Beats Per Minute.

BPP Bits Per Pixel.

BPS Beats Per Second.

BSS Blind Source Separation.

BT body temperature.

BVP Blood Volume Pulse.

CNN Convolutional Neural Network.

CV Computer Vision.

DIP Digital Image Processing.

DNN Deep Neural Network.

DSP Digital signal processing.

FFT Fast Fourier Transform.

FPS Frames Per Second.

HOG Histogram of Oriented Gradients.

HR Heart rate.

HRV Heart rate variability.

HSV Hue Saturation Value.

ICA Independent Component Analysis.

ITU International Telecommunication Union.

LED Light-Emitting Diode.

MAE Mean Absolute Error.

ML Machine Learning.

MMOD Max-Margin Object Detection.

PCA Principal Component Analysis.

PPG Photoplethysmography.

PSD Power Spectral Density.

RGB Red Green Blue.

RMSE Root Mean Square Error.

RMSSD Root Mean Square of the Successive Differences.

RoI Region of Interest.

RPM Remote Patient Monitoring.

rPPG Remote Photoplethysmography.

RQ The Respiratory Quotient.

RR Respiration Rate.

SBP Systolic blood pressure.

SD Standard deviation.

SMA simple moving average.

SNR Signal-To-Noise Ratio.

SpO^2 Oxygen Saturation.

Chapter 1

Introduction

1.1 Overview

Heart rate (HR) is the periodicity of the heart pumping blood through the arteries in 60 seconds. It is also an indicator of one's cardiovascular strength. Over the years, researchers have developed various devices and techniques related to the measurement of cardiac activity, specifically HR.

Photoplethysmography(PPG) is an affordable and simple HR measurement technique that relies on the readings of an optical sensor placed on the subject's skin. PPG offers non-invasive HR estimation through the utilization of the blood volumetric fluctuations occurring within the small blood vessels located in the body's tissue.

Remote photoplethysmography(rPPG) is another similar approach to PPG that has a different signal acquisition process. Instead of requiring the attachment of an optical sensor to the skin, rPPG benefits from the visual data obtained from a camera sensor to measure the blood volume pressure(BVP) changes. Therefore, as its name implies, it offers remote and non-contact HR measurement.

Over the years, new rPPG models and approaches have emerged. Despite these developments, the limitations and challenges related to the acquisition of the rPPG signal and HR estimation have restricted further expansion and utilization of rPPG in real life.

With the majority of rPPG studies being conducted in lab-controlled environments, the issue of noise in realistic conditions remains open. In controlled studies where illumination is ideal and candidates have no or minimal movement, the rPPG methods face a minimal challenge.

Having said that, this is not the case for uncontrolled studies where the environment is noisy and subjects are moving. In such scenarios, fluctuating illumination accompanied by subject motion challenge HR prediction.

This thesis aims to study the effect of realistic experimental settings on the acquisition of rPPG signals and subsequent HR measurement.

The multidisciplinary essence of rPPG has turned it into a crossing point between image and signal processing. The materialization of machine learning(ML) solutions has facilitated comprehensive image and signal processing operations that once did not exist.

In continuation of prior studies, in this work a novel unsupervised solution for the acquisition and processing of rPPG signals is implemented. This work ventures to offer a broader look at the issue of noise in realistic scenarios. The results of this study will provide further insight into real-world rPPG practices and will contribute to future studies.

1.2 Problem statement

Up until recently, most of the rPPG experiments conducted were too focused on lab-controlled measurements. Over the past couple of years, this paradigm has shifted, and realistic scenarios are becoming more popular.

Because of rPPG reliance on a camera sensor, the collected signals are prone to noise of all kinds. The noise, whether originating from the environment or subject, has the potential to deteriorate the quality of collected signals. Hence, it can lead to inaccurate HR predictions, which places rPPG in a weaker spot in comparison to PPG.

Because of such complications, rPPG measurements will continue struggling in predicting accurate HR, unless limitations of such are recognized and their effects are minimized or neutralized.

In that sense, addressing the rPPG challenges or offering any solutions concerning noisy environments mandates a thorough comparative analysis that is beyond lab-controlled settings. To gain such insight, an easy and applicable rPPG measurement setup can be a good starting point for the materialization of such analysis.

1.3 Motivation

Owing to researchers' growing interest in remote HR measurement techniques, rPPG studies have gained considerable traction. Despite the progress, PPG is still overshadowing rPPG on a daily basis and is the leading HR monitoring method.

To shift this paradigm, it requires primarily recognizing the factors that undermine the utilization of rPPG in realistic scenarios. The acknowledgement of these obstructive causes accommodates an opportunity to efficiently address them and opens the path to the materialization of rPPG on the daily basis.

Apart from addressing these prohibitive factors, it is also necessary to eliminate them or neutralize their footprint on the final results.

In the past few years, the fast pace of development in artificial intelligence (AI)-based solutions has outpaced their application in other fields. For example, in computer vision (CV), AI-assisted high-fidelity face mesh prediction[30] is a cutting-edge, open source and lightweight architecture that maximizes a deep regression algorithm to offer 478 facial landmark points[43].

The inclusion of such an architecture in the signal acquisition process can significantly affect the quality and diversity of the collected rPPG signals. In that sense, another motivation of this work is to implement state-of-the-art algorithms within the signal acquisition process to enhance the quality of collected signals.

1.4 Research questions

This thesis thrives in offering a deeper insight into the factors affecting the quality of camera-based vital signs extraction in noisy environments and seeks to fulfill the following research questions (RQ):

- RQ1: What are the deterring factors that challenge the accuracy of rPPG estimation in noisy environments?
- RQ2: What measures can be taken to enhance the quality of extracted rPPG data in noisy environments?
- RQ3: To what extent do disturbances influence HR estimation?

1.5 Research objectives

The proposal of the following research objectives (RO) addresses the above-mentioned RQs.

- RO1: Conduct a thorough review of previous investigations, followed by the classification of the methods developed, the datasets used, and the common challenges faced by rPPG extraction and noise.
- RO2: Proposing an offset solution to curb the effect of disturbances during the extraction of the rPPG signal.
- RO3: Implementing an rPPG approach on a noisy dataset and discussing the effect of noise on the final results.

1.6 Thesis structure

The thesis is organized into eight chapters as follows:

1-Introduction

2-Background

Targeted towards readers with minimal background on the scientific concept that rPPG is based on. The basics of blood optical characteristics, PPG, and pulse oximetry are shortly discussed in this chapter.

3-Image processing

With image processing being an inseparable part of rPPG studies, articulating the rudimentary image analysis concepts through visualization became a necessity. Through the course of this chapter, the basic concepts of image processing that drive the investigations of rPPG are briefly explained and depicted.

4-Machine learning

Defining ML and its subsets. Exploring and detailing the utilization of computer vision (CV) for face detection purposes. In addition, we briefly explain and implement three distinctive face detection approaches on an artificially produced image.

5-Remote photoplethysmography

The main emphasis of this chapter revolves around explaining rPPG in detail, discussing the mathematical theory behind its existence, and reviewing the literature. In general, this chapter offers a valid and up-to-date study of earlier works on rPPG, its approaches, and its corresponding datasets. To fulfil the objectives of the study, this chapter also scrutinizes the limiting factors facing the utilization of rPPG in realistic scenarios.

6-Materials and methods

The proposal and implementation of a modular architecture that seamlessly syncs with three face detectors and extracts HR. This chapter thoroughly explains the logic behind the inclusion of every process, while implementing it. At the end of this chapter, the reader will have an understanding of the rPPG mechanism and will be able to articulate methodologies that are beneficial to the implementation of remote and robust HR estimation in noisy environments.

7-Results and discussion

Disclosing the results of the proposed method on the dataset at hand, followed by a comparative analysis of the findings, and a discussion revolving the results and limitations of the utilized dataset.

8-Conclusion and future work

As the final chapter, finalizing the highlights of study, drawing conclusions, and addressing future possibilities are the shaping remarks of this chapter.

Chapter 2

Background

2.1 Preface

Vital signs are a set of clinical indicators unveiling the underlying status of the most basic functions of one's body. For example, heart rate (HR), known to be the most common indicator among physicians and medical professionals, helps clinicians during routine patient checkups and health assessments.

Biomedical signal is a term broadly used in academia and industry that refers to the manifestation and observation of physiological events by means of stationary and non-stationary signals [16].

Photoplethysmography is an example of a biomedical signal commonly used among healthcare professionals in clinical and nonclinical settings. In recent years, the use of photoplethysmography in smart wearables and fitness trackers has significantly increased health awareness among health enthusiasts. The following sections further discuss the enabling principle behind photoplethysmography, its mechanism, and application.

2.2 Optical properties of blood

The hemoglobin responsible for blood redness plays a crucial role in oxygen delivery throughout the human body [1]. These proteins oxygenate organs and tissues by detaching the oxygen molecule bonded to their iron, resulting in a change in color in the bed of the skin. The oxygenized skin tissue exposed to visible light absorbs green-blue wavelengths and reflects red-orange lights.

To the naked eye, observing such processes is not relatively easy, mainly due to inconsistencies in tissue oxygenation levels across different organs, In contrast, an optical sensor (photodetector) has a much higher spectral peak sensitivity, and its utilization has the potential to reveal such subtle color changes [2].

Implementation of such sensors is not without setbacks. Factors such as age, physical activity, and skin tone are among the challenges of any optical method. From the difficulties imposed by technical execution to those of data collection and analysis, lies numerous obstacles that require specific solutions [2, 3].

2.3 Photoplethysmography

Photoplethysmography (PPG) is an easy, economical and noninvasive method of screening vital signs that requires a photodetector and a light source to interpret the blood brightness variation in the bed of tissue into a viable biosignal.

The function of PPG benefits from the variations in light absorption by the hemoglobin proteins of blood in the microvascular bed of tissue. To utilize PPG, it is essential to have an optical sensor, such as a photodetector, which detects the periodic pattern of blood flow created by the heartbeat as it pumps blood through the veins and arteries. This pattern is crucial for accurate measurement of various physiological parameters, such as HR and oxygen saturation.

Studies have demonstrated that the quality of the readings obtained by PPG not only relies on the optical properties of blood, but also depends on other parameters, such as skin tone, obesity, skin structure, skin temperature, and measurement environment [4, 5].

2.3.1 Pulse oximetry

The most well-known use case of PPG is the pulse oximeter device, which enables non-invasive monitoring of physiological signs, easily and at a fraction of the cost of other techniques. With no exception, the majority of fitness trackers and health wearables utilize some form of PPG sensor to offer remote physiological measurements such as HR.

A pulse oximeter uses an optical sensor accompanied by a set of infrared or LED lights to display the difference between oxygenated blood vs non-oxygenated one. The optical biosignals yielded through this method contain valuable information about physiological vital signs, such as HR, blood volume pressure (BVP), and heart rate variability (HRV).

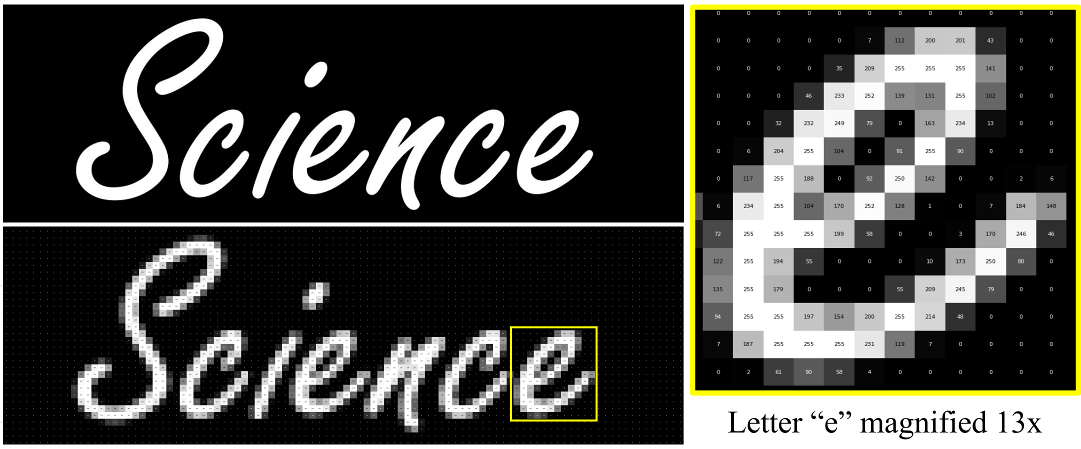
Chapter 3

Image processing

Image processing is a vast and rapidly growing field that plays a crucial role in a variety of industries, including medical imaging, remote sensing, and computer vision. At its core, image processing involves the manipulation and analysis of images using computational techniques. This chapter explores and visualizes the basic concepts of image processing.

3.1 Pixel and pixel value

In digital imaging, the pixel is the smallest component of a digital image stored in a binary format. The term first appeared in scientific papers after the publication of two different articles [7] in SPIE proceedings in 1965 and has a very context-specific definition. The pixel value is a measure of pixel brightness, which relies on the image type and color space to indicate the pixel color. In digital images, where each pixel is a small component of a digital image, a pixel value is the magnitude of each pixel. Figure 3.1 depicts various perceptions of an image in a digital device.



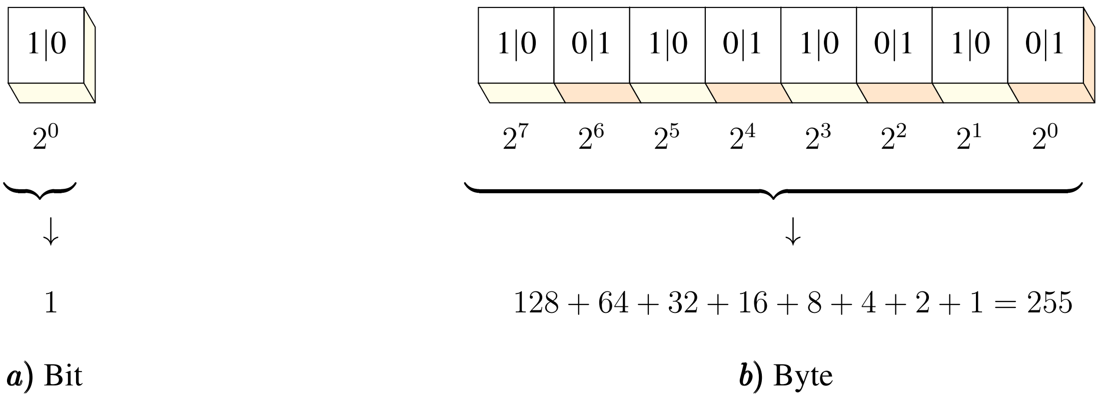
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	7	112	200	201	43	0	0
0	0	0	0	35	209	255	255	255	141	0	0
0	0	0	46	233	252	139	131	255	102	0	0
0	0	32	232	249	79	0	163	234	13	0	0
0	6	204	255	104	0	91	255	90	0	0	0
0	117	255	188	0	92	250	142	0	0	2	6
6	234	255	104	170	252	128	1	0	7	184	148
72	255	255	255	199	58	0	0	3	170	246	46
122	255	194	55	0	0	0	10	173	250	80	0
135	255	179	0	0	0	55	209	245	79	0	0
94	255	255	197	154	200	255	214	48	0	0	0
7	187	255	255	255	231	119	7	0	0	0	0
0	2	61	90	58	4	0	0	0	0	0	0

Letter “e” magnified 13x

Figure 3.1: Top-left is the mode by which a digital device illustrates an image. Top-bottom map’s pixel brightness intensities. The highlighted yellow area shows the relationship between pixel value and brightness intensity.

3.2 Data in digital devices

’Bit’ is the smallest and most fundamental unit of measurement that defines a logical state (0 or 1). ’Byte’, on the other hand, is the concept of arranging bits and is the universal standard for measuring and addressing digital data. One byte includes eight blocks of bits, where each bit is either 1 or 0. From a mathematical stance, n bits generate two to the power of nth different combinations (2^n). Figure 3.2 demonstrates possible results in the arrangement of bit and byte building blocks in a digital device.



a) Bit

b) Byte

Figure 3.2: **a)** The total number of combinations that one-bit yields are 2^1.**b)** The total number of combinations that one-byte yields is 2^8, denoting the existence of 256 ways that 1 and 0 can arrange in 8 blocks of bits.

3.2.1 Digitization

A digital image belonging to the Euclidean space is a function of two variables, $f(x, y)$ in the Cartesian system. Digitization is the process of transforming this function into binary data, and it is achieved by using a digitizer to sample and quantize an analogue image into a digital one. In other words, it is the vehicle that facilitates the transition from analog to digital. As illustrated in 3.1, digitizing an image produces a matrix of discrete numbers, where each matrix element describes a point on the Cartesian system and holds an integer value ranging from 0 to 255 [8, 10].

3.2.2 Sampling

Sampling is the process of converting an analogue function (intrinsically continuous) into a non-continuous (discrete) one. Sampling digitizes the coordinate values of an analog signal and is the preparatory step to quantization [8]. In $f(x, y)$ where x is the time axis and y is the amplitude, sampling discretizes time while allowing continuity of amplitude, dictating the resolution of the digitized image. Signal sampling depends on the dimension of the collected signal [8]. For example, in the case of a 2-d signal (comprising data points and time stamps), the signal reconstruction process relies on a single frequency component where $f_s \geq 2f_m$(signal sampling theorem).

3.2.3 Quantization

Quantization is the consecutive step after sampling that converts the variables of a continuous function into a noncontinuous (discrete) one, which digitizes the amplitude values of an analogue signal. Quantization controls the gray aggregation degree in a digitized image by discretizing the amplitude values and maintaining the continuity of time.

Unless finite empty bits are available, storing an infinite value on any digital device is an impossible task. Hence, to store and retrieve an image in digital space, it requires quantizing it to a finite amount of bits that can be stored on a digital storage [8].

Although sampling and quantization are crucial to harness the versatility and robustness of digital systems, to some extent they attenuate image quality during the conversion process.

The effect of aliasing, and the quantization noise in the digitized image introduces a certain degree of degradation in the converted image. Therefore, it is necessary to define a proper sampling interval and quantization degree by which only digital images are produced within acceptable quality metrics.

For instance, in the case of a digital video where scenes are rapidly alternating, a lower quantization level and a higher sampling interval ensure acceptable video quality. The same principle is also applicable to videos having a lower alternation per frame. Such images demand a higher quantization level and lower sampling rate [8].

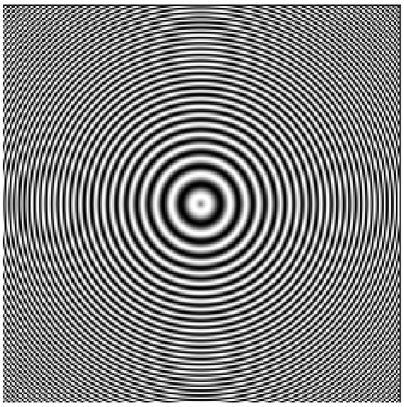


Figure 3.3: Circular waves on each corner depicting the effect of aliasing in an image.

3.3 Digital image

A digital image can be defined as an array of small integers called “pixels” recorded in computer-readable binary format. A discrete image belongs to Euclidean space, meaning that it follows the same geometric principles governing the Euclidean plane.

In a 2-D space, the width, and length are the geometric dimensional measures that determine the placement of the pixel elements, thus the array associated with an image (pixel array) is a matrix of m columns × n rows. A 2-D matrix of integers, such as the one presented below, renders a $w{\times}l$ sized image in the Euclidean plane [8].

$$I_{(X,Y)} = \begin{array}{c|c} & \text{m columns} \longrightarrow \\ \text{n rows} \downarrow & \begin{pmatrix} I_{(0,0)} & I_{(0,1)} & I_{(0,2)} & I_{(0,3)} & I_{(0,4)} & \cdots & I_{(0,y-1)} \\ I_{(1,0)} & I_{(1,1)} & I_{(1,2)} & I_{(1,3)} & I_{(1,4)} & \cdots & I_{(1,y-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ I_{(x-1,0)} & I_{(x-1,1)} & I_{(x-1,2)} & I_{(x-1,3)} & I_{(x-1,4)} & \cdots & I_{(x-1,y-1)} \end{pmatrix} \end{array}$$

Each matrix entry represents the Cartesian coordinates associated with the corresponding picture elements that form an image.

Figure 3.4, illustrates the formation of an image in the Cartesian system based on the intensity of the brightness of the pixels. It is worth noting that the x and y in the $I_{(x,y)}$ entry are two independent variables and do not interfere with one another. In a digital image, the total number of picture elements of an image is the display resolution, quoted as width×height.

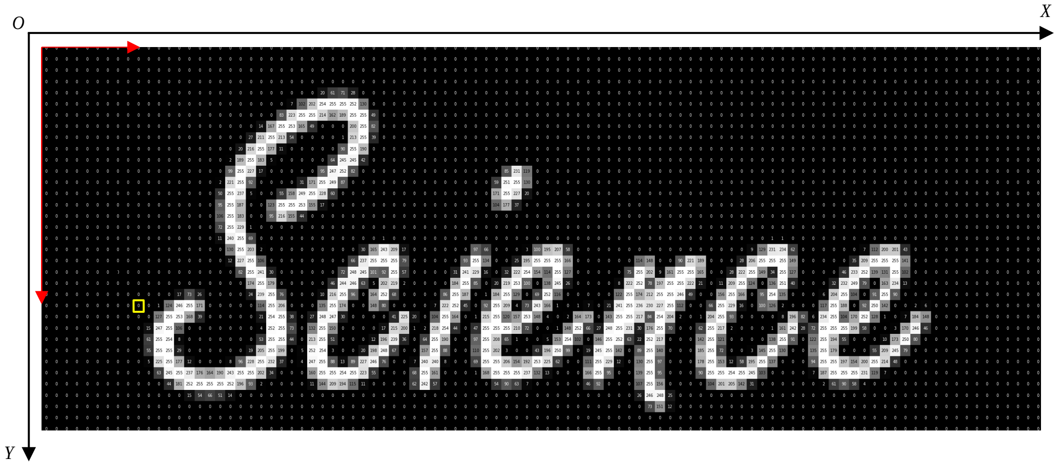
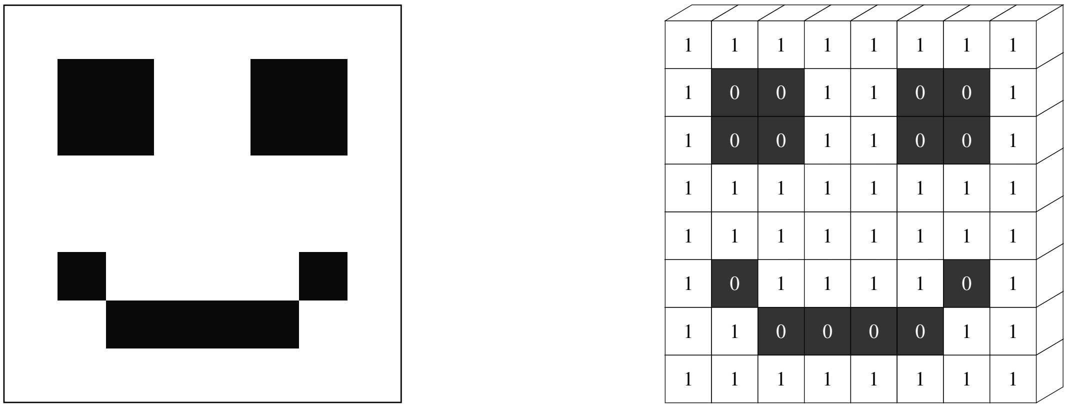


Figure 3.4: In the Cartesian coordinate system, the points where x and y axes congregate is regarded as the origin (O) and has coordinates of $(0,0)$.The entry $I_{(10,24)}$, highlighted with yellow, 10 is the distance from point O on the $X - axis$, while 24 is the distance from point O on the $Y - axis$.

3.3.1 Binary image

A binary image is the simplest form of an image, in which pixels are only black and white. Such images are often referred to as bitmaps. In a bitmap, the value of each pixel is either 1 or 0 denoting the binary format. Compared to other kinds of images, Bitmaps occupy the least storage space in digital devices. Having said that, these images are expected to have a higher resolution to illustrate an effective depth range [8]. The pixel value in binary images is a one-bit number, signified either as background or foreground. Figure 3.5 demonstrates that the pixel values assigned to the white and black pixels of a binary image are 1 and 0 respectively.



1	1	1	1	1	1	1	1
1	0	0	1	1	0	0	1
1	0	0	1	1	0	0	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
1	0	1	1	1	1	0	1
1	1	0	0	0	0	1	1
1	1	1	1	1	1	1	1

Figure 3.5: A bitmap (left) followed by its building blocks (right).

3.3.2 Grayscale image

A grayscale image is a mode of displaying picture elements of a discrete image by their brightness intensity [10]. As Figure 3.6 shows, these images comprise all shades of gray and differ from binary images. Compared to color images, grayscale images take less place in storage. In comparison to bitmaps, they do convey more information as a result of having an equal intensity between their red, green, and blue components. Each pixel of a grayscale image takes one byte of data or 8 bits of information, denoting 256 possible colors, where 0 is pitch black, 255 is white, and everything in between is different shades of gray.

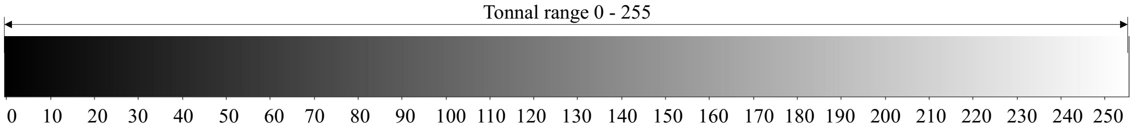


Figure 3.6: Grayscale dynamic range.

3.4 Color Model

Color model, also known as color space, is a method of explaining and specifying the arrangement of colors across all the picture elements of an image. Color models are a set of mathematical models and transformations that control and regulate colors in images [8, 9]. The development of various color spaces, such as CIE 1931, XYZ, CMYK, CMY, CIELUV, HSLUV, HSV, YCbCr, and YUV have opened the path to a range of image and video processing tasks in a variety of settings [11]. Despite these advances, RGB is still the most favorable model and the go-to standard when it comes to imaging. Figure 3.7, illustrates the conversion of BGR to RGB color spaces through the replacement of the blue and red axis.

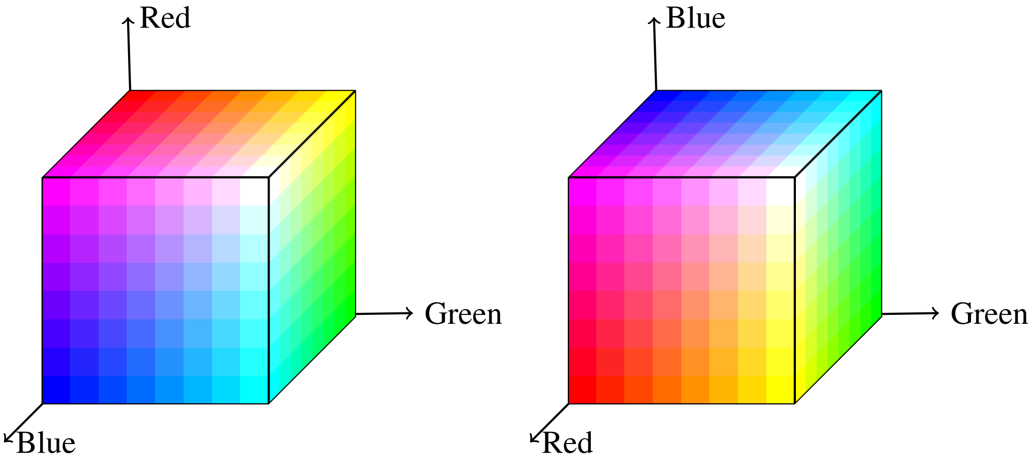
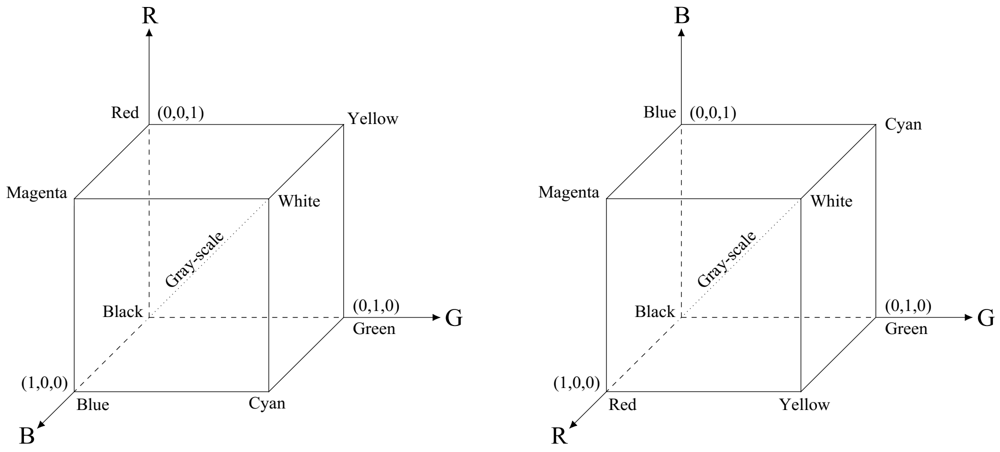


Figure 3.7: BGR and RGB color cubes.

The RGB and BGR color models feature cubical structures in the 3D space, and the colors are specified by their coordinates on the cube planes, Figure 3.8.



b) BGR Colour model

a) RGB Colour model

Figure 3.8: BGR and RGB color models.

3.5 Channel

A channel is a feature of colored images that caches one color data from the primary color components of a color model (e.g., RGB or CMYK). In other words, it is the grayscale image of a color image with similar dimensions. While channels of an image are autonomous from each other, their attributes depend on the color model.

For instance, in the RGB color model, channels are red, green, and blue. Contrastingly, in the HSV model, channels are hue, saturation, and brightness. Figure 3.9 decomposes an image in the RGB color space into its sub-channels.

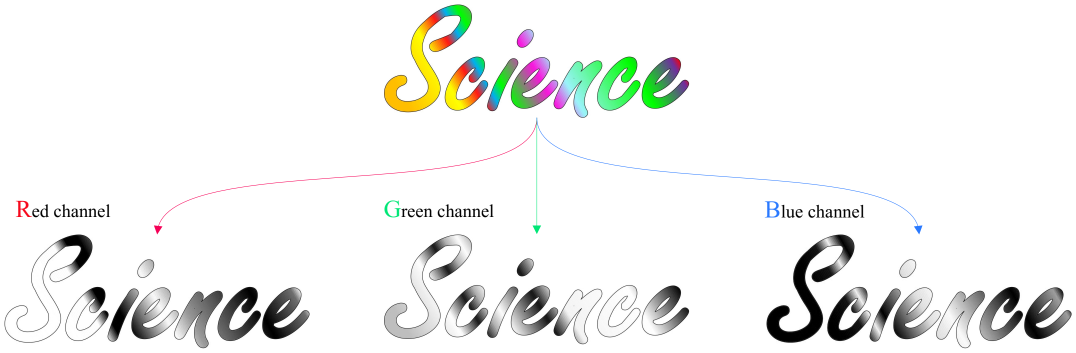


Figure 3.9: The combination of the red, green, and blue channels forming a color image in the RGB space.

3.6 Color depth

As mentioned in Section 3.2, computers use bits to store, retrieve, transfer, and manipulate digital data. In computing and digital communication, where images are parsed into pixel components, and each pixel is cached as bits of information, the depth of the color represents the number of bits that each pixel caches on the disk space.

Color depth has a unit of bits per pixel (bpp), and its value hinges on the image color space and the number of channels. For instance, in grayscale images where only one color channel exists, the color depth is 8 bpp, implying the possibility of having 256 color combinations. That being said, in the RGB color model with 3 channels, the color depth is 24 bpp, denoting 2^{24} possible color combinations [8].

Figure 3.10, presents the building blocks of an RGB color model next to its color depth.

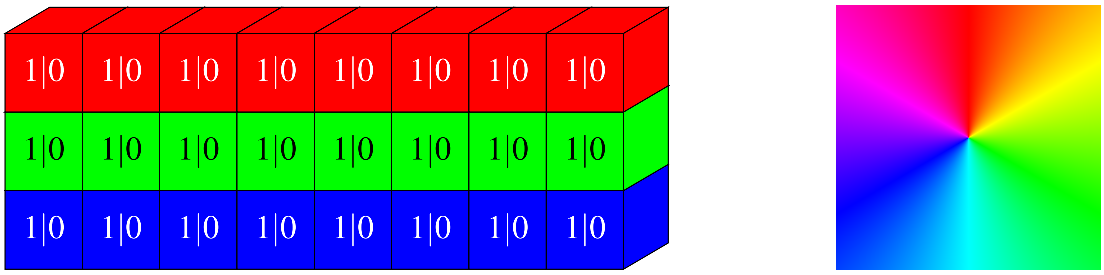


Figure 3.10: Byte building blocks (left). Different shades of primary colours (right).

3.7 Color image

An image comprised of primary colors and their linear combinations is the simplest definition of a color image. Figure 3.11 illustrates a color image across different color spaces. The pixel organization in a color image follows the same principles of binary and grayscale images with an addition of being influenced and controlled by a color model [10]. As depicted, not all color spaces are eye-friendly and convey meaning, therefore the definition of color is more of a concept rather than a fact in digital images.

For instance, in the RGB color space, the red, green, and blue colors and their linear combination creates a color image, whereas in the HSV color space, colors are defined by the intensity levels of hue, saturation, and brightness.

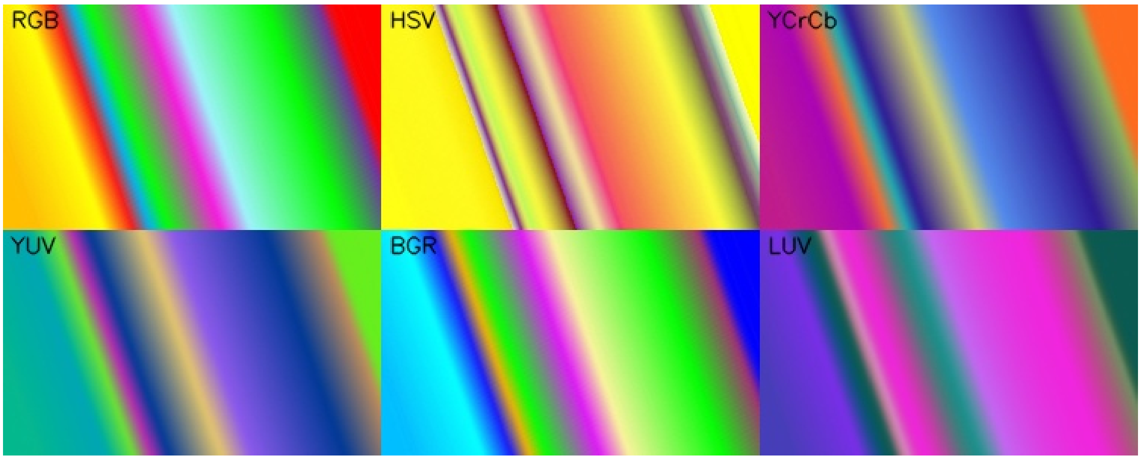


Figure 3.11: Comparison of colors in different color spaces.

A colored image in the RGB space is characterized by 3 matrices in the two-dimensional Euclidean space. Figure 3.12, compares the perception of an image in the eye of a human (a) versus a computer (b).

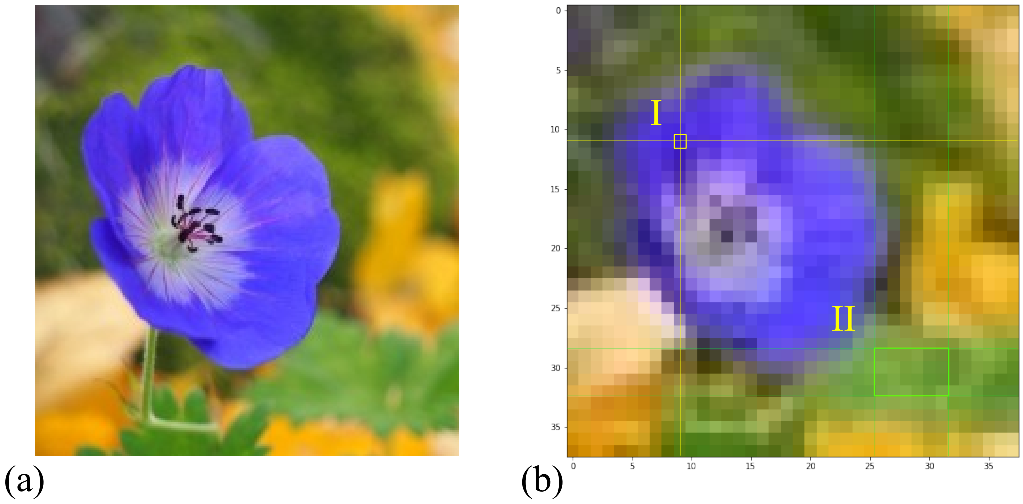


Figure 3.12: In the image (b), the yellow region cached as an array of integers (I_{RGB}) sitting at coordinate (9, 11) hosts one purple pixel, and the green region sitting at (II_{RGB}) hosts an array consisting o different shades of green.

Matrix $(I_{3\times1})$ 3.1 is a typical example of pixel components in RGB space.

$$I_{RGB} = [\boxed{63}\ \boxed{29}\ \boxed{212}] \tag{3.1}$$

- pixel value of the red channel
- pixel value of the green channel
- pixel value of the blue channel

Unlike the above matrix, region $(II_{6\times4})$ encapsulates a set of coordinates and covers an area of 24 picture elements. As depicted below, it includes all the pixels residing between the vertical axis 25 to 32 and the horizontal axis 29 to 33, and is characterized by the (II_{Red}), (II_{Green}), and (II_{Blue}) matrices in a 2D space.

$$\begin{array}{ccc} II_{Red} & II_{Green} & II_{Blue} \\ \begin{matrix} 68 & 60 & 59 & 66 & 55 & 50 \\ 72 & 67 & 57 & 62 & 53 & 52 \\ 60 & 66 & 55 & 55 & 49 & 48 \\ 67 & 72 & 71 & 59 & 54 & 65 \end{matrix} & \begin{matrix} 166 & 163 & 160 & 165 & 150 & 144 \\ 176 & 172 & 161 & 164 & 151 & 147 \\ 165 & 169 & 160 & 159 & 150 & 147 \\ 165 & 170 & 174 & 165 & 156 & 170 \end{matrix} & \begin{matrix} 124 & 126 & 128 & 133 & 116 & 110 \\ 129 & 129 & 120 & 122 & 105 & 103 \\ 116 & 124 & 115 & 112 & 99 & 99 \\ 119 & 126 & 129 & 118 & 108 & 121 \end{matrix} \end{array}$$

The above matrices help computers identify the components of an image and their intensity in a region of interest (RoI). The elements of these matrices also reveal the channel dominance of the pixel components.

The area in the region II is an example of the aforementioned, as it is mainly composed of green shades. By comparing the matrices, it is clear that the intensity of the green pixel components is well above other channels, denoting the dominance of the green channel.

3.8 Spatial domain

The spatial domain is an enabling tool in image processing, which is defined as the mode of representing an image by its pixel values[8]. Direct manipulation of pixel values through various operations, such as grayscale transformation, neighborhood averaging, discrete formulation, image smoothing, edge-preserving, and histogram equalization fall under spatial domain territory. As illustrated by Figure 3.13, the amplitude profile of an image is one of the instances of spatial domain applications in image processing.

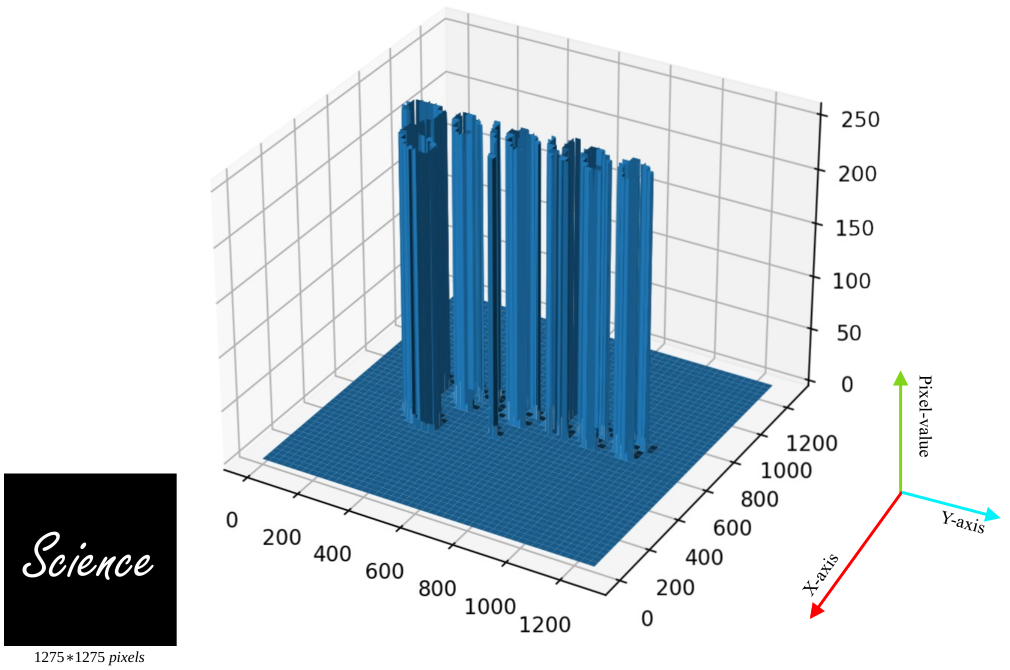


Figure 3.13: The x-axis and y-axis are respectively the representatives of width and height of the black 1275×1275 image. The z-axis (green) is the pixel value intensity, for RoIs in white, this value equates to 255 and for black areas it equates to 0.

3.9 Spatial resolution

Spatial resolution, also known as pixel density or image resolution, is the presentation of pixels of a desired area within the image frame by which an object or scene is captured [8]. It is measured by counting every pixel towards x and y axis distance unit, and it dictates the quality and fineness of details.

Depending on the nature of an imaging task, the spatial resolution can vary between fractions of nm to physical dimensions greater than km. For example, microscopic images obtained by electron microscopy have spatial resolutions of approximately 0.1 nm.

Fine reduction in spatial resolution results in the appearance of pixel blocks in an image. Although this might be unpleasant to look at, in some applications, like camera-based remote patient monitoring (RPM), a few pixel blocks are all needed to conduct an experiment.

3.10 Digital video

A digital video is a sequence of time-varying images obtained or broadcast successively over a time window. The frequency at which this phenomenon occurs is the frame rate expressed in frames per second (fps), and it defines the concentration of distinct still image sequences.

While an image is a spatial distribution of pixel intensities and a two-dimensional signal, a video is a spatial distribution of pixel intensities propagating over a temporal domain and therefore a three-dimensional signal, Figure 3.14.

Although 24 fps is the most common frame rate employed nowadays, the hardware and software properties of the video capturing or displaying device can still influence the fps value. Concerning the quality of a video, resolution, codec, bit rate, frame rate, color depth, and bit control mode are the factors that directly determine the fate of a video.

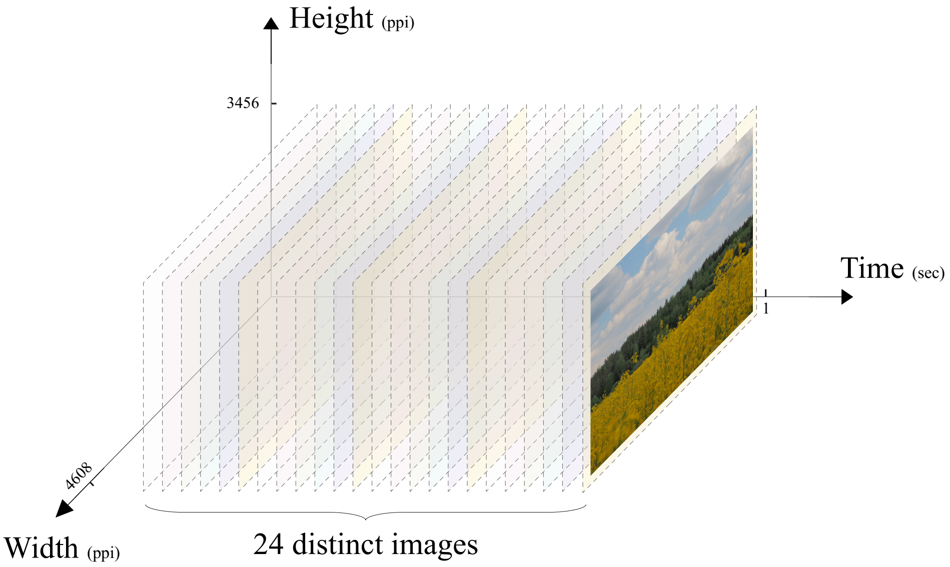


Figure 3.14: Demonstration of one-second digital video captured at 24 fps over its temporal domain.

Chapter 4

Machine learning

4.1 Preface

Owing to the development of sophisticated mathematical algorithms, faster processing units, and advanced data storage technologies, machines can now execute a range of tasks on an overwhelming volume of complex datasets, such as biomedical signals [118]. As a result of these significant technical advances, new innovative solutions have emerged that once did not exist, including machine learning.

4.2 Machine learning

Machine learning (ML) is a method of teaching computers to learn and make predictions or decisions based on data, without explicitly programming them to do so. It involves the use of algorithms and statistical models to analyze and interpret data, and identify patterns and trends.

ML actualizes automatic learning and improvement through training and validation processes of data entries (training dataset) and delivers a predictive and highly scalable model from scratch [22]. The objective of ML systems revolves around the use of data in describing what has happened; predicting how it will happen, or even prescribing necessary actions.

The field of ML is often divided into four subcategories. Supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning [22]. For the context of this thesis, we focus on supervised and unsupervised learning.

4.2.1 Supervised learning

This category bases its operation on training a model with classified (labelled) examples. Supervised learning enables a number of capabilities in ML models. Some of the key capabilities that supervised learning empowers include: improved accuracy and reliability, enhanced predictive power, increased flexibility and adaptability, and greater transparency and interpretability. The preliminary idea of supervised learning is to present the machine with adequate example inputs (training data) accompanied by desired outputs (test data) and expect it to determine the correlation between the two [22].

In supervised learning, classification and regression are two common methods in practice. The classification goal is to predict the class or category of an input data point. Figure 4.1 provides a comprehensive overview of the supervised learning pipeline, including the raw data inputs, the algorithm utilized for processing, and the resulting model output.

Supervised learning is typically done by training a model on a labeled dataset, where the output is already known, and using that model to make predictions on new, unseen data. Face detection is an example of a classification problem in supervised learning, where the input data would be an image, and the algorithm would predict whether or not the image contains a face.

Regression, on the other hand, aims to predict a continuous numerical value, such as a price or a probability. This is typically done by training a model on a labeled dataset, where the output is already known, and using that model to make predictions on new, unseen data [22].

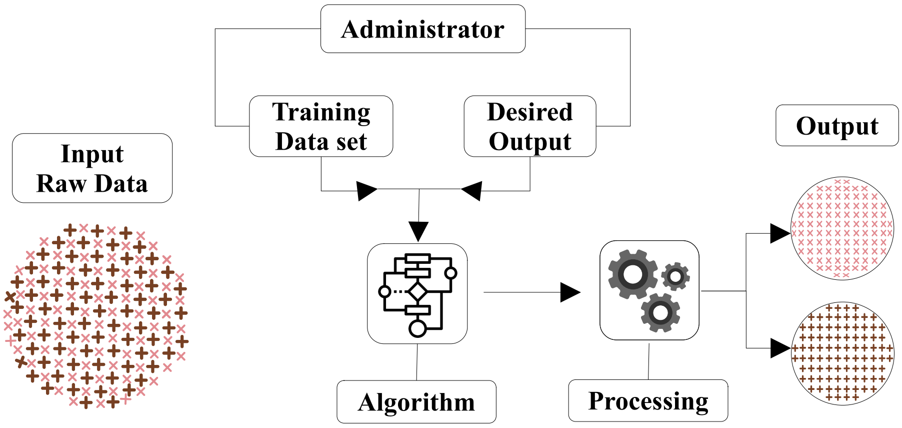


Figure 4.1: An schematic overview of a supervised learning pipeline

Classifier

A classifier is a ML model that is trained to predict the class or category of a given input. This means that the model is trained on a labeled dataset, where the correct class or category for each input is provided. The model uses this training data to learn the relationships between the input data and their corresponding classes or categories, and then uses this knowledge to make predictions on new, unseen data.

There are several benefits to using classifiers in ML. Classifiers can be used to make predictions on new, unseen examples. They can handle complex data. They can be fine-tuned by adjusting their hyperparameters to achieve better performance and efficiency [23]. Because of classifiers, ML pipelines are now capable of allocating a class label to data entries in real-time [32, 28] and can process multiple tasks simultaneously (synchronous recognition of facial features) [29]. In studies that require some form of image recognition, specifically face detection, the utilization of a well-trained classifier speeds up the pace of the study and reduces the errors during the detection process.

4.2.2 Unsupervised learning

Unlike the previous approach, unsupervised learning seeks to find concealed patterns and trends in datasets whose entries are labelless. The focus of algorithms in this approach is directed towards the recovery of vast structures within the given inputs. Its capacity in detecting resemblances and inconsistencies has made it an optimal method in image recognition, segmentation, as well as exploratory data analysis and signal processing.

Among many approaches within unsupervised learning, clustering, dimension reduction, and association rule mining are considered to be the most important areas [22].

Some real-world applications of unsupervised learning are in computer vision for object recognition; anomaly detection for finding atypical data entries, and in biomedical imaging for segmentation, classification, and detection of tumors.

Independent component analysis (ICA) and principal component analysis (PCA) are two of the most repetitive and practical techniques within unsupervised learning and dimensionality reduction[22]. In the following sections, the two will be discussed and their applications and performance will be assessed in the analysis of biosignals.

4.3 DNN

A deep neural network (DNN) is a class of artificial neural networks (ANNs), by which a computer recognizes objects and patterns [22]. The term "deep" in DNNs implies the depth characteristic of these models, stemming from a multi-layered architecture.

The depth of a DNN, which is translated to the number of layers a DNN consists of, determines the complexity of patterns it can recognize. Deeper networks bring about more power in recognizing more complex patterns.

DNNs have different types, such as Convolutional Neural Networks (CNNs) or shortly ConvNets, Generative Adversarial Networks (GANs), and Recurrent Neural Networks (RNNs).

In the past couple of years, ML methods have shifted significantly towards DL-based techniques. Widely used in certain areas from natural language processing (NLP) to image and speech recognition[1], DL techniques have demonstrated extraordinary results. Our daily lives have become so acquainted with examples of DL applications that we might not notice how deeply It has influenced the world.

4.3.1 CNN

As its name implies, a CNN is a class of neural networks featuring the convolution operation. CNNs are the standard method of artificial intelligence (AI) for computer vision tasks, such as face detection, image classification, segmentation, and more [22, 28].

Continuously applied to all pixels of an image, a convolution operation extracts certain features of an image, providing valuable information about a given region of interest. This operation enables a learning system by which the computer learns not only from one pixel but also from the information of surrounding pixels [20].

A CNN, being a DNN by nature, embodies specific layers in its architecture including input and output layers from bottom to top, as well as convolution, pooling and fully-connected layers in the middle. The arrangement of CNN layers is based on the problem, image dataset and accessible computational power [22]. For instance, AlexNet is made up of 8 layers while VGG16 incorporates a deeper 16-layer design, resulting in significantly higher accuracy in the ImageNet image classification challenge [24].

[1] A DNN is the basis of most DL algorithms in image recognition.

4.3.2 Kernel

Convolution operation is a mathematical calculation performed on pixels of an image using a filter (a kernel) moving over the image pixels. The kernel is a matrix of specific numbers (weights) applied to the input image pixels by a convolution operation.

The output of this operation is a new set of pixel values representing a specific pattern in the image [33, 22]. The learning process for a computer in a DL pipeline is to obtain the weights in the convolution operation during countless running of specific algorithms to reduce the errors of the model in recognizing a specific pattern.

4.4 Computer vision

Computer vision (CV) is a subset of computer science and digital image processing that enables image perception for deriving valuable visionary and non-visionary data from visual inputs in computers and digital devices. CV uses AI to work around image data and solve problems without the need for supervision [34].

Albeit, human vision is quite incredible, but it is limited to certain disadvantages embedded in its physiological structure. For instance, measurements of dynamic image attributes, and detecting and tracking anomalies at a consistent pace without mistake, are impossible through the eyes. Whereas CV not only empowers such measurements but also assists in the manipulation of visual data in different ways.

In addition to a static image and dynamic processing, CV also accommodates quantitative and numerical analysis, night-vision systems, uninterrupted iterative execution, recognition of subtle changes and patterns, object detection, and remote sensing.

4.5 Face detection

Face detection is a subset of CV that materializes the placement and identification of human faces in digital visual data. In recent years, substantial progress has been made in the development of face detection classifiers, underscoring the importance of continued research in this area. The development of such classifiers has sparked the possibility of conducting countless studies, including rPPG.

Rationalization of these classifiers in healthcare [34] has become the preparatory step towards the realization of remote non-contact HR measurements. They facilitate infrastructure for ultimate data collection. In the following, some of these algorithms are briefly introduced and put to test.

4.5.1 Viola-Jones

The object detection framework developed by Viola and Jones [32], developed in 2001, is an ML-based algorithm that classifies images under the value of simple features, reminiscent of Haar-like features. The training of the Viola-Jones algorithm on a collection of manually labelled face and non-face images (4916 images) with a fixed resolution of 24× 24 pixels resulted in a face detection classifier referred to as haarcascade. As an object detection-based algorithm undeterred by the location and scale of objects in images and videos, haarcascade offered fast, real-time frontal face detection, and tracking [32]. In practice, Haar-like features mimic the convolutional kernels. Figure 4.2 depicts some of the standard Haar features and their application [32]. The implementation of haarcascade has rationalized an array of studies in CV . Its application in the field of biomedicine has actualized non-invasive remote photoplethysmographic measurement. For instance, a wide range of studies concerning non-invasive RPM and remote HR extraction consider Haar as their main probe for the retrieval of cardiac data [6, 25, 26, 27].

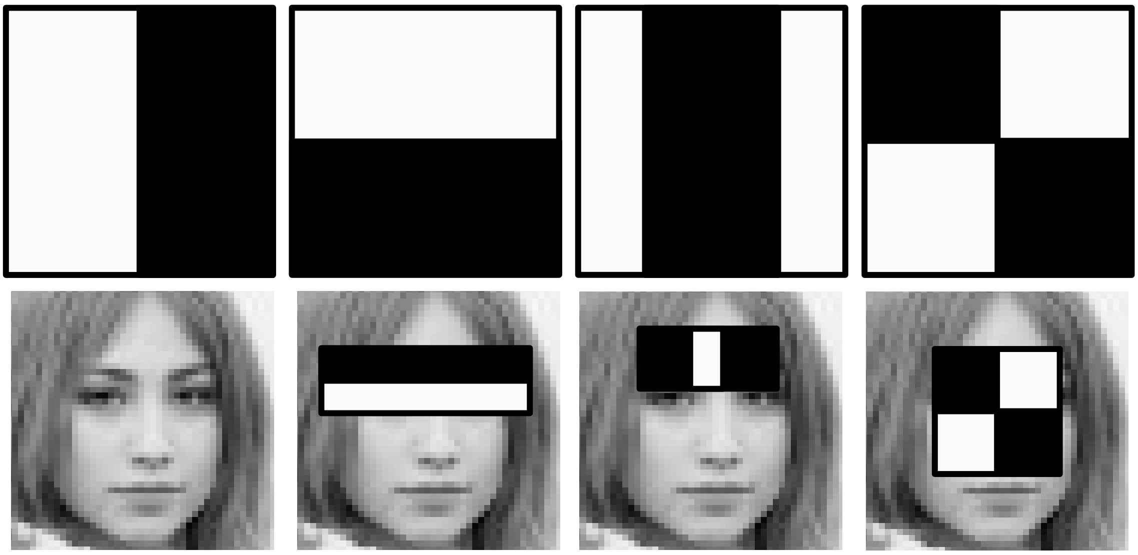


Figure 4.2: From top left to top right are the Haar features suitable for the detection of edges (the first and second), lines (the third) and diagonals (the fourth). From bottom left to bottom right, the first image is a 24×24 pixels training sample and the rest are instances of Haar features focusing on the detection of facial properties, such as eyes (second image), nose (third image), and cheeks (third image).

Haarcascade was once thought to be a strong face detection algorithm, but it did not live up to the reputation. Its incompetency against rotation and acute facial postures [129], on top of its vulnerability against false positives [130, 130], followed by its demanding nature for fine parameter tuning in complex scenarios prevent it to become the ultimate face detection solution [131].

4.5.2 Dlib

Dlib [35], developed in 2009 is a robust, open-source ML-based toolkit.It provides various implementations for kernel-based methods, Bayesian networks, and tasks, such as clustering, anomaly detection, and more [35]. Dlib offers a pre-trained face detection algorithm that acts like a mapping function and enables facial feature annotation.

This pre-trained model often referred to as Dlib facial landmarks encapsulate various facial regions including eyes, chin, nose, and jawline. By default, it offers 5 and 68 different facial points [35], but there is also an extended version which offers 81 facial points and includes forehead as well [36]. Dlib accomplishes face detection through two different approaches, namely HoG and MMOD.

HoG

HoG [39], known as the histogram of oriented gradients combined with a sliding window mechanism, pyramid, and the linear support vector machine (SVM), caters to instantaneous frontal face detection and some non-frontal face estimation [40]. As an object detector algorithm, HoG does not require extensive computation power, and as a CPU-based module, it is relatively fast and capable in a variety of settings [38]. Similar to Haar, HoG practicality is also limited by facial postures at acute angles. Figure 4.3 visualizes and compares the facial landmark pointing system based on HoG on an artificial intelligence-generated fake person [37].

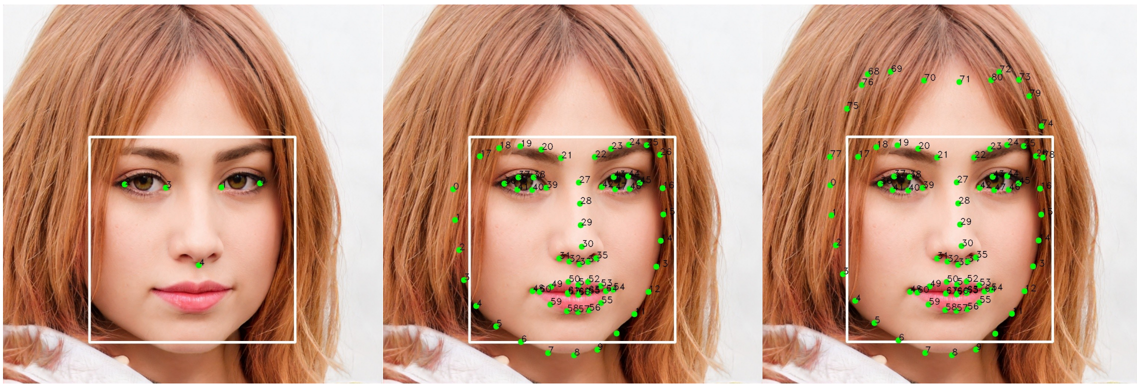


Figure 4.3: From left to right, AI-generated images annotated with green dots represent the 5, 68, and 81 landmark points. The numbers next to each green dot are also representatives of the landmarks and hold the coordinates corresponding to the facial feature or point. The white rectangle, referred to as the bounding box, indicates the presence of a face within its frame.

MMOD

Dlib maximum-margin object detector (MMOD) [38] is a CNN-based face detection algorithm that exploits deep neural networks (DNN) for the analysis of visual data. This approach leverages Dlib-toolkit [35] combined with CNN features to detect faces and offers much more accuracy than HoG. The implementation of CNN algorithms in this method has given it an edge against HoG and other non-CNN solutions. Similar to HoG, MMOD also offers a pre-trained face detection model (referred to as weights) for the retrieval of the face and its features. Not only MMOD achieves face detection on frontal faces but also performs well in complex scenarios whereby faces are at acute angles [38]. In terms of performance, MMOD relies on a graphics processing unit (GPU) and offers extreme computational power. Although MMOD does not have facial landmark annotation, it does produce coordinates corresponding to the face. Because of dlib's open-source algorithm, it is possible to use the produced coordinates and highlight the detected face as an object within a bounding rectangle, Figure 4.4.

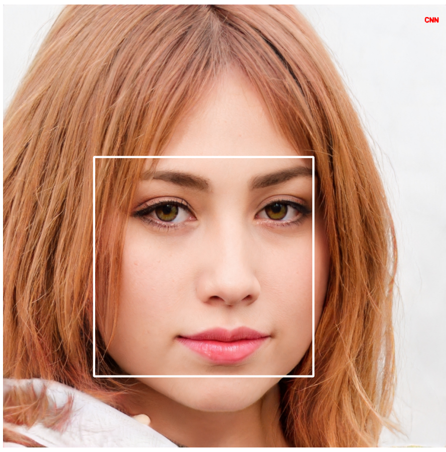


Figure 4.4: Dlib MMOD face predictor produces coordinates from four corners of the face. These coordinates are then utilized to draw a bounding box around it.

4.5.3 MediaPipe

MediaPipe [28] is an ML framework that empowers the creation of modular perception pipelines and is an enabling tool in theorizing arbitrary sensory data. MediaPipe takes streams of information, such as image data and delivers perceived narratives like face detection and landmark streams, object localization, and tracking. MediaPipe was developed with the aim of serving both as a research and development platform and as a source for ready-to-use ML applications. Because of that, it allows rapid prototyping and utilization of production-ready ML applications [28].

MediaPipe face detection and face mesh are two of the most prominent and ready-to-use ML applications concerning the detection and annotation of the face and facial landmarks [28, 29, 30, 31]. Face mesh and detection do not require high computational power and can deliver promising results regardless of the main type of processing unit. In the following, the two are briefly explained.

MediaPipe face detection

MediaPipe face detection is a profound and lightweight face detector that benefits from the famous sub-millisecond neural face detection algorithm designed for mobile GPUs (BlazeFace) to offer real-time, ultra-fast, and on-the-spot face detection [41]. MediaPipe face predictor offers ultra-high stability and performance in complex scenarios. Being immune to acute head angles has given it an edge concerning face profile visual data. In applications demanding accurate facial RoI data entries, the MediaPipe face detector is a very suitable candidate. By default, it facilitates multi-face detection and comes with 6 facial landmarks. It also profits from a CNN architecture combined with a single shot detector and an enhanced alternative of a non-maximum suppression algorithm in delivering super-realtime face detection.

Figure 4.5 portrays MediaPipe face detection in action. As illustrated below, a total of 6 facial landmark points have been detected. The degree of face predictor accuracy referred to as face confidence equating to 0.95, denotes 95% of certainty in the prediction.

Face number: 1

Face confidence: 0.95

Face bounding box coordinates:

xmin: 0.163
ymin: 0.315
width: 0.616
height: 0.616

Right eye:
x: 0.370
y: 0.499

Left eye:
x : 0.630
y : 0.488

Nose
x: 0.531
y: 0.653

Lips
x: 0.523
y: 0.773

Left cheek
x: 0.199
y: 0.550

Right cheek
x: 0.729
y: 0.531

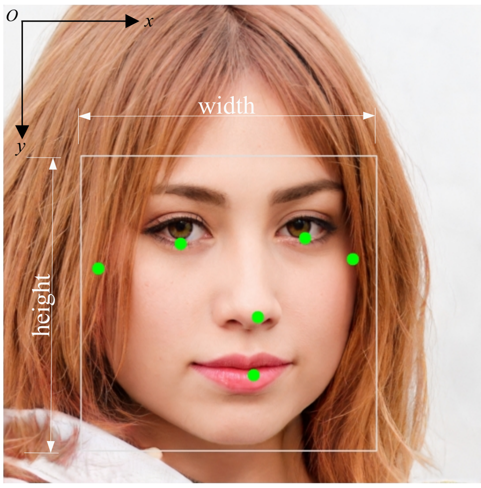


Figure 4.5: The point o on the top left corner of the image marks the coordinates $(0, 0)$. Each facial feature annotated with a green dot has a coordinate of (x, y) in the Cartesian system. The white bounding box around the face assures the presence of a face.

MediaPipe face mesh

MediaPipe face mesh also known as the attention mesh is a concurrent high-fidelity face reticulation prediction model that takes advantage of a compact statistical analysis technique (Procrustes analysis) to navigate an efficient and strong mobile logic [30]. Attention mesh utilizes neural networks and operates smoothly on the central processing unit (CPU) as well as GPU. In comparison to the former face predictors, the attention mesh delivers the highest number of facial landmarks, totaling 468 points. It extracts the face through the utilization of MediaPipe's super-realtime face detection tool. It then applies a feature extractor for mapping the face. Once done, it divides the model into several sub-models, of which one of them predicts and delivers all the face mesh landmarks.

Its diverse facial landmarking system enables accurate and detailed access to any desired facial region. Figure 4.6 illustrates attention mesh in action. As earlier demonstrated in MediaPipe face detection, each detected landmark had a coordinate of (x, y). In Attention mesh [29, 31], because of the mesh 3D topology architecture, the algorithm yields a depth parameter as well, signifying coordinates of (x, y, z). Concerning head rotation combined with back-and-forth head movements, the depth feature enhances accurate tracking of any RoI within the face bounding box. Attention mesh implements these coordinates jointly with the canonical face model to build a 3D reticulation topographic layer on top of the given image [44]. Because of its open-source architecture, it is possible to derive the mesh coordinates for other purposes as well. In applications and studies demanding high-fidelity RoI extraction, the attention mesh is a great match.

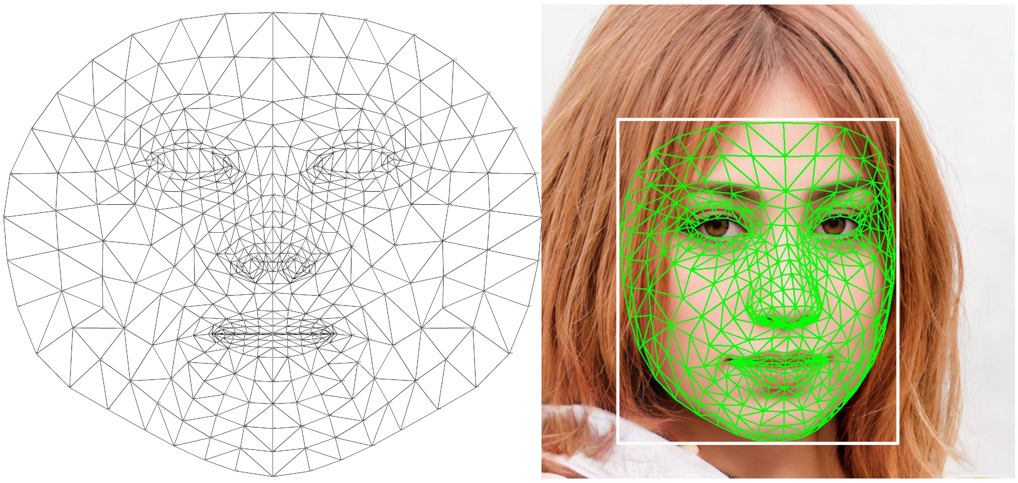


Figure 4.6: Canonical face model is a 3D topographic map of the facial landmarks (left figure) [45, 30, 44]. The intersection of contours is indexed with numbers ranging from 0 to 468, whereby each index holds a coordinate of (x, y, z). The white rectangle bounding box encapsulates the face mesh and the facial landmark points (right figure).

Chapter 5

Remote photoplethysmography

5.1 Preface

Remote photoplethysmography (rPPG) [6], also known as video-based photoplethysmography is a non-invasive physiological signs measurement technique that utilizes an inexpensive camera to detect the volumetric chan-ges of blood through tracking skin pixel intensity fluctuations. rPPG extends the boundaries of transillumination through the proper use of signal and image processing techniques. Benefiting from the conversion of images to three-dimensional digital signals, rPPG seeks to transform such signals to two-dimensional time series data suitable for signal processing and HR retrieval.

In recent years, rPPG has become a topic of interest among researchers and academia. With the expansion of CV-based solutions, the perspective of rPPG estimations has undergone a significant transformation [117, 98, 87, 123]. Understanding the mathematical theory behind rPPG is crucial, as it serves as the foundation for subsequent approaches and challenges in the field.

Without a sound comprehension of the mathematical principles, it would be difficult to develop accurate and reliable rPPG estimation approaches or to address any issues that may arise in its implementation.

5.2 The mathematical theory behind rPPG

In 2016, Wang et al. [67], proposed a mathematical model by which the algorithmic principles supporting rPPG were further elaborated. Their proposed concept, known as "*skin reflection model*" explains the signal received by the imaging device as a composition of specular and diffuse reflections off of the skin surface.

The skin reflection model, enabled by Shafer's dichromatic reflection model (DRM) [42] application for the revival of scenes geometry, assumes that the illumination source intensity fluctuates despite having a fixed spectral composition. It also claims that the intensity of light captured by the camera module is contingent upon the distance between the illuminating source and the skin surface, as well as the distance between the skin surface and the camera sensor, as illustrated in Figure 5.1.

The skin color that a camera observes is constantly shifting over time. The skin reflection model explains this variation through the change in three qualitative observations. Understating these rudimentary basics facilitates better approaches in rPPG estimation and simplifies countermeasures in handling noise-originating challenges.

The change in the illumination intensity is influenced by the specular and diffuse reflection (the absorption level of the skin tissue), as well as the change in the distance between the camera sensor, skin, and light source. The change in the motion variation is influenced by pulsatile body movements that originate from the heart pumping oxygenated blood through the arteries, accompanied by the body's natural movements.

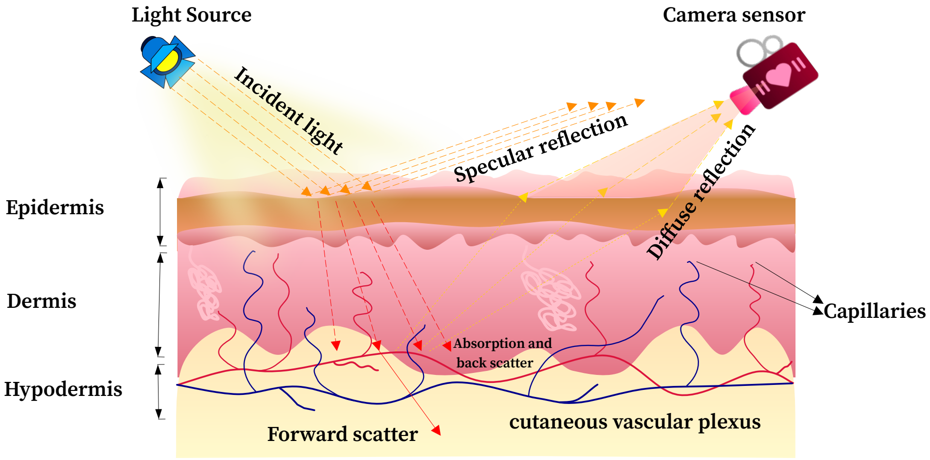


Figure 5.1: Skin reflection model

This model proves that the sum of these temporal variations corresponds to the brightness radiance degree observed by the camera sensor. Equation 5.1 best explains DRM and the skin reflection model [67].

$$\mathbf{C_k}(t) = I(t) \cdot \Big(\mathbf{v_s}(t) + \mathbf{v_d}(t)\Big) + \mathbf{v_n}(t) \tag{5.1}$$

where the RGB channel of the $k - th$ skin pixel is $\mathbf{C_k}(t)$. Brightness intensity is $I(t)$. Specular and diffuse reflection are, $\mathbf{v_s}(t)\ and\ \mathbf{v_d}(t)$ respectively. The camera quantization noise is $\mathbf{v_d}(t)$.

As Equation 5.1 shows [67], the specular and diffuse reflection components of the dichromatic reflection model are responsible for regulating the brightness intensity. Although specular reflection does not contain pulsatile data, it can still affect camera observations due to its influence on the geometric configuration between the camera module, skin surface, and illuminating source. This is caused by body movements, which can alter the readings.

$$\mathbf{v_s}(t) = \mathbf{u_s} \cdot \Big(s_0 + s(t)\Big) \tag{5.2}$$

where $\mathbf{u_s}$ is the Illumination band color direction parameter, s_0 is the stationary division, and $s(t)$ is the alternating division.

The presence of melanin amino acids in the skin tissue causes the diffuse reflection to experience an explicit chromaticity [67]. As Figure 5.1 illustrates, once the incident light hits the skin surface and enters the tissue, it scatters and undergoes some degree of absorption. The impact of volumetric fluctuations of blood over time influences the aforementioned processes as time-driven events and establishes $\mathbf{v_d}$ as a time-sensitive function.

$$\mathbf{v_d}(t) = \mathbf{u_d} \cdot d_0 + \mathbf{u_p} \cdot p(t) \tag{5.3}$$

where $\mathbf{u_d}$ is the skin tissue color direction parameter, d_0 is the stationary reflection power, and $\mathbf{u_p}$ is the RGB channel comparative pulsatile power, and $p(t)$ is pulse data.

The rearrangement of $\mathbf{C_k}(t)$ can be achieved by substituting Equation 5.2 and 5.3 into Equation 5.1, resulting in:

$$I(t) \cdot \Big(\mathbf{u_s} \cdot \big(s_0 + s(t)\big) + \mathbf{u_d} \cdot d_0 + \mathbf{u_p} \cdot p(t)\Big) + \mathbf{v_n}(t), \tag{5.4}$$

Furthermore, it is also possible to create a separate equation, specifically dedicated to stationary skin reflection, by integrating the stationary divisions of DRM components into one component.

$$\mathbf{u_c} \cdot c_0 = \mathbf{u_s} \cdot s_0 + \mathbf{u_d} \cdot d_0 \tag{5.5}$$

where $\mathbf{u_c}$ is the skin reflection color direction parameter and c_0 is the reflection intensity (power).

The substitution of Equation 5.5 in the rearranged $\mathbf{C_k}(\mathbf{t})$ Equation 5.6, dismantles $\mathbf{I}(\mathbf{t})$ into stationary I_0 and non-stationary $I_0 \cdot i(t)$ divisions [67].

$$\mathbf{C_k}(t) = I_0 \cdot (1 + i(t)) \cdot (\mathbf{u_c} \cdot c_0 + \mathbf{u_s} \cdot s(t) + \mathbf{u_p} \cdot p(t)) + \mathbf{v_n}(t). \tag{5.6}$$

As Figure 5.1 illustrates, the signal captured by the camera module is the combination of specular and diffuse reflections. Knowing that the specular reflection only conveys parallel light rays, it becomes evident that diffuse reflection must be the source of pulsatile information. The optical tendency of hemoglobin in absorbing light alters the integrity of the reflected light from the bed of tissue and acts like a pulsatile time-stamp on the collected image sequences.

rPPG seeks to extract this pulsatile data $p(t)$ from the k-th skin pixel observed by the camera module in the RGB channel. Equation 5.6 is the skin reflection model that expresses the reflection of individual skin pixels in a captured video as a time-changing event[1] in the red, green, and blue color channels.

In the 3rd section of "*Algorithmic Principles of Remote-PPG*", Wang et al. [67] demonstrate the utilization of the mathematical formulation of skin reflection model in evaluating the strength and effectiveness of other existing rPPG methods. For instance, assuming that the studies based on spatial averaging had enough camera arrays concentrated on the skin's desired area and the obtained RoI held enough number of pixels, the value of $\mathbf{v_n}(\mathbf{t})$ is so insignificant that it can be neglected.

By considering the visible skin pixels of RoI, it is possible to reformulate Equation 5.6 as follows:

$$\mathbf{C}(\mathbf{t}) \approx I_0 \cdot (1 + i(t)) \cdot (\mathbf{u_c} \cdot c_0 + \mathbf{u_s} \cdot s(t) + \mathbf{u_p} \cdot p(t)) \tag{5.7}$$

where $p(t)$, $s(t)$, and $i(t)$ are the AC components of the rPPG signal. These attributes correspond with the volumetric changes of blood between systolic and diastolic cardiac cycles. The amplitudes of the AC components are much smaller than those of the DC components, and as a result, their products are often negligible and can be disregarded. As demonstrated below, expanding the earlier equation into its subcomponents allows the separation of negligible parts.

$$\begin{gathered}\mathbf{C}(\mathbf{t}) = I_0 \cdot \mathbf{u_c} \cdot c_0 \ + \ I_0 \cdot \mathbf{u_s} \cdot s(t) \ + \ I_0 \cdot \mathbf{u_p} \cdot p(t) + \\ I_0 \cdot i(t) \cdot \mathbf{u_c} \cdot c_0 \ + \ I_0 \cdot i(t) \cdot \mathbf{u_s} \cdot s(t) \ + \ I_0 \cdot i(t) \cdot \mathbf{u_p} \cdot p(t)\end{gathered} \tag{5.8}$$

By simplifying the above equation, the products of $p(t) \cdot i(t)$ and $s(t) \cdot i(t)$ are neglected.

$$\mathbf{C}(\mathbf{t}) \approx I_0 \cdot \mathbf{u_c} \cdot c_0 \ + \ I_0 \cdot \mathbf{u_s} \cdot s(t) \ + \ I_0 \cdot \mathbf{u_p} \cdot p(t) \ + \ I_0 \cdot i(t) \cdot \mathbf{u_c} \cdot c_0 \tag{5.9}$$

[1] Considering that body movement and pulsatile blood are time-driven events.

The assumption of disconnectedness between skin pixel placements on the captured image and different color vectors (belonging to color spaces) certifies that $\mathbf{C}(\mathbf{t})$ the approximation is indeed the ultimate RoI spatial RGB average.

5.3 rPPG approaches

With respect to noise handling, rPPG approaches can be categorized into two distinct groups: DL-based solutions and non-DL solutions. This categorization of rPPG techniques serves to provide a useful framework for understanding the strategies and underlying principles employed to tackle noise in rPPG analysis. Over the years, rPPG approaches have been continuously assessed for their performance and robustness in addressing challenges such as illumination variation, motion artifacts, and noise [86, 87].

5.3.1 Non-DL rPPG solutions

Non-DL solutions are the cornerstone of rPPG studies, and their contributions commence from the early rPPG investigations and extend far into very recent studies.

Green [6], SSR [88], POS [67], PBV [56], CHROM [55], and LGI [74] are some of the most well-known and novel instances of methods belonging to these criteria. In the following, these models are briefly discussed.

Green

This method [6], is argued to be one of the simplest and most effective remote HR measurement approaches that benefit from the optical properties of blood in absorbing green wavelengths [89]. The evaluation of results from the RGB channels favored the green channel over the red and the blue channels due to fewer signal artifacts. Thus providing the strongest PPG signal responsible for hemoglobin oxygen absorption [90].

Green offers signal acquisition through spatial averaging of green channels over a region of interest. Green seeks the dominant frequency component by extracting the spectral contents of the facial region affiliated with the highest Oxyhemoglobin absorption.

In Equation 5.10, y(t) sums up the theory explaining the Green method.

$$x(t) = x_{RGB}(t) = x_r(t) + x_g(t) + x_b(t) \quad \Longrightarrow \quad y(t) = x_g(t) \tag{5.10}$$

where x(t) is the RGB temporal traces, $x_r(t)$ is the averaged red traces, $x_g(t)$ is the averaged green traces, and $x_b(t)$ is the averaged red traces.

Blind source separation

Blind source separation (BSS) is a computational data processing system that focuses on the recovery of overlooked signals from the linear mixture of several sources without any previous knowledge of the combination process [91].

BSS utilizes a statistical model for basing the assumption of having a combination of linear or nonlinear unidentified underlying variables, whereby the combination factors are not known. In signal processing, where measurements are obtained as a collection of parallel waves or time-indexed events, like the captured rPPG signals from the skin surface, BSS separates the mixture of unknown traces into distinguished signals [22].

In scenarios whereby unknown source signals $S(t) = (s_1(t), s_i(t), \ldots, s_n(t))^T$ are observed as the linear combination $x(t) = (x_1(t), x_j(t), \ldots, x_m(t))^T$. BSS bases its assumption on the fact that the sources are not correlated in a statistical sense and seeks to estimate the mixing matrix $A = [a_{ij}] \in \mathbb{R}^{m \times n}$ responsible for the production of $x(t)$, Equation 5.11.

$$x(t) = A \cdot s(t) \tag{5.11}$$

The unmixing matrix B, Equation 5.12 is the countermeasure that BSS takes to undo the combination of sources for the recovery of original signals $y(t)$, Equation 5.13, 5.14, [91].

$$B = [B_{ij}] \in \mathbb{R}^{n \times m} \tag{5.12}$$

$$y(t) = B \cdot x(t) \tag{5.13}$$

$$y(t) = (y_1(t), y_i(t), \ldots, y_n(t))^T \tag{5.14}$$

Figure 5.2 illustrates the role of a BSS model in the extraction of known signal sources from a linear mixture of them.

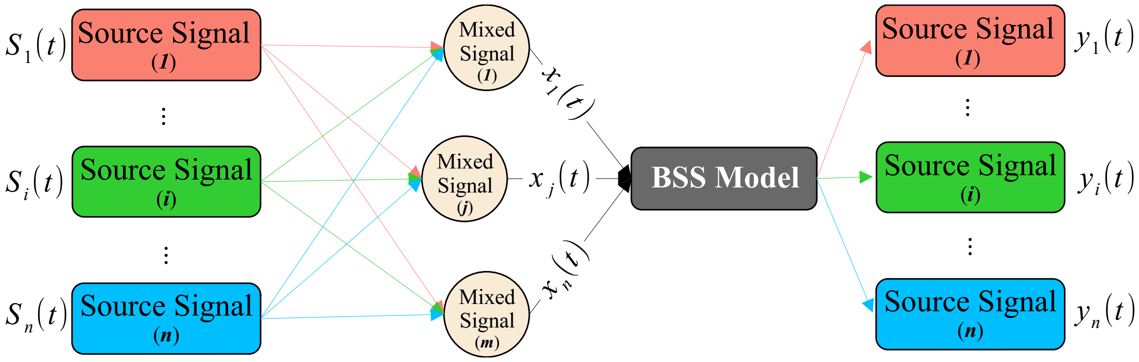


Figure 5.2: The red, green, and blue rectangles on the left depict the known source signals, the circles in between are an illustration of the linear combination of them, and the gray process box represents the BSS algorithm followed by separated source signals.

The suitable BSS approach

Considering the rPPG signal as a mixture of unknown source signals, one can utilize BSS to decompose the rPPG signal into n uncorrelated source signals to obtain HRV and HR values. Principal component analysis (PCA) and independent component analysis (ICA) are two of the most well-known BSS techniques.

ICA and PCA methods can be utilized to achieve this target [92], but the selection of a suitable BSS approach depends greatly on the rPPG studying environment. For instance, in rPPG measurements during physical activities, PCA is proven the superior rPPG method, but in a broader context, ICA is by far more resilient in non-fitness scenarios [67]. Both ICA and PCA are great BSS techniques, but when it comes to rPPG measurements, factors such as motion and examining environment should be considered in advance [67].

CHROM

Chrominance-Based rPPG addresses the shortcomings affiliated with PPG unforeseeable normalization misconceptions that originate from the specular reflection at the surface of the skin. As discussed earlier in 5.1, specular reflection does not carry any pulse signal, yet its union with the diffuse reflection plays a role in observing colors through the lens of a camera. This union, affected by the angles between the light source, skin, and the camera sensor, is not consistent over time. The movement of the participant facing the camera effectively unbalances the harmony of angles, hence disturbing rPPG measurements.

CHROM briefly eliminates such disturbances (stemming from the specular component) through the normalization of colour difference channels. Equation 5.15, expresses how CHROM creates two standardized and separate orthogonal chrominance signals X$_s$, Y$_s$ from the raw RGB traces x$_{rgb}(t)$ despite the colour of the illumination source [55].

$$\begin{aligned} &X_s(t) = 3 \cdot x_{r(norm)}(t) - 2 \cdot x_{g(norm)}(t) \\ &Y_s(t) = \tfrac{3}{2} \cdot x_{r(norm)} + x_{g(norm)} - \tfrac{3}{2} \cdot x_{b(norm)} \\ &\alpha = \tfrac{\sigma(X_s(t))}{\sigma(Y_s(t))} \ , \ \ S_{rPPG} = X_f(t) - \alpha Y_f(t) \end{aligned} \tag{5.15}$$

whereby,α is the fractional expression of computed standard deviation (σ) for X$_s$, Y$_s$ components, and S is the rPPG signal, X$_f(t)$ and Y$_f(t)$ are alternatives of band-passed filtered components X$_s(t)$ & Y$_s(t)$.

PBV

PBV, standing for pulse blood volume, is a color-based model that utilizes the comprehension of RGB color vectors in extracting pulse data. In a normalized RGB space, PBV exploits the differentiation of arterial blood spectra and skin tissues lacking blood, which subsequently results in observing the striking fluctuations along an explicit axis. PBV uses this information in building a motion-robust rPPG algorithm. With reference to performance, PBV accomplishes far better than ICA, PCA, and CHROM methods [56].

SSR

Spatial subspace rotation (SSR), also known as 2SR, is another novel rPPG algorithm that concentrates on the estimation of skin pixels'subspace temporal rotation in extracting pulse data from RGB videos.

In comparison to other methods, SSR is highly data-driven. For signals obtained from well-defined skin masks, SSR outperforms ICA, CHROM, and PBV in categories concerning skin tone, recovery of HR after training, illumination variation, and motion-artifacts. SSR eliminates the requirements associated with pulse-related and skin-tone of ICA and other techniques [88].

POS

POS, an abbreviation for "plane orthogonal to skin", is an alternative to CHROM that in essence hypothesizes the singularity of an illumination source accompanied by a consistent spectrum. In computing HR, POS considers factors associated with the physiological characteristics of skin reflection and optical parameters.

In terms of robustness in non-fitness investigations, POS outperforms almost all the model-based methods except one incident. SSR remains the sole model whose performance is on par with POS, mainly due to sharing similar features in their data-preprocessing algorithms.

With respect to investigations revolving around algorithmic principles of rPPG, POS offers the removal of specular reflections originating from the surface of the skin. The so-called "skin reflection model" is an extensive contribution of POS that explains rPPG methods through mathematical conventions [67].

LGI

Local group invariance (LGI) is a solid rPPG method specifically designed for HR estimation in the presence of rigid disturbances in facial videos. The study backing this model explicitly investigated factors such as head movement, realistic lighting scenarios, facial expression, exercise, and talking in a variety of uncontrolled settings including, indoor public gyms, indoor offices, and outdoors. LGI focuses on the rearrangement and a denser distribution of the blood volume signal energy in a vector space. In the process of evaluating the accuracy of LGI against other methods, the authors in [74]introduced a dedicated uncompressed database of indoor and outdoor videos collected under realistic illumination, activity, and facial expressions [74].

5.3.2 DL-based rPPG

In recent years, the advent of AI-powered solutions has gained considerable traction in rPPG studies. DL techniques can assist rPPG investigations in almost every step of the workflow. From ROI detection and selection to the introduction of end-to-end pipelines [98, 99], DL solutions enable concomitant measurement concerning the most common challenges facing rPPG algorithms, such as motion and illumination variations.

Although conventional methods have pioneered the majority of rPPG studies, too often they are complex and difficult to utilize. DL rPPG methods concentrate on the elimination of unnecessary steps or unionizing them as a one-stop solution.

In comparison to conventional approaches, these methods have faster computation response due to abolishing face detection, skin segmentation, the transformation of color space, face admission, face tracking, facial features segmentation, decomposition of source signals, and signal post-processing steps [98].

HR-CNN [100], STVEN+rPPGNet [99], Deep-rPPG [104], 3D CNN [101], PhysNet [103] , DeepPhys [98], and Meta-rPPG [102] are some acclaimed rPPG DL algorithms. Figure 5.3, sets up the workflow design of the STVEN+rPPGNet algorithm. As illustrated, STVEN+rPPGNet directly opposes the typical steps of conventional approaches, and establishes its superiority through the fusion of a video quality enhancement mechanism (STVEN) immediately followed by a rPPG signal recovery model (rPPGNet) [99].

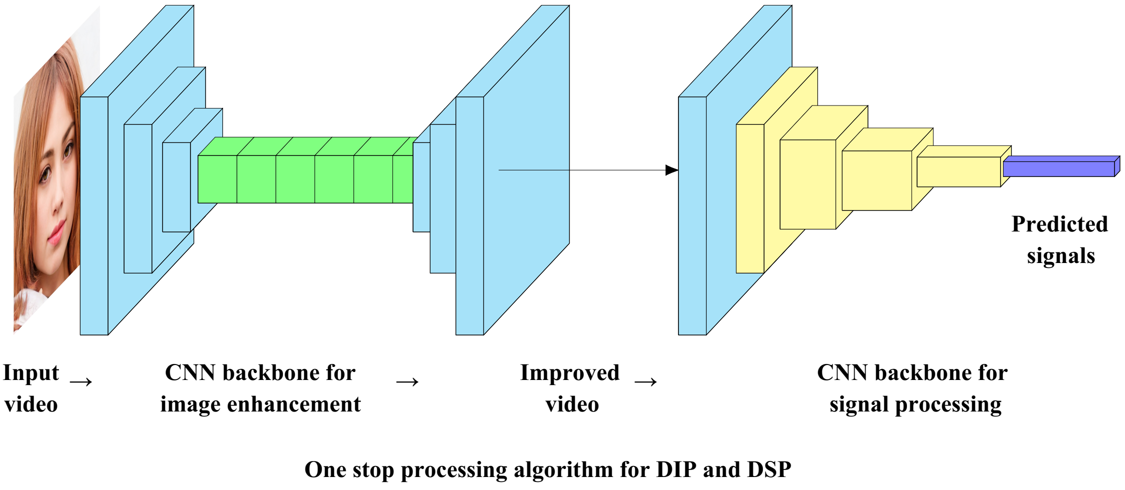

Predicted
signals
Input video →
CNN backbone for image enhancement →
Improved video →
CNN backbone for signal processing
One stop processing algorithm for DIP and DSP

Figure 5.3: A schematic overview of a 3D CNN layer that improves video quality (STVEN) combined with a network for predicting signals (rPPGNet) [99]. The synchronized harmony between the two models results in the prediction of viable HR signals.

Table 5.1, categorizes some former investigations respecting rPPG methods descending from DL solutions, based on topic, year, highlights, and trialed datasets.

rPPG studies descended from DL solutions				
Ref	Topic	Year	Highlights	Dataset
[102]	E2E* supervised learning (Meta-rPPG)	2017	Convolutional encoder transductive meta-learner, rPPG estimator and synthetic gradient generator	MAHNOB-HCI UBFC
[100]	E2E* system immune to illumination variations and motion artifacts (HR-CNN)	2018	A two-step CNN embracing raw signal extractor and HR evaluator	COHFACE PURE MAHNOB-HCI
[98]	Motion analysis attention mechanism (DeepPhys)	2018	Skin reflection model, CAN*, enhanced angular velocity and range of head rotation	MAHNOB-HCI
[105]	E2E* system in realistic scenarios (Deep PPG)	2019	large-scale dataset with a variety of realistic scenarios. Exploration of the dataset through various CNN* architectures.	PPG-DaLiA
[101]	Synthetic training (3D-CNN)	2019	Addressing the shortcomings of real-life data in biomedicine, building an automated artificial data friendly toolkit (Heart Track)	UBFC-RPPG
[99]	Video compression loss safety valve (STEVEN)	2019	rPPGNet	MAHNOB-HCI
[103]	Reconstruction of precise rPPG signals (PhysNet)	2019	DSTNs*, rPPG signal from raw facial video	MAHNOB-HCI OBF
[106]	HRV estimation	2020	DSTNs* for remote HR and HRV estimation from uncompressed facial videos	MAHNOB-HCI
[107]	SPAD* cameras (SPAD rPPG)	2020	DL and conventional DSP	Private dataset
[104]	Generalized rPPG skin ROI (Deep-rPPG)	2020	DSTNs*, light weight estimation network, wider applications	MAHNOB-HCI PURE COHFACE
[108]	Configurable rPPG pipeline (Face2PPG)	2022	Face stabilization, dynamic ROI, OMIT-QR* based RGB to rPPG	UBFC-RPPG MAHNOB-HCI PURE, LGI COHFACE

Table 5.1: An overview of some novel rPPG DL solutions. **SPAD**: single photon avalanche diode, **E2E**: end-to-end, **CAN**: convolutional attention network, **CNN**: convolutional neural network, **OMIT**: orthogonal matrix image transformation, **DSTNs**: Deep Spatio-Temporal neural Networks

5.4 Noise in rPPG

Ever since the introduction of rPPG, much of the efforts have been directed towards enhancing the efficiency of proposed methods by minimizing the effect of disturbances. The importance of noise in rPPG investigations has led to the development of various techniques and datasets. As previous sections pointed out, in rPPG studies, noise appears in different forms and originates from various sources. The following sections are a further elaboration of the early works concerning noise and rPPG estimation, along with the factors affecting the rPPG results.

5.4.1 Early works in rPPG noise reduction

The initial attempts in rPPG dates back to 1994 when Costa et al. [46] attempted to extract vital signs from the skin tone fluctuations through the utilization of two optical techniques. The first method was non-invasive and fully remote, which focused on speckle patterns to extract pulsatile signals. In contrast, the second method was an invasive yet remote technique which required the attachment of a small mirror on the skin. Even though this was a pioneering study, the findings were inadequate, inefficient, and lacked quantitative analysis concerning noise and related matters.

A decade later, in 2005, Puri et al. [47] pioneered the novel idea of monitoring computer operator sentiment state non-invasively and remotely by collecting sensory data via a compact thermal camera. The gathered sensory data from face temperature distribution were analyzed to extract the blood flow from the forehead frontal vessels corresponding to the peripheral nervous system (PNS). Although the focus of their study was on the correlation between stress and blood flow, in the years after, it inspired other studies [48, 49] and contributed to laying down the fundamental groundwork for future execution, research, and development of rPPG methods. One of the unspoken innovations of this study was the effective implementation of RoI selection and tracking to mitigate the noise caused by subject motion.

In 2006, Takano et al. [50], demonstrated the possibility of simultaneous extraction of RR and HR by employing a CCD camera to acquire a time-lapsed sequence of images from fourteen female participants. Their non-invasive and contact-free approach focused on the variations in median image illumination intensity of RoI occurring in 30 seconds timeframe. Using separate image and signal processing software, a set of tasks, including RoI brightness assessment, power spectrum analysis, and filtering were executed that resulted in the extraction of RR and HR. Among pioneering rPPG studies, their work was the only one that directly addressed the issue of illumination in realistic life-like scenarios. Their study openly discussed the illuminating conditions as the critical step towards quality image analysis in rPPG estimation. In addition, their study also considered the issue of sudden illumination changes and the fact that daily activities occur in such conditions rather than in lab-controlled illuminating settings.

In the same year, empowered by the implementation of fast Fourier transform (FFT) and thermal images, Garbey et al. [51] introduced another contactless and non-invasive HR measurement. To quantify HR, their method concentrated on the recovery of frequency components with elevated energy levels and applying an adaptive estimation on obtained mean FFT derived from body temperature fluctuations. Even though their study took to account the effect of noise introduced by the environment and unstable blood flow, it failed concerning the noise caused by involuntary muscular contraction, a motion artifact-related disturbance.

In the following year, a dedicated dual-wavelength contact-free rPPG mechanism was also developed [52]. The device was capable of measuring HR and S_pO^2 levels through a system composed of a metal-oxide camera and two alternating arrays of LEDs. The proposed device was trialed on a total of ten subjects in an uncontrolled setting with realistic illuminating conditions. To create an alternating illumination scenario, the LEDs emitted each wavelength of light. Despite the novelties, their approach required participants to remain motionless during signal acquisition and proved to be inefficient in handling the noise originating from the subject's motion.

Despite the developments, it was not until 2008 that Verkruysse et al. explored the idea of non-contact photoplethysmography imaging, through the utilization of an inexpensive digital camera in an ambient lighting scenario [6]. During their study, volunteers were requested to sit, stand, or hold a still position in front of a camera while having their faces recorded. The collected videos were then transferred to a computer for further image and signal processing measures. The implementation of a technique, called "spatial averaging" on the recorded videos led to the discovery of several viable frequencies in the RGB channels. In addition, they also discovered that the signal from the green channel is the strongest of all color channels. This discovery proved the correctness of hemoglobin optical properties regarding their peak absorption for green-blue wavelengths [53].

They even took the initiative of further investigating other possible body parts and noticed that the facial area is not the only region that holds signals. Given that the PPG signal was measured across other areas, it became evident that the signals retrieved from the face, explicitly from the forehead, were stronger. The implementation of spatial averaging prompted a notable reduction in the degree of camera quantization malfunction and amplified the signal-to-noise ratio (SNR). In addition to prior noise-related issues like motion, their study for the first time addressed homogeneous illumination as the root of shading artifacts and in response suggested the utilization of a higher-resolution camera.

In the following years, spatial averaging stimulated numerous studies and unleashed the true potential of colored videos in the revival and retrieval of biosignals [55, 56, 54, 49]. Owing to the growing interest in rPPG research, around the same period other innovative non-contact physiological signs measurement techniques started to appear that were visionary data independent and relied on microwave sensors for the quantification of HR and HRV [57, 58]. With the majority of the proposed methods being prone to motion artifacts, supervision of human operators, and extortionate running and maintenance costs, further developments were becoming more complex.

For the first time in 2010, in an attempt to address and cover the shortfalls of earlier works, Poh et al. [59] proposed a novel procedure that focused on the separation of mixed RGB signals obtained from facial colored video recordings into independent components. The implementation of a laptop built-in webcam, the execution of an automated face-tracking algorithm, and the use of the blind source separation (BSS) technique to untangle noise from unknown source signals on the RGB channels of facial videos were among the novelties of their methodology. The above-mentioned measures contributed to the reduction of manual supervision, automated the process, and lowered motion-illumination artifacts. The comparison of BVP readings from an FDA-approved sensor confirmed the accuracy of their proposal. Prompting the recognition of their work as the first accurate camera-based solution for contact-less HR surveillance capable of simultaneous measurements.

In the years after, [59] inspired numerous studies and laid down a comprehensive rPPG strategy [54, 60, 61, 62]. The incorporation of advanced high-definition camera module in hand-held devices propelled further expansion and development of rPPG methods. The initial techniques needed physical contact with the edge of the user's finger to extract a reliable HR [63, 64, 66].As smartphone sensors became more reliable, new innovative methods emerged that rely on facial readings obtained from the smartphone camera [65].

5.5 rPPG Influencing factors

As noted previously, disturbances induced by illumination variation and movement of subjects are two of the most frustrating challenges facing rPPG algorithms. In the following, some well-known and less-heard-of these troubling factors are discussed.

5.5.1 The effect of video compression algorithms

In 2017, McDuff et al. [109] conducted a systematic investigation respecting the effect of different video compression algorithms and codecs on the performance and accuracy of rPPG measurements. Regarding the effect of video bit-rate on the BVP signal-to-noise ratio (SNR), the preliminary evaluations on x264 and x265 video codecs delivered promising results and indicated a significant decrease in SNR between uncompressed raw videos and compressed ones. An increase in the rate of the video compression constant factor (CRF) equates to a linear decrease in SNR and bit rate, suggesting the effectiveness of x265 codec over x264.

5.5.2 Pigmentation of participants skin

The skin tone of a participant is a central factor in determining how strong the AC component of a signal is obtained [86]. This is because rPPG depends on the wavelength of the reflected light off the subject's skin. While light skin tones are normally more reflective, yielding stronger recorded signals, dark skin can challenge the rPPG algorithm in obtaining signals. One main challenge is the higher negative impact of any possible noise on the AC component of the RGB raw signal.

5.5.3 Digitization

One of the most underestimated factors concerning rPPG measurements lies within the data acquisition process. Although sampling and quantization are crucial in leveraging the versatility and robustness of digital systems, to a certain extent they do attenuate the image quality during the conversion process [110].

The noise resulting from digitization is rarely calculated. In addition, low sampling directly affects and causes aliasing, hence it should be taken into consideration concerning rPPG studies [111]. With the majority of algorithms and camera systems designed for none-rPPG purposes, it becomes evident to contemplate these attributes before and after initiating the data acquisition.

5.5.4 Scene illumination

One of the primary elements impacting rPPG is the scene illumination, because of the RGB pixel intensity is a fundamental factor in this technique. A substandard lighting setup forces extra work for algorithm optimization and demands more effort to compensate for the reduced accuracy. The quality of the obtained pulse signals heavily relies on the illumination variations of the environment. For instance, some studies have demonstrated that obtaining pulse signals under natural sunlight compromised the accuracy of the rPPG performance [112, 71].

Another example is the effect of screens (e.g., of a smartphone) that diminishes the quality of a recorded signal by introducing a significant amount of noise. Artificial light sources are diverse and range from simple light bulbs to infrared and fluorescent lamps [114]. Also, a combination of multiple artificial illumination modalities is proposed for certain scenarios [115]. Although artificial lighting methods provide more control over scene illumination and are significantly less prone to illumination variations, still an immaculate condition is not guaranteed [116].

5.5.5 Motion

Movements of the ROI (as a result of the minor movements of the face), or the camera introduces some noise to the RGB raw signal, known as the motion artifact. In such cases, the motion imposes illumination variation on the ROI, causing instability in the amplitude of the RGB raw signal. Additionally, the ROI detection algorithm cannot find the measurement area, leading to a significant loss of information in the RGB raw signal.

5.5.6 Camera parameters

The primary device in a rPPG study is the video recording camera. Resolution and video sampling rate are the two most overlooked attributes of a camera in the execution of a rPPG experiment. These parameters ultimately affect the integrity of the RGB raw signal. Camera resolution determines the number of pixels a camera is capable to record. In DSP, more pixels translate to more information, providing higher possible quality of raw RGB signal. The Video sampling rate, expressed in fps and Hz, is the frequency at which the video signal is recorded. In other words, the number of still images (frames) captured by the camera in the unit of time (seconds) determines the video sampling rate. The reported sampling rate of the majority of rPPG studies ranges from 20 fps achieved by low-cost cameras to 60 fps by premium cameras.

5.6 rPPG datasets

Studies conducted on rPPG require a dataset to evaluate their methods on. While many publications in the literature have employed a private/custom dataset, utilization of publicly available datasets has also become a routine procedure. Another common practice in the rPPG private/custom data gathering is to use an RGB camera placed at a certain distance (normally one meter) from subjects. Depending on the aim of the studies, certain tuning and adjustments are applied to lighting conditions, indoor/outdoor capturing settings, and motion involvement.

Although establishing a brand-new private dataset enhances the evaluation quality, a few challenges hinder their application. A fundamental barrier to the use of private datasets is the ethical issues concerning sensitive data gathering. Plus, even with an appropriately followed ethical procedure, the challenge of finding participants for the data collection exists. Conclusively, in spite of the many benefits of custom data collection, the use of publicly available data saves time and resources in rPPG studies.

Aside from the properties and approaches concerning the collection of the data, the conditions in which the data is a structured matter as well. In that sense, rPPG datasets are categorized into two subclasses of lab-controlled and realistic databases. For instance, MAHNOB-HCI and COHFACE datasets are lab-controlled, meaning that the effect of illumination changes and movement disturbances are reduced for minimum possible noise.

Such datasets are suitable for fine-tuning and evaluating new rPPG solutions. Contrastingly, realistic datasets are prone to more noise, yet provide more insight into real-life scenarios such as driving and physical activities. In the following, settings and features of popular rPPG datasets as well as the common process of preparing a private/custom dataset are briefly discussed. Ultimately, by the end of this section, the criteria by which a realistic dataset was preferred is further elaborated.

5.6.1 MAHNOB-HCI

MAHNOB-HCI dataset is a series of video recordings collected from 30 participants (13 males and 17 females) in the age range of 19-40 [68, 70]. MAHNOB-HCI method is a multimodality dataset in which the data comes from face video, audio, eye gaze and nervous system signals. The videos were captured using 6 cameras with dimensions of 780×580 pixels and a frame rate of 60fps [69].

Physiological signals included EEG, ECG, respiration pattern as well as galvanic response and temperature of the facial skin. Although this dataset is a proper instrument for emotion recognition and implicit tagging rPPG studies, but because of the extreme level of video compression, the videos in this dataset sustain a degree of noise artifacts [68, 73].

5.6.2 COHFACE

The COHFACE dataset comprises 160 video recordings obtained from 40 participants, 12 of whom are female and 28 male, with an age distribution centered around +35 years [71, 72]. Because of using two different lighting setups to simulate realistic illumination conditions, the videos captured in this dataset appear more natural, compared to the MAHNOB-HCI dataset. The length of videos in this dataset is about 1 minute and the video dimensions are 640 × 480 captured at 20fps using a Logitech HD webcam C525.

The physiological metrics aimed in this data preparation include respiration patterns and blood volume pulse measurements. The major setback of this dataset is the limited skin tone variation of subjects [71]. Another drawback of this dataset is that it had undergone heavy compression, which inevitably led to the introduction of noise artifacts [123, 73].

5.6.3 UBFC-RPPG

Also known as the UBFC dataset, the idea behind this data collection approach revolves around the implementation of a game scenario to induce a raise in the heat rate [76, 77]. A mathematical game with a time limit was supposed to increase the heat rate of the participants during the video acquisition. Hence, a pulse oximeter was used for ground truth information acquisition.

43 two-minute videos recorded by a Logitech C920 HD pro camera with a dimension of 640 × 480 at a 30 fps frame rate are included in this dataset. Although the idea behind this data collection method is innovative, the lack of varying lighting conditions can be considered a drawback [76].

5.6.4 PURE

The PURE dataset was created based on a controlled motion variation and artifact introduction while head movements are captured [78, 79]. This dataset provides information on conditions where illumination inconsistencies are associated with the non-contact pulse measurement from videos. A series of 60 uncompressed videos with about 1-minute duration were acquired from 2 males, and 8 females. The captured Videos have a dimension of 640 × 480 at 30fps frame and the camera was 1 meter away from the participants.

The recordings were done in six different setups including steady mode, talking condition, minimum rotation, intermediate rotation, slow translation, normal translation, and speedy translation. Accompanied by the camera, was a CMS50E pulse oximeter employed for ground truth data acquisition (for HR and S_pO^2) [78].

5.6.5 OBF

Oulu Bio-Face (OBF) is another rPPG dataset that aims to fill the gaps with earlier proposed databases [80]. OBF encloses a diverse quantity of facial videos accompanied by instantaneous ground truth physiological data. The OBF dataset stands apart from other datasets due to its inclusion of data from patients with Atrial Fibrillation (AF), whereas other datasets primarily consist of data from healthy participants.

In OBF, data acquisition occurs in two separate 5-minute sessions (resting and exercising respectively), of which a 5 minutes exercise break (climbing stairs) was included in between. The use of RGB and NIRV camera sensors has resulted in having two types of sampling ratio and resolution, 1920×2080 at 60 fps and 640 × 480 at 30 fps respectively.

5.6.6 LGI-PPGI-DB

LGI-PPGI-DB as a multi-session dataset solves some problems associated with the single-scenario data collection methods [74, 75]. The main characteristic of this dataset is the administration of four different conditions for video acquisition: 1) resting mode with static lighting and no motion, 2) moving mode with static lighting and motion in the head, 3) exercise mode using a bike ergometer and finally 4) conversation mode with head motion and natural fluctuating lighting condition.

A total of 25 participants of different ethnicity (20 males and 5 females) between 25–42 years of age were included in the data collection process. Videos with 1–5 minutes duration were obtained by a Logitech HD C270 webcam at a 25fps frame rate. Plus, the ground truth data was achieved by employing a PPG signal from a CMS50E device [74].

Dataset structure

The LGI rPPG dataset in total has 200 minutes of uncompressed video data. At the time of conducting this study, only four candidate files were available to access. Table 5.2

details the ID number, recording scenario, size, and length of the videos in this dataset. Tree chart 5.4 illustrates the structure of the original dataset [75].

Dataset content				
Name	Scenario	Duration	Size	Total duration
Id 1	Gym	06:37 sec	9,167 Mb	10 min and 23 sec
	Resting	00:59 sec	1,692 Mb	
	Rotation	1:14 sec	1,717 Mb	
	Talking	1:33 sec	2,152 Mb	
Id 2	Gym	05:22 sec	7,421 Mb	8 min and 35 sec
	Resting	01:06 sec	1,521 Mb	
	Rotation	1:05 sec	1,504 Mb	
	Talking	1:12 sec	1,676 Mb	
Id 3	Gym	03:50 sec	5,314 Mb	7 min and 43 sec
	Resting	01:12 sec	1,681 Mb	
	Rotation	1:20 sec	1,861 Mb	
	Talking	1:21 sec	1,877 Mb	
Id 4	Gym	05:04 sec	7,008 Mb	8 min and 44 sec
	Resting	01:11 sec	1,644 Mb	
	Rotation	1:10 sec	1,632 Mb	
	Talking	1:19 sec	1,822 Mb	

Table 5.2: A total of four scenarios equating to 49.712 GB or 35 minutes and 25 seconds of uncompressed video data captured at 25 fps in AVI codec with auto-exposure on.

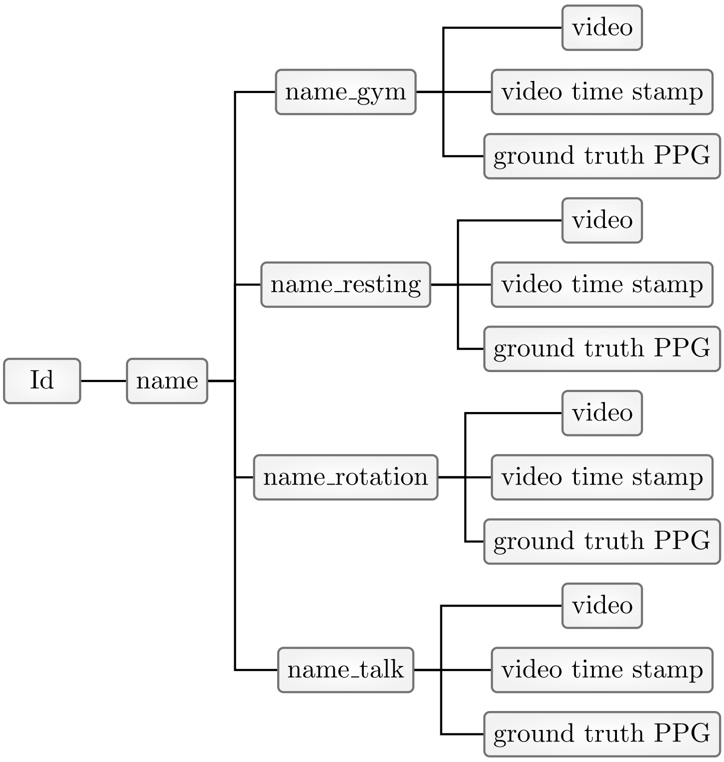


Figure 5.4: The directory structure of the LGI-PPGI-Face-Video-Database

Table 5.3, summarizes the rPPG datasets based on name, sampling ratio, number of videos and participants, the length of videos, and the criteria to which they belong.

rPPG Datasets						
Ref	Dataset	Sampling Rate (Hz)	Statistical Population	Collection of Videos	Duration	Scenario
[71]	COHFACE	20	40	160	1 Min	Controlled
[74]	LGI-PPGI-DB	25	25	100	1 Min	Realistic
[76]	UBFC-RPPG	30	50	50	1 Min	Controlled
[81]	HKBU-MARs V1	25	8	120	12 Sec	Controlled
[82]	HKBU-MARs V2	20-30-50	12	1008	10 Sec	Controlled
[83]	VIPL-HR	30	107	2,378(VIS) 752(NIRV)	30 Sec	Hybrid
[68]	MAHNOB-HCI	60	30	120	40 Min	Controlled
[84]	OSF	25	3	160	Not Specified	Controlled, 7 LCs
[85]	MMSE-HR	25	140	560	20 Sec	Controlled
[78]	PURE	30	10	60	1 Min	Controlled
[80]	OBF	60-30	106	212 (RGB-NIRV)	5 Min	Realistic

Table 5.3: VIS: visible light videos, NIRV: near-infrared video, Hybrid: a bridging dataset between controlled and realistic scenarios, HKBU-MARs: The HKBU 3D Mask Attack with Real World Variations, LCs: lighting Conditions

Chapter 6

Materials and Methods

6.1 Preface

In this chapter, we present a modular solution for remote, video-based HR extraction in noisy environments using rPPG. Furthermore, we will apply our introduced approach step-by-step on a realistic rPPG dataset and measure HR. Our goal is to offer a countermeasure solution concerning the effect of noise during rPPG signal acquisition. We will also evaluate the effect of noise on the computed HR and compare the accuracy of predicted HR against reference data.

We evaluate the proposed rPPG approach by using the LGI-PPGI realistic dataset that encompasses both indoor and outdoor environments and subjects with different skin tones. The dataset includes a range of noise levels, such as varying lighting conditions and motion artifacts, to simulate real-world scenarios.

The presence of noise and motion artifacts in the acquired rPPG signals pose a significant challenge for accurate HR computation. In order to address this issue, we utilize and apply three different face detectors, namely OpenCV's Haarcascade, Dlib, and MediaPipe to detect the face region in the acquired noisy videos.

We then convert the detected RoIs into workable time-series signals for further processing and analysis. Considering that the face detection methods operate distinctively from another, using and comparing them will enable us to see their effect as the signal extraction probe in rPPG measurement in noisy scenarios.

6.2 Proposed architecture

Our proposed method is composed of six different modules and initiates with data preprocessing, followed by RoI, signal extraction, signal preparation, signal processing, and HR extraction. Each module covers a handful of measures and techniques. Diagram 6.1 articulates an overview of the workflow and the placement of each module in the planned architecture.

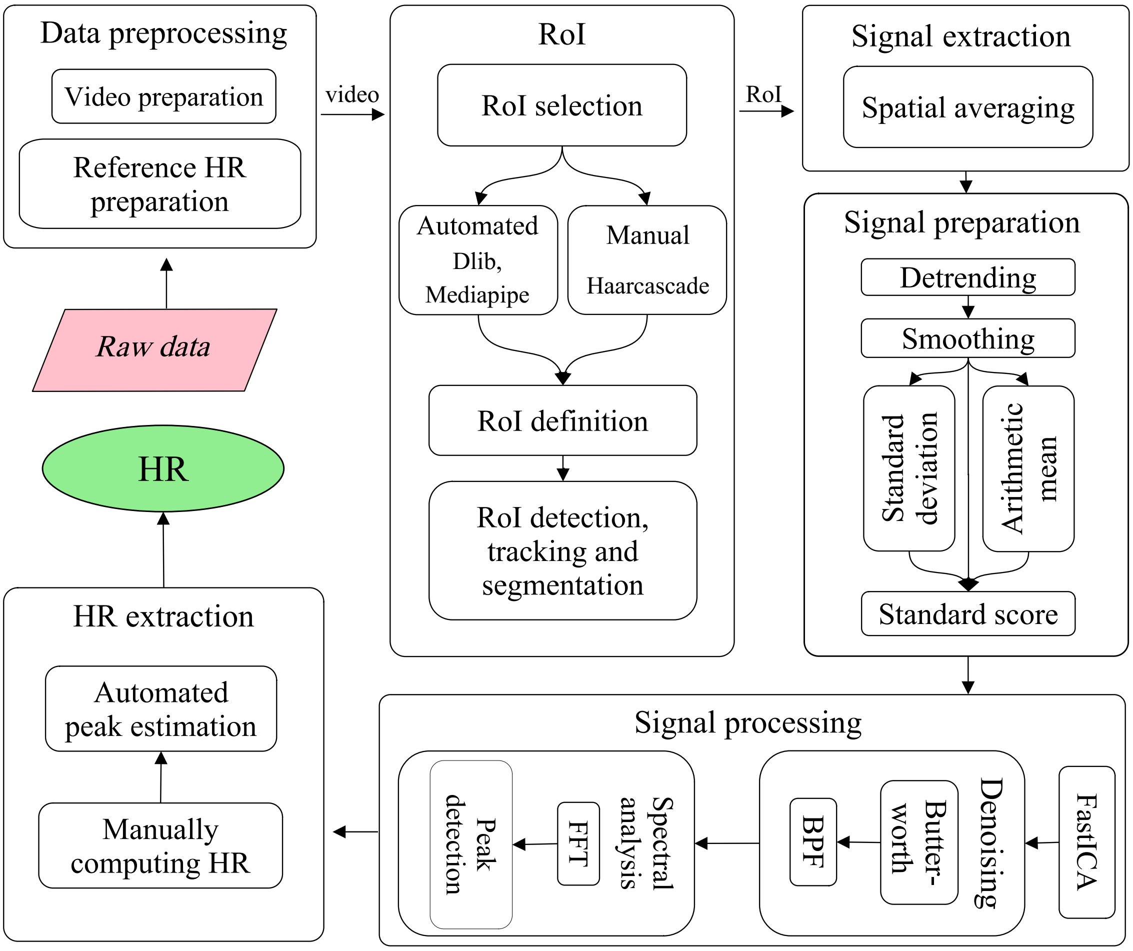


Figure 6.1: The schematic overview of the Workflow.

With respect to the experiment design, our proposed workflow seeks to fulfil and deliver solid results concerning the ROs of this thesis.

6.3 Data preprocessing

Data preprocessing as a whole enables managing the relevancy and homogeneity of the raw collected data and eases the process of data classification[22].

6.4 Video preparation

In this module, we load the video data into the Python programming environment with the help of the open source computer vision library (OpenCV).

6.4.1 Color model conversion

This process ensures that, the color space of our video data is suitable for OpenCV operating environment, as well as selected face detection approaches in the RoI module.

Considering the fact that we are evaluating three face detection approaches within our modular architecture in parallel (not simultaneous), as a precautionary measure, at this step we make sure that each algorithm receives the video data in its required format.

OpenCV color space

OpenCV, by default, operates within the BGR color space. Our dataset, however, contains videos in the RGB color space, thereby requiring the conversion of these videos to the BGR space upon loading them into the OpenCV environment.

Face detectors color space

The choice of color space for loaded video data not only affects the operation of OpenCV, but also has an impact on the performance of specific face detection algorithms.

For instance, the Haarcascade classifier requires the data to be in grayscale format, thus necessitating the conversion of the BGR color space to grayscale. In contrast, the Dlib face detection algorithm is flexible in its color space requirements and can operate with video data in most color spaces.

In the following, we will further elaborate the various color model conversion approaches that are necessary for the operation of our RoI selection module.

Grayscale conversion

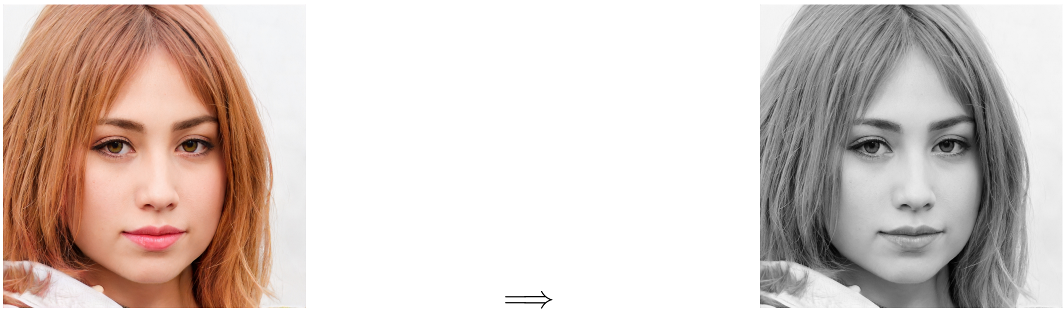
The conversion of a 3-channel n-dimensional array into a single-channel n-dimensional array is called a grayscale conversion. The conversion of an image (or video frame) to grayscale reduces the computational burden on both hardware and software by simplifying the mathematical operations involved.

As a DIP method, we can utilize grayscale conversion to implement convoluted processes on colored images (video data) in a shorter time span. With regard to the bulky colored videos of our dataset, the inclusion of this step guarantees a stable video processing.

Equation 6.1, known as the weighted average Y, is a derivative of the famous luminance conversion that balances the pixel values across all the primary color channels[12].

$$Y = 0.299 \times R + 0.587 \times G + 0.114 \times B \tag{6.1}$$

where Y represents the luminance or brightness of the image. The coefficients behind R, G, and B are used to calculate the luminance (Y) value from the red (R), green (G), and blue (B) color values of an image. The coefficients are based on the relative sensitivities of the human eye to different colors. As Figure 6.2 shows, we can convert an image from RGB color space to the grayscale if only we apply the above equation to the matrix of colored pixels 3.7.



(a) Original input RGB image

(b) Grayscale output image

Figure 6.2: In some older face detection algorithms, like Viola-Jones, the pre-trained classifier like Haarcascade only works with the grayscale images.

By converting the RGB video data to grayscale through the grayscale conversion process, we can ensure the proper functioning of the Haarcascade and Dlib face detectors.

RGB ⇔ BGR conversion

The RGB ⇔ BGR color space conversions are based on a permutation matrix that interchanges the red and blue columns of the RGB/BGR matrices and creates a new matrix.

$$\underbrace{\begin{bmatrix} r & g & b \end{bmatrix}}_{RGB} \times \underbrace{\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}}_{Permutation-matrix} = \underbrace{\begin{bmatrix} b & g & r \end{bmatrix}}_{BGR} \implies$$

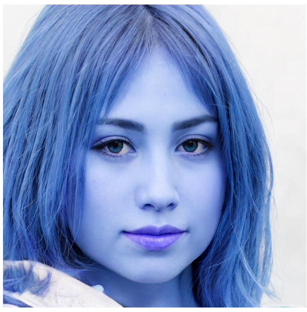


Figure 6.3: BGR

$$\underbrace{\begin{bmatrix} b & g & r \end{bmatrix}}_{BGR} \times \underbrace{\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}}_{Permutation-matrix} = \underbrace{\begin{bmatrix} r & g & b \end{bmatrix}}_{RGB} \quad \Longrightarrow$$

Figure 6.4: RGB

Unlike older face-detection algorithms, MediaPipe face detection only functions with image data of RGB format. As noted, OpenCV requires RGB to BGR conversion, but BGR is not a suitable color space for MediaPipe. To address this issue, we utilize the permutation matrix corresponding to the conversion of BGR to RGB over the video data and convert the BGR videos to RGB. Then, we feed in the RGB videos to MediaPipe and upon the completion of MediaPipe face detection, we convert the video data back to BGR.

6.5 Region of Interest

As its name implies, region of interest (RoI) refers to the desired area containing a certain aspect or attribute. In rPPG estimation, RoI often refers to territory inside the frames of a rapidly moving or still picture. Considering that we plan to extract HR data from face, the next step is to select a RoI. Given that the chosen face detectors have varying capability with respect to detection and tracking of RoI, in the following, we further elaborate RoI selection and its existing approaches.

6.5.1 RoI selection

Perhaps one of the most important steps in rPPG measurements corresponds to the quality of the collected signal. High-quality data extraction is a critical step towards any data analysis. In rPPG studies where data is acquired through non-contact means of measurement, RoI is the ultimate source for the extraction of raw physiological data.

In this study, we classify RoI selection approaches, as automated and manual. This classification stems from the face detection capability in detecting the desired facial landmarks. Meaning that, if a face detection algorithm offers direct access to facial landmarks, it falls under the criteria of automated. However, if it does not provide such feature, it is considered as manual.

Automated

This segment benefits from robust and instantaneous pattern recognition classifiers in detecting, tracking and localizing facial features. Supported by the development of ML solutions dedicated to CV, countless facial recognition classifiers have been developed [122]. With the rise of intuitive and high-quality open-sourced algorithms, toolkits such as Dlib [35] and MediaPipe [28, 29, 30, 31] have simplified RoI selection. While some studies opt for these toolkits, other solutions also exist [124, 126, 125].

Manual

Unlike the automated-based solutions, another alternative is to manually select the RoI. Any video is a composition of multiple frames having a width×height dimension in the Euclidean space, propagating through the vector of time. Once an RoI is selected from the video frame, it is possible to address or segment it using RoI pixel coordinates.

As Figure 6.1 depicts, in the RoI module, the preprocessed videos are loaded into designated RoI selection approaches. While Dlib and MediaPipe offer automated access to facial landmarks and RoIs, for the Haarcascade we manually separate and extract RoI.

6.5.2 RoI definition

An appropriate RoI that offers a strong and consistent BVP signal is critical towards HR estimation. A well-defined RoI contributes to the extraction of maximum pulsatile data, and by doing so increases the likelihood of accurate measurements. Regardless of what RoI selection we are opting for, to maximize the strength of the extracted signals, we need to define a facial RoI that promises sufficient blood circulation.

It's worth noting that the face is a key RoI that allows for consistent extraction of the BVP signal, however, not all facial features are necessary for this process. Because of the natural movements and disturbances within the face, such as blinking, nasal vestibule movement, and frowning, it is important to carefully select the specific facial features to use for quality signal retrieval [87].

Forehead as the RoI

To select a suitable RoI we preliminarily conducted a number of investigations on different facial attributes. We then analyzed obtained results from various facial features, including the forehead, glabella, root of the nose, infraorbital triangle, nasolabial furrow, eyes, and cheeks. The analysis of results showed us that the forehead contains the strongest signal of all regions.

The noise generated by subject motion, the motion of facial features, and illumination variations imbalances on subjects faces have led us to believe that forehead has a good potential to be the site of raw signal extraction.

To prevent any biased assumptions, we also reviewed similar studies concerning the suitability of RoI and found that the forehead has an established record for reflecting the activity of the heart muscle[127].

6.5.3 RoI detection, tracking, and segmentation

In rPPG studies, RoI detection, tracking, and segmentation are essential in the signal processing pipeline, and their existence relies on the features that a face detector offers. These steps enable us to extract raw signals from the available data and assess the performance of face detectors in noisy conditions. Having knowledge about these steps assists us in precisely locating and extracting the necessary information from the dataset at hand and computing the HR. In the following, we will further expand the role of these steps in our pipeline.

RoI detection

RoI detection is the process of identifying the location of the face in the acquired images. In our pipeline, we utilize Viola-Jones, Dlib, and MediaPipe face detection classifiers to detect the subject face in every frame. The use of multiple face detectors empowers us to better evaluate the role of face detectors in challenging lighting and pose conditions in video-based HR extraction in noisy environments.

RoI tracking

In the utilized dataset, due to changes in facial expression and head movements, the subject's face region either moves or changes its appearance. Therefore, we have to consider a mechanism that facilitates the extraction of data from the same location of the face region in every frame. In our proposed architecture, the module representing this mechanism is known as RoI tracking.

By default, the selected face detectors do offer a tracking feature up to a number of faces. Our videos on the other are captured in an uncontrolled manner, meaning that people other than our subjects are present throughout some scenarios. During our preliminarily investigations, we noticed that in such cases, the face detector rapidly jumps from the subject face to other faces present in each frame, hence provoking the integrity of the data extraction. To overcome this issue, we fine-tune the parameters of our face detectors in a way that the face detectors only detect one face at a time. To further reinforce it, we additionally introduce a term that only admits the face that is closest to the camera.

RoI segmentation

Considering that we are going to select the forehead as the ultimate site of signal extraction, we need to segment the forehead from the rest of the face. In rPPG studies, RoI segmentation is the process of separating the skin regions of the face from other non-skin regions such as hair, eyes, and background. RoI segmentation is necessary because non-skin regions can introduce noise and artifacts into the rPPG signals, which can negatively impact the accuracy of the HR computation.

Face detection and forehead extraction

As stated earlier, we plan to evaluate the suitability and resilience of three different face detectors, namely, Haarcascade, Dlib, and MediaPipe in the extraction of rPPG signal in noisy scenarios.

The above-mentioned face detectors by default do not offer facial features segmentation, but given that they have open-source algorithms [32, 35, 28], it is possible to utilize secondary solutions that lead to the customization of their source code, therefore adding these features.

In the following sections 6.5.3, 6.5.3, and 6.5.3 we further elaborate and visualize our approach concerning the detection, tracking, and segmentation of forehead within each of the selected face detectors.

In our evaluation, we we run our pipeline in three separate sessions, each utilizing one of the selected face detectors. The face detection approaches do not operate concurrently, but for evaluation purposes we do execute them in parallel.

HaarCascade (Manual)

This method commences with feeding the pre-trained Haarcascade classifier to the algorithm. Because the pre-trained classifier was initially trained with grayscale images, utilization of this classifier demands the conversion of video color space from BGR to gray (see Equation 6.1). Hence, we initially perform a color space conversion and then feed the video to our face detector. Upon the detection of the face, the algorithm generates coordinates corresponding to the top, bottom, left, and right corners of the face. We will then employ these coordinates to separate the face from the rest of the image and bound it to a box-like frame.

As mentioned in earlier chapters, the cascade classifier does not offer a facial landmark-pointing system, hence some creativity is needed to separate the defined RoI from the rest of the face. Figure 6.5 illustrates the mechanism by which the coordinates of the forehead are found and used for RoI extraction.

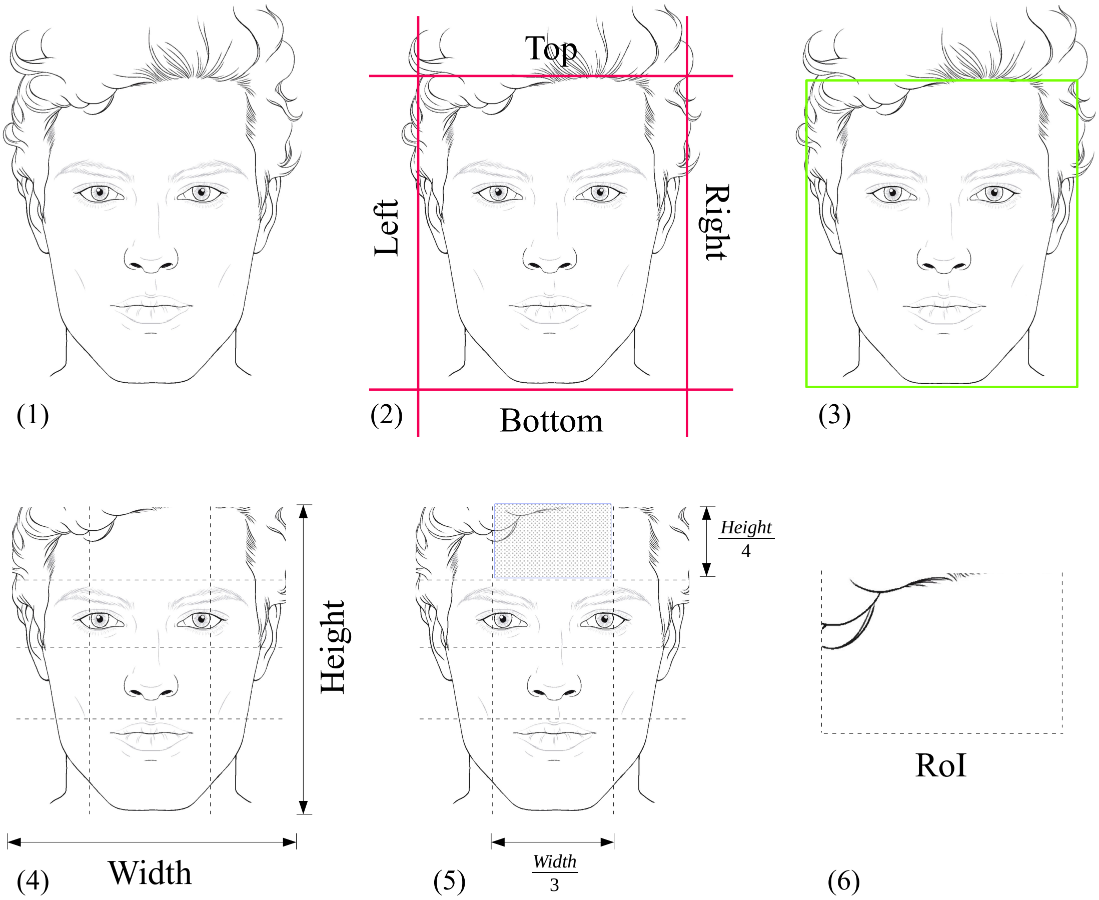


Figure 6.5: The 1st image illustrates a person present in a video frame. The 2nd image portrays the face detection classifier in action, as shown, once a face is detected, the classifier yields the coordinates corresponding to the four corners of the face. The 3rd image illustrates the bounding box containing a face. The 4th image divides the face horizontally into 4 and vertically into 3 regions. The 5th image hosting the dashed area highlights the defined RoI coordinates by their beginning and ending points, vertically and horizontally. The 6th image hosts the extracted RoI from the original frame.

Dlib-HoG (Automated)

This method starts with loading the Dlib frontal face detector and feeding the weights of 81-face-landmark points to the shape predictor. Figure 6.6 illustrates the procedure by which Dlib detects the face and how we use the facial landmark points to access the coordinates of RoI and segment it.

Dlib pre-trained model also works with color images and OpenCV's default BGR color space format. Having said that, to lower the computational load of 3 color channels, we implement a grayscale conversion and feed the grayscale video to Dlib face detector.

Once a face is detected, the program generates coordinates corresponding to the top, bottom, left, and right corners of it. In addition, it simultaneously applies the shape predictor weights on the detected face and indexes the facial regions, Figure (3) 6.6.We then employ the landmark points as a guide for defining and creating a hoovering bounding box around the forehead and segment the forehead from the rest of the face, Figure (4,5,6) 6.6.

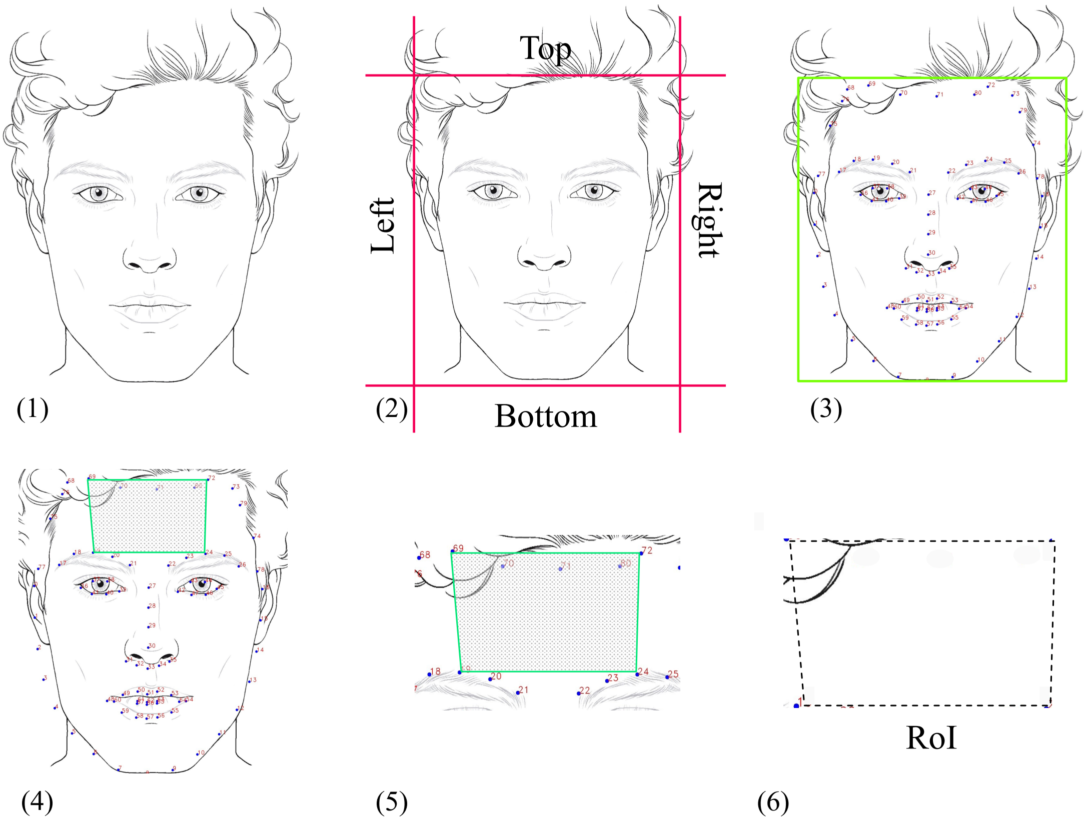


Figure 6.6: The 1^{st} image illustrates a person present in a video frame. The 2^{nd} image portrays Dlib-HoG face detector in action. As shown, once a face is detected, the face detector yields the coordinates corresponding to the four corners of the face. The 3^{rd} image illustrates the Dlib shape predictor utilized with Dlib pre-trained 81 facial landmark points in action. As demonstrated, the face is overlaid with facial landmark points and the bounding box containing it. The 4^{th} image highlights the dashed area containing the forehead. The 5^{th} image hosting the dashed area signifies the indexes of facial landmark points (69, 72, 19, 24). The 6^{th} image hosts the extracted RoI from the original frame.

Attention Mesh (Automated)

This method starts with fetching the MediaPipe face detector module. Unlike previous methods, MediaPipe does not require loading weights beforehand. But Instead, it demands loading the face mesh and drawing functions if manipulations beyond face detection are required. MediaPipe also does not require color space conversion, and by default functions on videos and images of RGB color models. After loading the program with the visual data, the face detector detects the face and retrieves the coordinates corresponding to the top, bottom, left, and right sides of it. The face mesh function then utilizes the face coordinates to detect the facial features. Once facial features are detected, the drawing function tessellates the facial landmarks and yields a canonical face model. The indices of the canonical face model (the facial landmark points) bear coordinates of (x, y, z), corresponding to 468 regions of the face. We then employ these indices to access the forehead region and extract it from the rest of the frame. Figure 6.7 illustrates the above-mentioned stages in consecutive order.

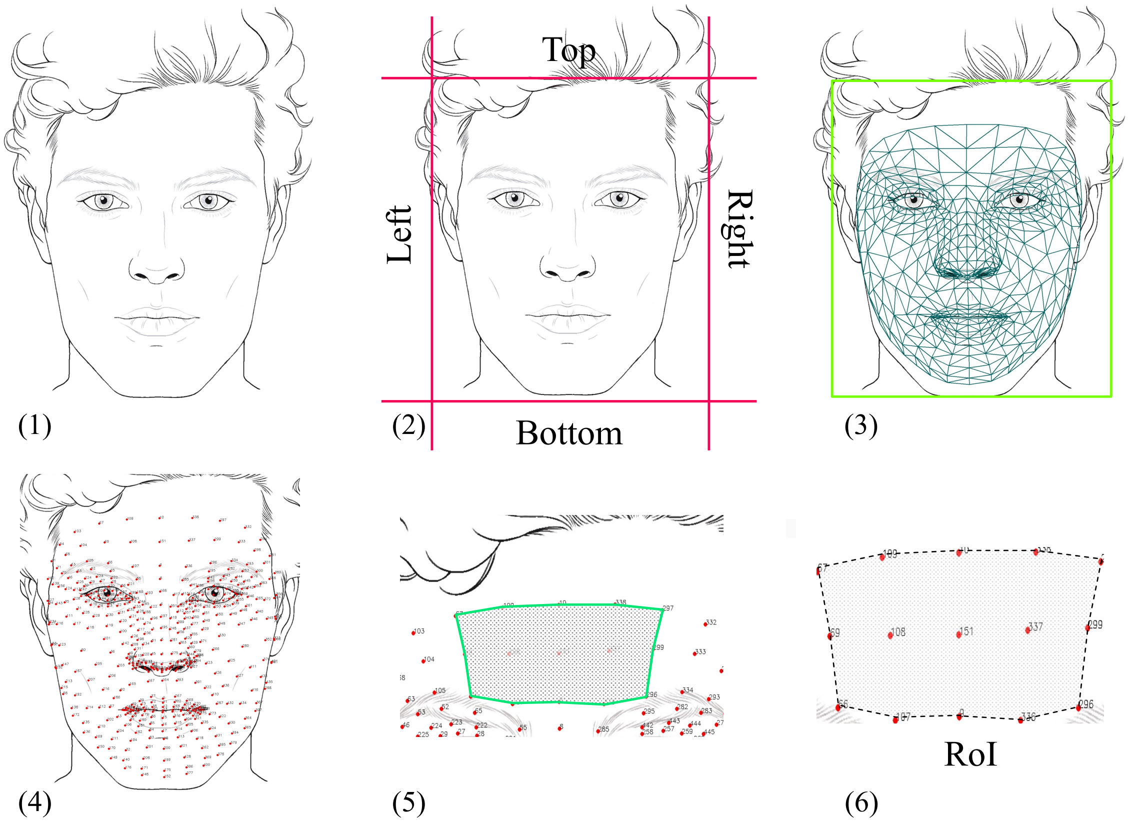


Figure 6.7: The 1st image illustrates a person present in a video frame. The 2nd image portrays the MediaPipe face detector in action. As shown, once a face is detected, the MediaPipe face detector yields the coordinates corresponding to the four corners of the face. The 3rd image, illustrates the face overlaid with attention mesh in the face bounding box. The 4th image highlights the indices of the canonical face model on top of the facial regions. The 5th image hosting the dashed area signifies the indexes of facial landmark points (9, 107, 66, 69, 67, 109, 10, 338, 297, 299, 296, 336). The 6th image hosts the extracted RoI from the original frame.

6.6 Signal extraction

As demonstrated earlier, once a face is detected, we track and segment the forehead. The simultaneous occurrence of detection, tracking, and segmentation processes triggers the extraction of raw image color data from the RoI in parallel. We then utilize an averaging technique called arithmetic mean (AM) [121] to convert the extracted forehead signals into workable traces. Using the Equation 6.2, we can calculate the arithmetic mean for the dataset a ={a$_1$, a$_2$, $\cdots$, a$_n$}.

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i \quad \Longrightarrow \quad \frac{a_1 + a_2 + \cdots + a_n}{n} \tag{6.2}$$

where A is the arithmetic mean, n is the number of values, i is the index of dataset values, and a is the dataset value.

6.6.1 Spatial averaging

Knowing that a color video is composed of 3 color channels, it becomes evident that the RGB image data of the forehead pixels at time t resembles the HR signal.

At this step, we utilize AM to obtain an indication of the central tendency of values that we recently extracted. Using the Equation 6.2, we compute the average of pixel traces spatially and across each time frame (averaging the RGB pixels of each frame or so to say averaging the values of the pixels' assembly matrix like the one in 3.7). Once done, we store the raw RGB signals comprising X$_r(t)$, X$_g(t)$, and X$_b(t)$.

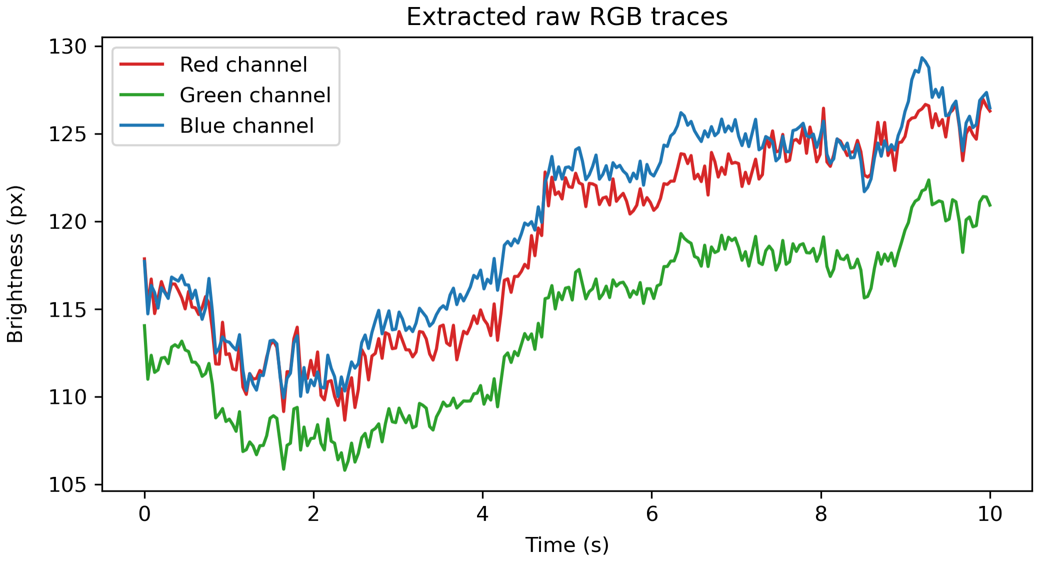


Figure 6.8: The extracted raw RGB traces for one of the videos in our dataset, where each trace is assigned to a color illustrating the brightness intensity of pixels in their corresponding channel during a 10-second time span.

6.7 Signal preparation

After obtaining the $\mathrm{X}_r(t)$, $\mathrm{X}_g(t)$, and $\mathrm{X}_b(t)$ traces, we need to ensure that the collected signals are worthy of processing by the BSS model that we have opted for. Hence, we implement the sequence of the following methods throughout this section to prepare them for the BSS model.

6.7.1 Detrending

Trend is a change in the mean during a period of time. Detrending is the event of removing an element from the dataset which has led to a specific trend. This is typically done to better identify and analyze any remaining variations in the signal that may be of interest, such as periodic fluctuations or random noise. Additionally, detrending helps us to stabilize the variance of a signal, making it easier to detect small changes or variations. the result of detrending is a set of points in a plot with no distinguishable trend [19].

By taking a closer look at Figure 6.8, an uptrend is highly noticeable. At this step we need to remove the parts of the data that has contributed to forming this uptrend, or in other words, detrend it. Figure 6.9 illustrates the plot of the detrended raw RGB traces.

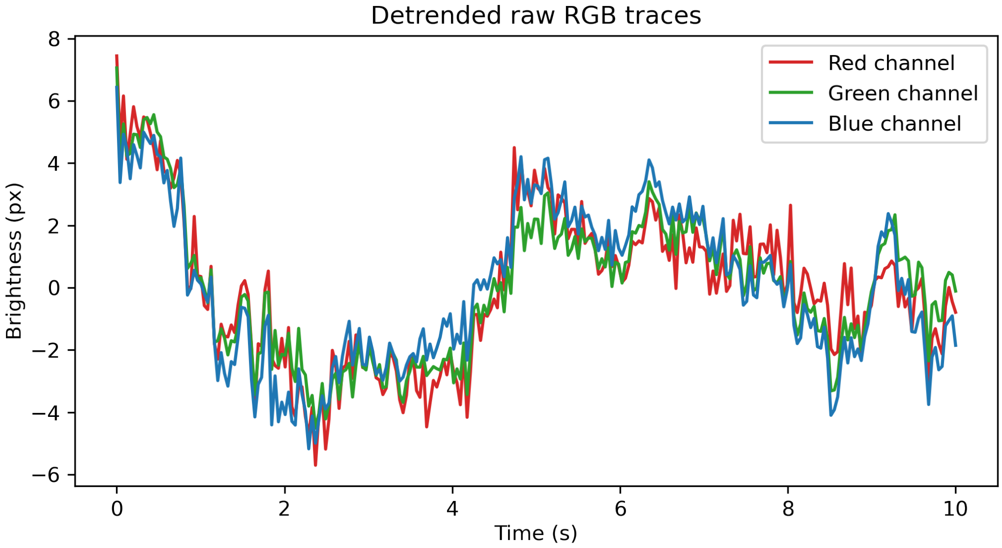


Figure 6.9: A visual example of detrending in time-series data analysis.

6.7.2 Smoothing

Smoothing is a mathematical function that reveals long-term trends of a signal by clearing the signal from short-term extreme fluctuations. In DSP, the inclusion of smoothing methods, known as low-pass filters, often occurs in the form of moving average filters [20, 17]. As seen in Figure 6.9, sudden spikes are evident all over the detrended RGB traces, making it difficult to observe its overall direction.

At this stage, in order to enhance the signal-to-noise ratio and mitigate sudden spikes, we will utilize a moving average filter as explained below.

Moving average

Moving average (MA) filter is a prominent statistical method used on data with identical periods (i.e., time series data). MA bases its calculation on the mean of a specific subset of the whole data points. The "moving" term implies the continuous calculation of the average of the n values from the starting point to the signal output at any given moment.

MA filters are useful in tasks where random noise reduction is required, like rPPG estimation. Aside from removing the noise, it also provides a broad image of the signal trend, which can be used as a predictive analysis technique in different use cases [20].

As stated, upon the completion of the detrending process, we utilize a simple moving average (SMA) with a 5-frame rolling window to eliminate noise and obtain a cleaner version of our signals. Figure 6.10 plots the smooth-detrend version of the RGB traces over a 10-second time span. As shown, the intensity of smaller fluctuations are significantly decreased.

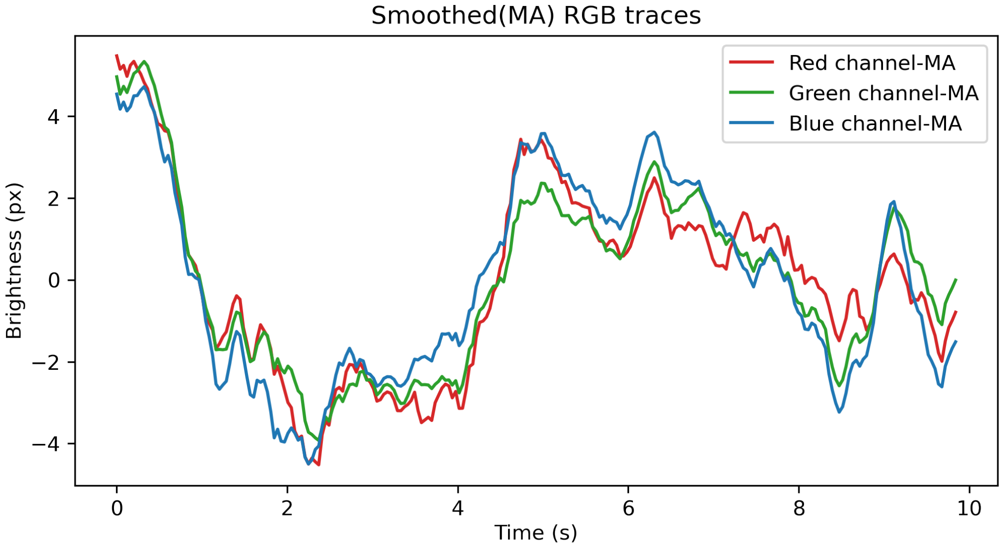


Figure 6.10: Smoothed RGB traces of the previous detrended plot.

6.7.3 Standard deviation

Pre-processing signals for directing them into ICA requires the implementation of certain statistical methods to reduce the noise and signal artifacts. As a statistical measurement, standard deviation (SD) is a scale indicating how much values are scattered from the mean [119]. To fulfil the requirements of the final step within this module, we implement the Equation 6.3 to compute the SD of the smoothed-detrended RGB traces.

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{6.3}$$

where s is the sample SD, N is the observations total number, x$_i$ is the value within the data distribution, and $\bar{x}$ is the sample mean.

6.7.4 Arithmetic mean

In order to prepare our data entries for the standardization, at this step, using the Equation 6.2, we compute the AM of the detrended-smoothed raw RGB traces.

6.7.5 Standard score

The standard score, also known as the z-score, is a statistical indicator that locates an observation placement around its distribution means, Equation 6.4.

$$z = \frac{x - \mu}{\sigma} \tag{6.4}$$

where z is the z-score, x is the observed value in question, μ is the mean of the distribution, and σ is the SD.

Standardization known as the normalization of the z-score is the process of obtaining the standard score of data entries. Standardization provides comparative information between specific values [120]. A necessary condition for the ICA is that the input signals are independent and identically distributed [91, 93, 97], in other words, the signals need to have zero mean and unit variance. To fulfill this requirement, at this stage we standardize our RGB traces and ensure that this necessary condition of ICA is met accordingly.

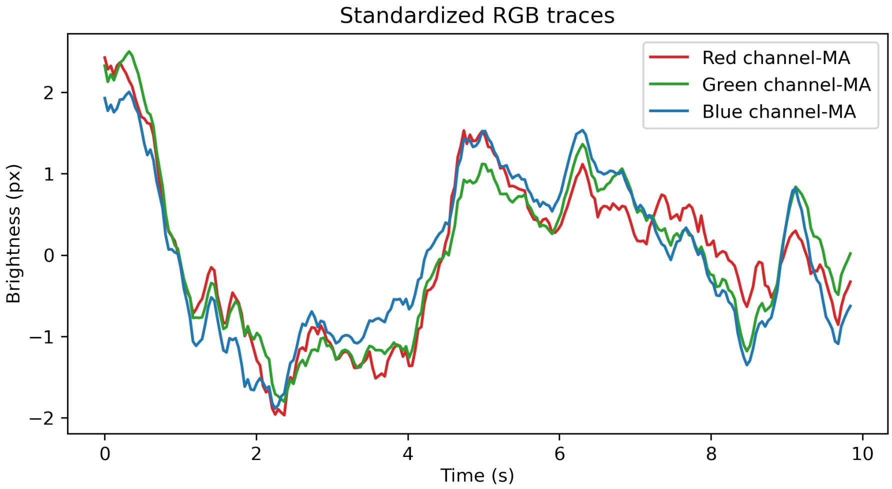


Figure 6.11: Standardized plot of the mean, detrended, and smoothed(MA) RGB traces.

6.8 Signal processing

As its name implies, signal processing is the act of analyzing, manipulating, and synthesizing signals for the purpose of observing hidden data features. We have various motivations behind the inclusion of this step in our proposed architecture.

For instance, the removal of noise or interfering components, separation of biomedical data from the mixture of unknown source signals, filtering disturbances, and most importantly the retrieval of HR data.

Along the way, we have also deployed other signal processing operations that lead to actualization of the above-mentioned processes, such as storing, reconstructing and transforming signals. In our proposed approach, the latter is considered as the basic operations within the process.

6.8.1 ICA

ICA seeks to uncover the set of independent signals or sources from the gathered mix of multidimensional data (raw RGB temporal traces). ICA bases its assumption on equal independency of the underlying signal variables [93].

In the case of Gaussian models where classical approaches fail, ICA has proven to be a richer method in revealing latent sources. There are no restrictions with regard to ICA input data, as it is fully capable of analyzing various sources originating from non-relating fields. In ICA, unlike FFT-based solutions, the components do rely on the structure of gathered data. That is, any changes in the sources of data impact the structure of data, therefore affecting the ICA components. This is why ICA is a case of discovery rather than fixed projections [95].

ICA materialization requires centering and whitening preprocessing steps. Assuming that a signal is already centered, the utilization of the Jacobian rotation matrix produces the whitening matrix that separates the original signal components from the mixed signal.

Expressed by Equation 6.5, ICA considers x$_{rgb}$(t) as the temporal mixed signal, a product of multiplying a random mixing matrix (A) by known source traces z$_{123}$(t), [94, 93].

$$\begin{aligned} x_{rgb}(t) &= A \cdot z_{123}(t) \\ z_{123}(t) &= A^{-1} \cdot x_{rgb}(t) \end{aligned} \tag{6.5}$$

where $x_{rgb}(t) = [x_r(t), x_g(t), x_b(t)]^T$, A is a 3×3 matrix consisting of coefficients, and $z_{123}(t) = [z_1(t), z_2(t), z_3(t)]^T$.

ICA transforms the challenge of source revival into the BSS-type problem of finding the de-mixing matrix ($W = A^{-1}$) responsible for the creation of x$_{rgb}$(t) in the first place, Equation 6.6.

$$z_{123}(t) = W \cdot x_{rgb}(t) \tag{6.6}$$

JADE [96] and FastICA [97] algorithms are two well-known stabilized and effective repeatable ICA implementations. We can employ the FastICA algorithm from the scikit-learn library to obtain an approximation of the source signals $z_{123}(t)$.

As discussed in 5.3.1 the implementation of ICA in rPPG leads to the segregation of raw RGB components from the ROI pixels and grants the retrieval of volumetric changes of blood within cardiac cycles from the skin surface.

FastICA

As earlier discussed, preprocessing data entries is crucial to simplification of any analysis. In that regard, ICA is no exception. Centering and whitening are the standard preprocessing stages for an ICA-backed solution. In dataset attributes that lacked predominant mean and covariance justification, centering and whitening are attenuating factors.

Thanks to the FastICA algorithm [97], we no longer need to separately implement these preliminary measures. As its name implies, FastICA provides a fast algorithm for ICA, and it includes all the necessary steps within itself.

Knowing that ICA expects an equal balance between the number of the source signals and the total of the observed traces, it became evident that the uncorrelated source signals are not more than three components S$_1(t)$, S$_2(t)$, and S$_3(t)$.

Once we pass the raw RGB signals through the FastICA algorithm, our pre-processed RGB traces will decompose into 3 uncorrelated source signals known as ICA components. The obtained ICA components are not consistent, and they may switch place every time, hence, we cannot label them separately and further processing is needed.

In the following, as a preparatory measure in the elimination of noise and other abnormalities within our raw ICA components, we will utilize a noise reduction operation, guaranteeing the refinement of workable data.

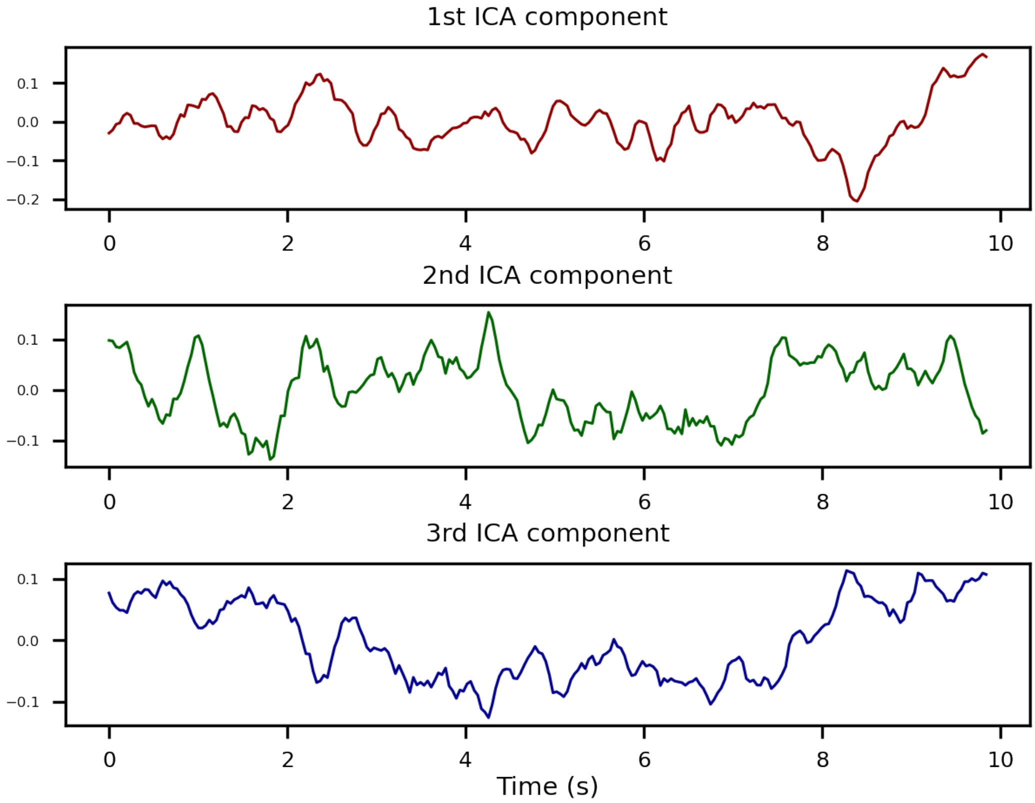


Figure 6.12: ICA components derived from the pre-processed RGB traces.

6.8.2 Denoising

As previously discussed in 5.4.1, the signal acquisition and the integrity of HR estimation in rPPG studies faces various altering factors, including noise and motion artifacts.

Denoising or noise reduction refers to the process of removing disturbances from signals and their segregation into their subcomponents. The specification for a noise reduction method lies in the criteria that it seeks to assist. A filter, as its name implies, is a discriminatory measure with a defined selection mechanism that fulfills a noise reduction process[18].

In anticipation of random inconsistencies stemming from noise and motion artifacts, we necessitate the inclusion of a denoising filter, which safeguards the removal or blockage of disturbances of non-cardiac origins. We use the upper and lower normal HR frequency ranges as the determining factors for our proposed filter.

During the early trials, we understood that not all filters comply with our objective. After many attempts and evaluations, it became clear that choosing an ideal band-pass filter (BPF) yields the most favorable outcomes.

To our knowledge, the BPF is capable of transmitting all the frequencies in between the normal HR frequency range while blocking the rest. In addition, the BPF also does not attenuate the frequencies and kept the integrity of ICA components.

Having said the above, at this stage, we retrofit a 5th order Butterworth BPF [21] to achieve our goal of eliminating noise and other disturbances from the ICA components.

We set the bounding frequencies for the BPF lowcut and highcut ranges to 0.75 and 4.5 Hz respectively. The sampling frequency f_s is also set to the sampling rate of the captured videos, equating to 25 Hz or fps, and the Nyquist frequency of half the f_s.

To avoid any phase distortions, the resulting BPF filter together with the ICA components are then feed into a digital filter in both directions (forward and backwards) to the signals [15]. This step, ensures that the measured HRs are within an acceptable domain representing the highest frequency magnitude.

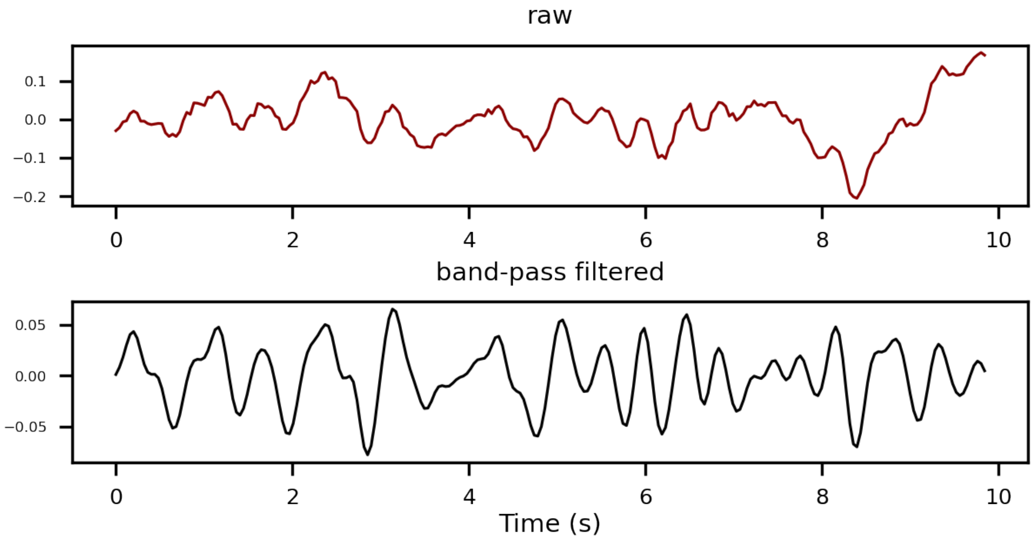


Figure 6.13: Raw vs denoised comparison of the 1st ICA component.

6.8.3 Spectral analysis

Spectral analysis also known as time-frequency analysis refers to the simultaneous study of signals in time and frequency domains. The spectral analysis seeks to retrieve key elements concerning amplitude, phase, and frequency of signal components [20].

The inclusion of spectral analysis in this step simplifies and enhances our comprehension of denoised ICA components. To achieve this objective, we implement frequency domain conversion through the following method.

FFT

In DSP, Fourier transform is a well-known algorithm that transfers a time-domain signal to its corresponding frequency-domain. To actualize the frequency domain conversion, we apply a discrete Fourier transform(DFT) algorithm known as the fast Fourier transform(FFT) [14] on the denoised traces [13].

Subsequently, we compute the absolute value of the FFT traces along with their corresponding frequencies and store them as frequency magnitude and amplitude for signals peak inspection. Figure 6.14 plots the frequency-magnitude response of our system.

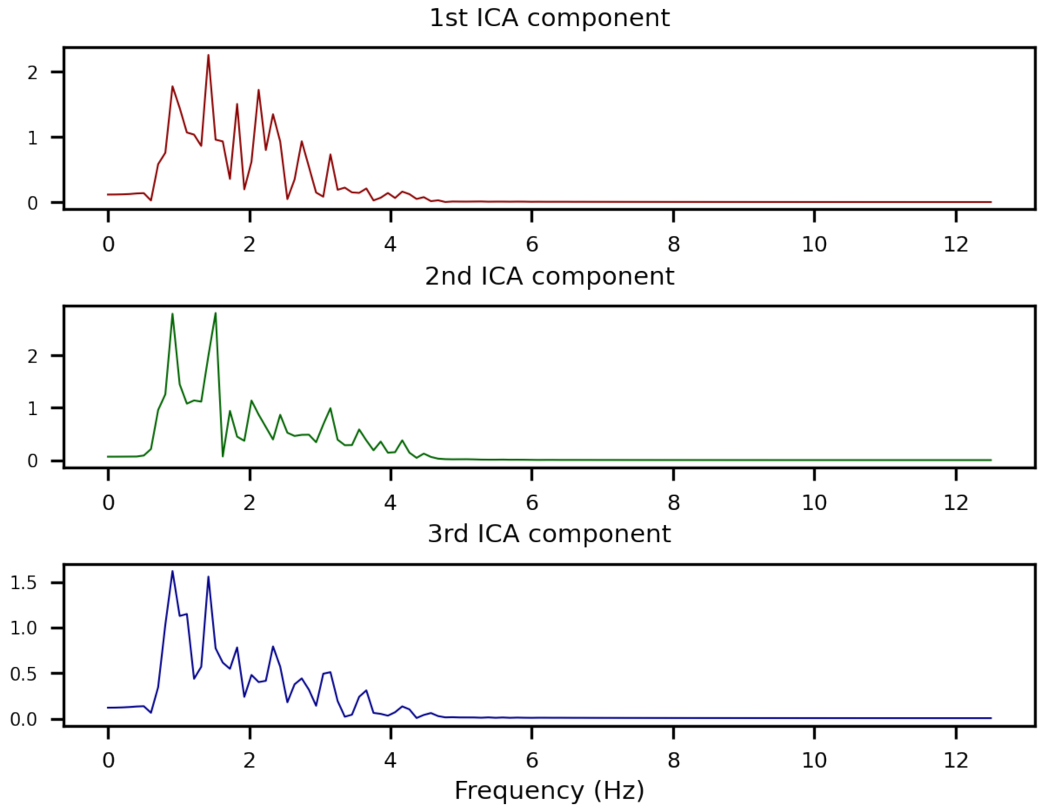


Figure 6.14: Smoothed ICA components through utilizing a hanning function.

Peak detection

In rPPG studies, peak amplitude detection is a prominent way in the detection of cardiac activity that is within a reasonable HR range. If signals are ideally prepared, the frequency of a viable HR should be on par with the frequency associated with the highest magnitude.

As Figure 6.14 illustrates, the frequency-magnitude plot of the second ICA component bears the highest magnitude among the other two components. Therefore, the frequency of its highest magnitude correlates to that of HR.

6.9 HR extraction

As mentioned earlier, the frequency of the highest magnitude equates to the frequency of HR. In that sense, once we detect the prominent amplitude peak among ICA components in the desired time frame (10 seconds), that frequency can be utilized to compute the HR.

6.9.1 Manually computing HR

With the frequency unit being in Hz (one-cycle per unit of time(s)), we can easily calculate the HR value by multiplying the highest magnitude frequency f with 60. To verify our claim, we manually tested our assumption on a handful of frequency-magnitude plots and compared our findings with the reference PPG. To our surprise, the results were in-line with those of HR values collected through physical contact.

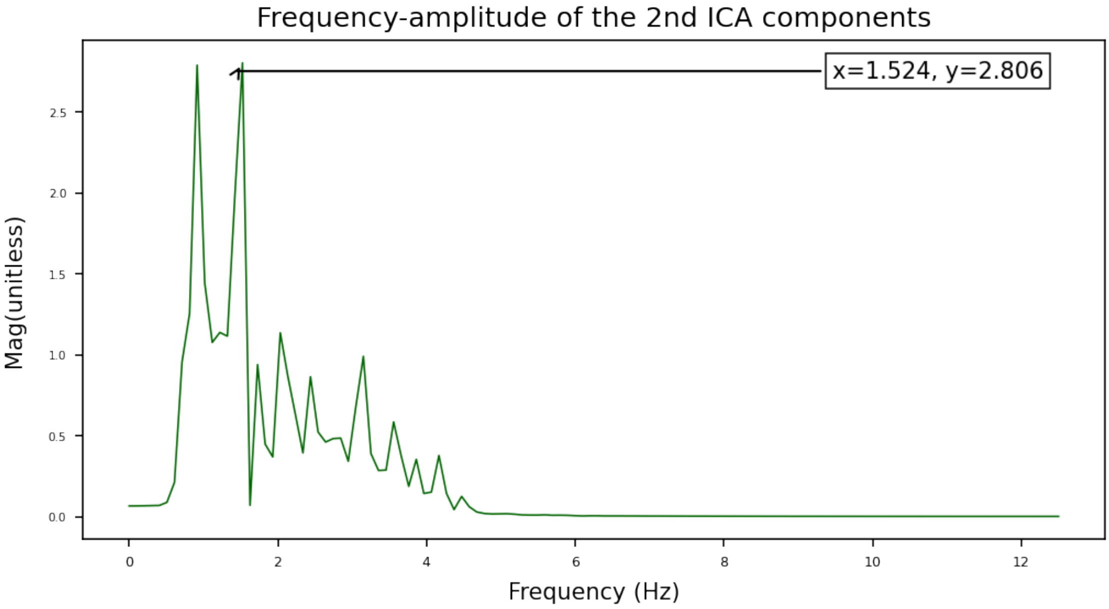


Figure 6.15: The 2nd ICA component bearing the maximum frequency magnitude.

As Figure 6.15 illustrates, the frequency of 1.524 Hz and magnitude of 2.806 correspond to the maximum peak coordinates. Equation 6.7 demonstrates how HR is calculated by only using the peak coordinate.

$$\begin{aligned} &HR = x \times f, \quad x = \ 1.524 \quad Hz \\ &HR = 1.524 \quad (Hz) \times 60 \quad (\tfrac{1}{Hz}) = \ 91.46 \quad bpm \end{aligned} \tag{6.7}$$

where HR is the total number of heart beats per minute, x is the frequency associated with peak magnitude, and f is 1 hertz equal to 60 oscillations per minute (one cycle per second).

6.9.2 Automated peak estimation

To automate the manual approach, we design a mechanism that initially finds the highest magnitude of ICA components and compares it against every other one within each time interval. Once it is clear which component bears the highest amplitude, the index number associated with that will be stored and utilized in accessing its frequency component. Instantly after, our automated algorithm uses that frequency to compute the HR. Figure 6.16 illustrates the automated peak estimation module 6.1 for one of the candidates in the gym scenario and in various time frames.

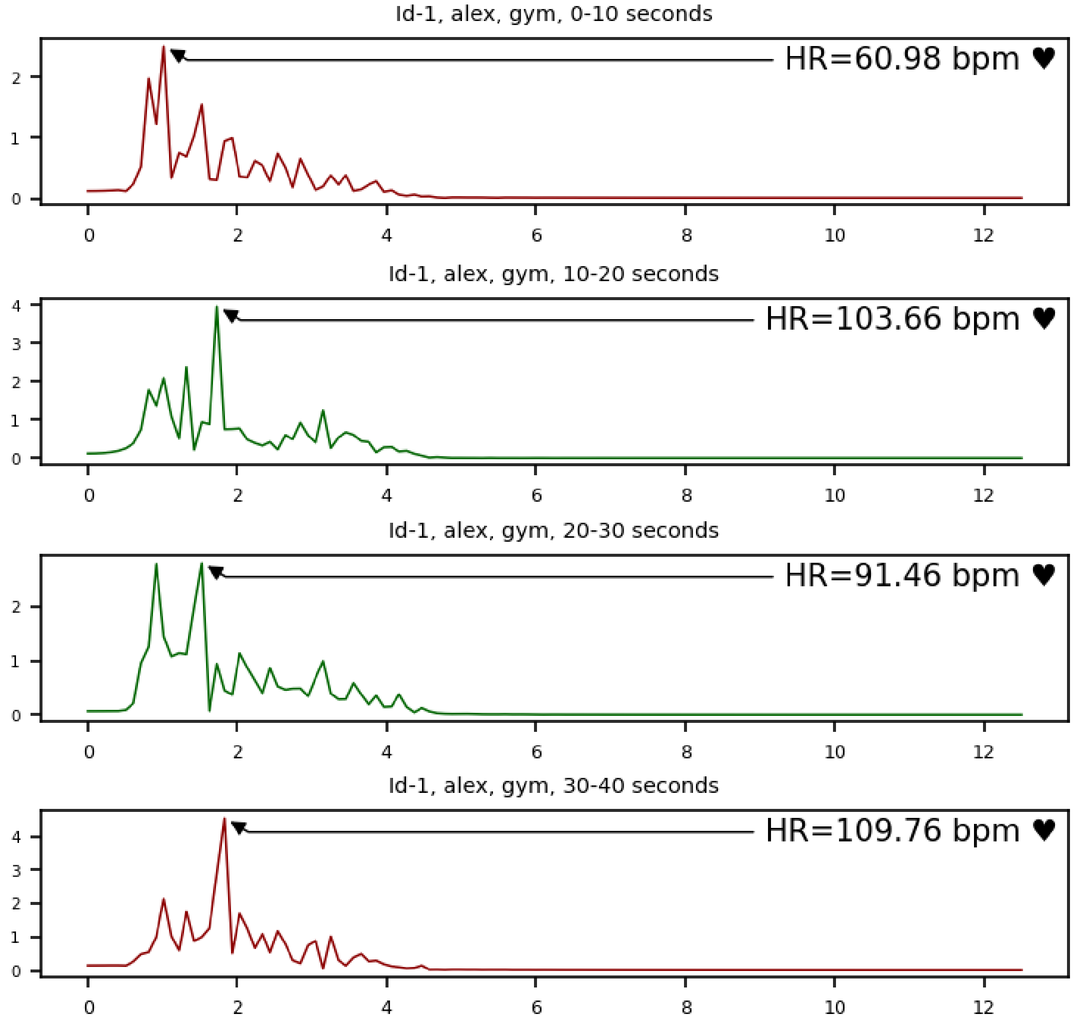


Figure 6.16: Automated peak detection combined with HR estimation during 10 sec time frames.

To verify the accuracy of the implemented automation mechanism, we compared the obtained HRs against the actual plots. Based on what we observed, there was no significant difference between the HRs computed manually against those automated.

In the process of obtaining results from the dataset at hand, we applied all the aforementioned procedures on a total of four candidates in 4 different scenarios. The next chapter thoroughly discusses the findings and compares them against the ground truth PPG data.

6.9.3 Reference PPG preparation

For analysis purposes, we initially extract the HR data along with time stamps from the dataset. Then we utilize the sampling frequency as the basis for the number of samples in every second and average the HRs into one value. As a result, HRs and time-stamps will be down sampled and stored as ground truth HR for future proposes. Figure 6.17 illustrates the structure of the XML reference file belonging to one of the candidates.

The authors of the LGI dataset have used a CMS50E PPG pulse oximeter to provide a synchronized ground truth PPG signal from participants' finger. The reference PPG signal was collected at an average sampling frequency of $f_s = 60Hz$. In addition, they have also captured time stamps and the device's pre-computed HR. All three parameters are stored in simple text-based XML files and can be found under the subdirectory:

id\name\name_scenario\ground truth PPG (cms50_stream_handler.xml)

```
<cereal>
    <value0>
        <value0>0</value0> # time-stamp 0 = one sixtieth of a second
        <value1>68</value1> # pulseoximeter computed HR = 68
        <value2>9</value2> # PPG signal amplitude
    </value0>
    .
    .
    .
    <value60>
        <value0>60</value0> # time-stamp 60 = 1 second
        <value1>68</value1>
        <value2>23</value2>
    </value60>
    .
    .
    .
    <value24117>
        <value0>24116</value0> # 401.93 seconds elapsed
        <value1>73</value1>
        <value2>44</value2>
    </value24117>
</cereal>
```

Figure 6.17: Every 60 entries of logged data in the XML files, equates to one second. As depicted, the XML file is a collection of three values at different instances of time. Value0 counts the frame numbers and can be considered as a time-stamp. Value1 holds the pulse oximeter computed HR. Value2 logs the PPG signal amplitude.

As Table 5.2 indicated, the duration of captured videos in different settings is not the same and varies from subject to subject. To simplify our analysis and overcome this issue, we use 10 seconds long sliding window with 1 millisecond steps throughout the process.

Chapter 7

Results and Discussion

7.1 Preface

The following chapter presents the results and discussion of our study. The discussion section provides an in-depth examination of the results, highlighting key findings and their implications. It also provides a critical evaluation of the methods used and the limitations of the study. In general, this chapter aims to provide a comprehensive understanding of the objectives, methods, and results of the study.

7.1.1 Processing environment

In this study, we utilized the Python programming language in combination with its relevant libraries for CV, DSP, and data analysis to facilitate the various operations and processes. These include data preparation, loading data, RoI detection, selection, and segmentation, raw data acquisition, data processing, and evaluation. A list of the most important libraries used can be found in the appendix (8.2).

In the last section of the previous chapter, we performed a number of perpetuity measures in the Python programming environment to simplify our access to the HRs collected through physical contact. In this chapter, we make use of the reference PPG data obtained in 6.9.2 to evaluate our results and measure the performance of various face detectors in noisy environments.

7.1.2 Evaluation metrics

The proposed architecture's performance is evaluated under various conditions using three metrics: mean absolute error (MAE), Equation 7.1, mean squared error regression loss (MSE), Equation 7.2, and Pearson correlation (r), Equation 7.3.

$$MAE = (\frac{1}{n})\sum_{i=1}^{n}|y_i - x_i| \tag{7.1}$$

$$MSE = (\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2 \tag{7.2}$$

where MAE is the average of absolute errors, MSE is the regression loss of the mean squared error (MSE), n is the total number of data points, y$_i$ is the predicted value of the i-th data point and x$_i$ is the actual value of the i-th data point.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{7.3}$$

where r is the Pearson correlation coefficient. n n is the number of data points in the dataset. x$_i$ and y$_i$ are the two variables correlated in the sample. $\bar{x}$ and $\bar{y}$ are the mean of x and y variables, respectively.

7.2 Experimental results

In this section, we present our results (HRs computed using rPPG) in a comparative format with reference data (HRs extracted from PPG). Using a series of graphs, we illustrate the consistency and accuracy of our pipeline in different scenarios and with three different face detectors (Haarcascade, Dlib, and Mediapipe).

Furthermore, for more diagnostic and advance analysis, we utilize the evaluation metrics mentioned in 7.1.2 to analyze the robustness of selected face detectors concerning the issues of noise and motion artifacts in each segment of our dataset.

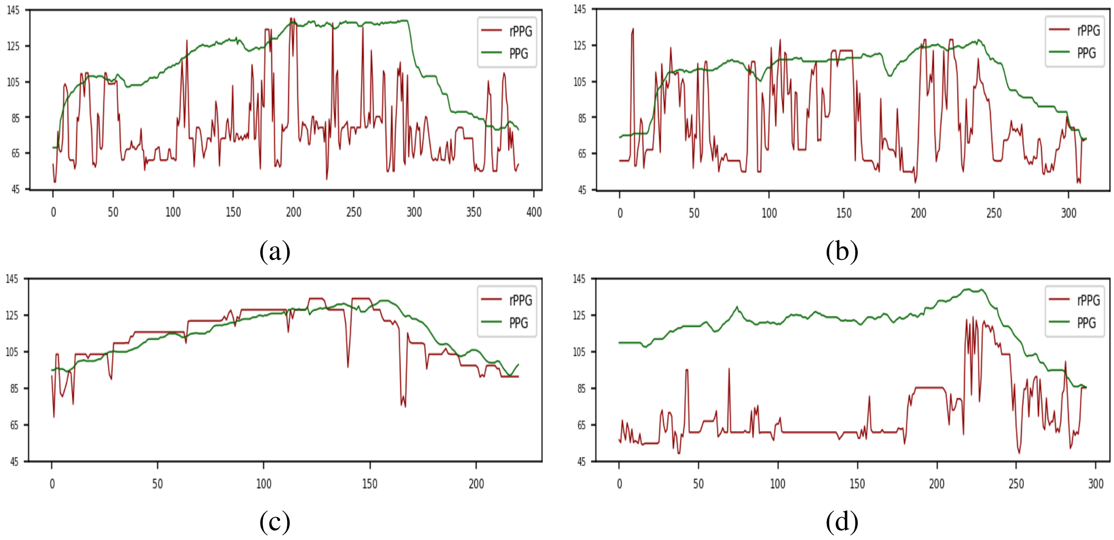
This comparison will provide valuable information on the effectiveness of rPPG for HR measurement and the suitability of different face detectors in noisy scenarios. In the following, we categorize our results based on scenarios (gym, resting, rotation, and talk). As we use three different face detectors, the results in each scenario are further classified into their method of face detection.

7.2.1 Scenario - Gym

The gym scenario is one of the four scenarios existing in LGI-PPGI dataset. Videos of this scenario are recorded from four different candidates during an exercise session on a stationary bicycle. In this section, we demonstrate the computed rPPG HR along with the corresponding reference PPG HR, using a series of plots. On the next page, we classify our results based on the face detection method utilized. For example, in the 'Haarcascade' subsection, there are four plots and each plot is labeled with an alphabet letter associated with the candidate's name. As illustrated, we have applied the same classification concept for the results obtained using Dlib and MediaPipe.

Haarcascade

Figure 7.1 plots the computed HR (colored red - rPPG) against reference HR (colored green - PPG) for all candidates with the Haarcascade face detector.



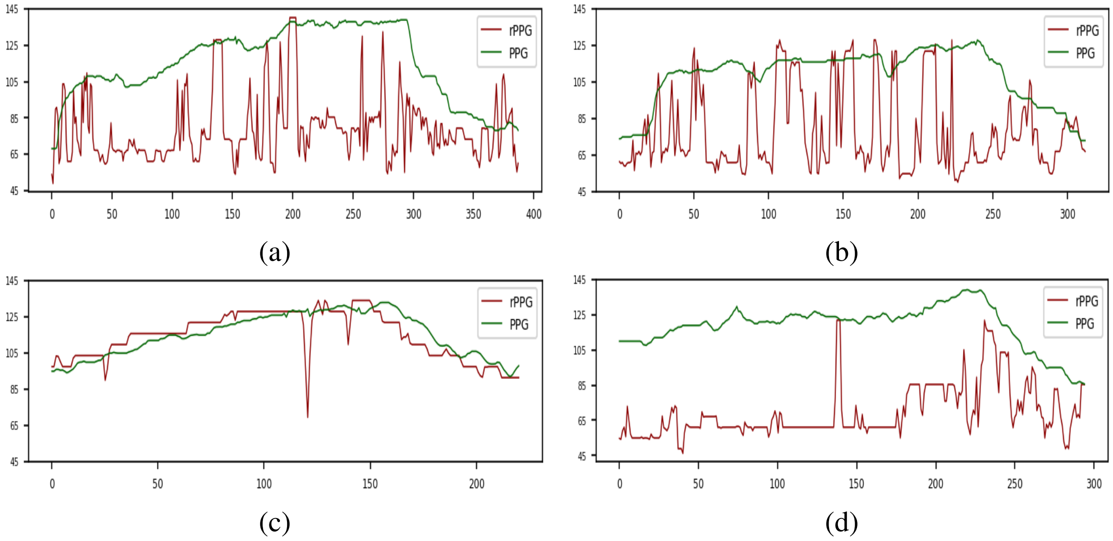
(a) (b)

(c) (d)

Figure 7.1: X-axis represents time (sec) and Y-axis represents HR (bpm). Plots labeled with a, b, c, and d are associated with candidate ID1, ID2, ID3, and ID4 respectively.

Dlib

Figure 7.2 plots the computed HR (colored red - rPPG) against the reference HR (colored green - PPG) for all candidates with the Dlib face detector.



(a) (b)

(c) (d)

Figure 7.2: X-axis represents time (sec) and Y-axis represents HR (bpm). Plots labeled with a, b, c, and d are associated with candidate ID1, ID2, ID3, and ID4 respectively.

MediaPipe

Figure 7.3 plots the computed HR (colored red - rPPG) against reference HR (colored green - PPG) for all the candidates with the MediaPipe face detector.

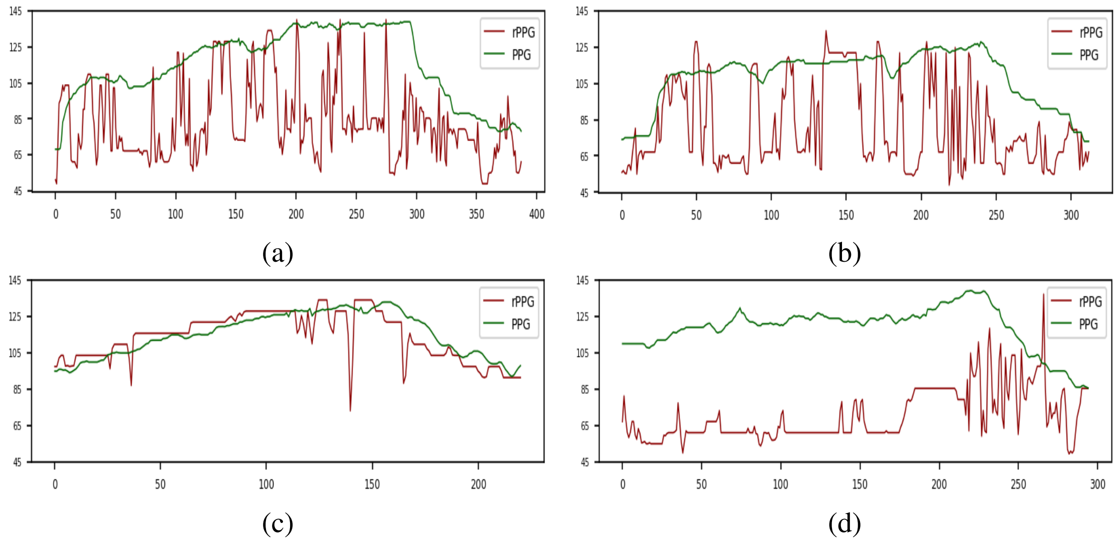


Figure 7.3: X-axis represents time (sec) and Y-axis represents HR (bpm). Plots labeled with a, b, c, and d are associated with candidate ID1, ID2, ID3, and ID4 respectively.

Performance summary

Table 7.1 summarizes the obtained metrics for all candidates in the gym scenario. Overall, the lowest values for MAE and MSE are achieved when the face detector is set to MediaPipe. With an average MAE of 29.04, and despite the presence of motion artifacts and illumination variation, MediaPipe offers the best results in this category.

Gym				
Candidate	Metric	Haarcascade	Dlib	MediaPipe
Id-1	MAE	37.60	37.37	33.38
	MSE	1829.32	1831.32	1566.98
	r	0.29	0.24	0.31
Id-2	MAE	28.90	32.21	29.26
	MSE	1251.01	1539.02	1293.94
	r	0.30	0.15	0.31
Id-3	MAE	6.10	5.53	6.01
	MSE	82.51	60.51	72.61
	r	0.79	0.80	0.77
Id-4	MAE	48.06	49.51	48.95
	MSE	2623.18	2740.81	2688.71
	r	0.26	0.22	0.07
Overall	MAE	30.16	31.15	29.04
	MSE	1466.50	1542.91	1405.56
	r	0.41	0.35	0.36

Table 7.1: Computed evaluation metrics for candidates at gym.

This scenario has the sharpest deviations between the predicted HR and GT HR. Compared to other segments of the dataset, these deviations at certain points and for some candidates are of a considerable magnitude, for instance Figures (7.1a and 7.2a)

Based on our observations, we noticed that some candidates began to sweat during the exercise session, such as ID3. Initially, we had the assumption that sweating undermines the function of our signal processing module. However, upon further examination, we noticed it had minimal impact on the quality of the extracted RGB signals and thus the results, as demonstrated in Figures (7.1c, 7.2c, and 7.3c).

Among the candidates in this scenario, ID3 has the least amount of body and facial movements. As depicted, the predicted HRs for this candidate is closely correlated with the GT HRs and serves as an indication of the impact of motion artifacts on the predicted HRs.

While the results for ID1 and ID2 have some correlation with the GT HR, as illustrated in Figures (7.1a, 7.1b, 7.2a, 7.2b, 7.3a, and 7.3b), the predicted HRs for ID4 were found to be less correlated with the GT HR, as demonstrated in Figures (7.1d, 7.2d, and 7.3d). As discussed in 5.4.1, the pigmentation of the participant's skin affects the rPPG readings. However, given that all participants have a similar skin pigmentation, it is unlikely that this factor is the cause of the uncorrelated results for ID4.

Upon analyzing and comparing videos of this scenario for all participants, we found that the environment illumination color temperature for ID4 is more tilted towards shades of cool white compared to the other candidates. At this point, other than subject motion and scene illumination, we cannot make any further assumptions to explain the case of ID4. Perhaps, if [74, 75] had disclosed more detail about participants' health background, our assumptions would have not been narrowed down to the above-mentioned observations.

The results of our study uncovers that motion artifacts, unstable illumination, variations in the angle of illumination, and reflections on the candidate's face are the primary factors negatively impacting the accuracy of the predicted HRs against the GT HRs.

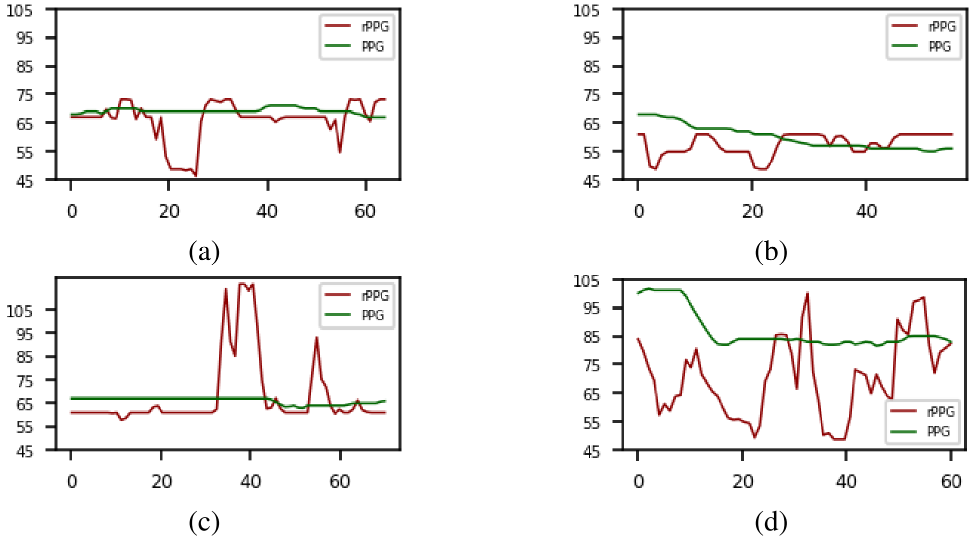
7.2.2 Scenario - Rotation

In the LGI-PPGI dataset, the rotation scenario is one of the four scenarios represented. The videos of this scenario are captured from four distinct subjects, while each subject rotates their head to the sides on two separate occasions. On the following page, we organize our results based on the face detection approaches used.

For instance, in the “Haarcascade” subsection, there are four plots, and each plot is labeled with an alphabet letter associated with the candidate’s name. As illustrated, we have applied the same classification concept for the results obtained using Dlib and MediaPipe.

Haarcascade

Figure 7.4 plots the computed HR (colored red - rPPG) against reference HR (colored green - PPG) for all the candidates with the Haarcascade face detector.

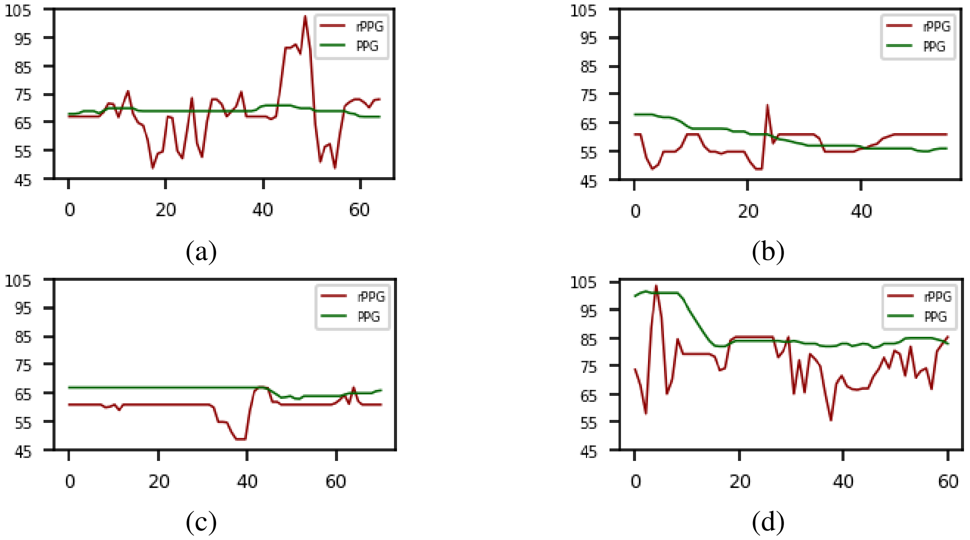


(a) (b) (c) (d)

Figure 7.4: X-axis represents time (sec) and Y-axis represents HR (bpm). Plots labeled with a, b, c, and d are associated with candidate ID1, ID2, ID3, and ID4 respectively.

Dlib

Figure 7.5 plots the computed HR (colored red - rPPG) against the reference HR (colored green - PPG) for all candidates with the Dlib face detector.

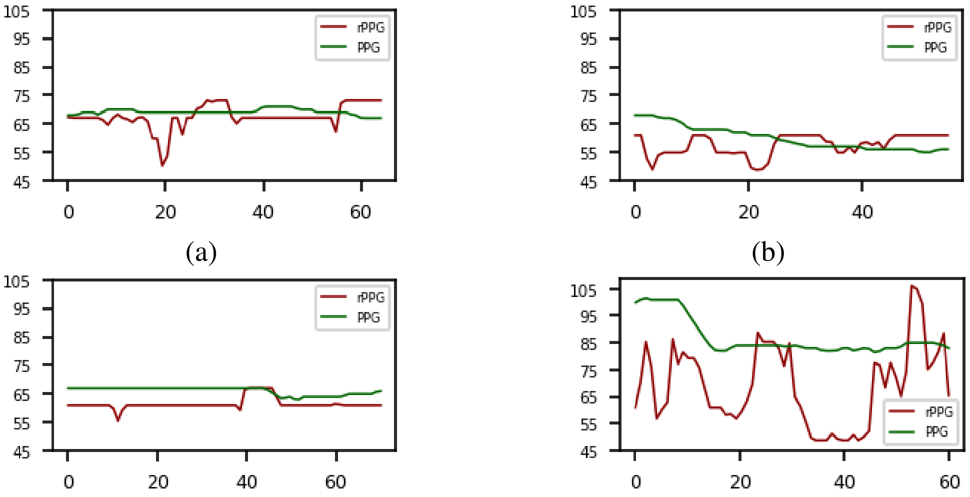


(a) (b) (c) (d)

Figure 7.5: X-axis represents time (sec) and Y-axis represents HR (bpm). Plots labeled with a, b, c, and d are associated with candidate ID1, ID2, ID3, and ID4 respectively.

MediaPipe

Figure 7.6 plots the computed HR (colored red - rPPG) against reference HR (colored green - PPG) for all the candidates with the MediaPipe face detector.



(a) (b) (c) (d)

Figure 7.6: X-axis represents time (sec) and Y-axis represents HR (bpm). Plots labeled with a, b, c, and d are associated with candidate ID1, ID2, ID3, and ID4 respectively.

Performance summary

Table 7.2 summarizes the evaluated metrics for all candidates in the rotation scenario. Overall, the lowest MAE and MSE values are obtained when the face detector is set to Dlib. With an average MAE of 7.43 and MSE of 102.38, Dlib stands slightly ahead of MediaPipe and well above Haarcascade in extracting quality data.

Rotation				
Candidate	Metric	Haarcascade	Dlib	MediaPipe
Id-1	MAE MSE r	5.10 58.22 -0.05	7.38 105.63 0.28	3.86 24.10 -0.30
Id-2	MAE MSE r	5.83 52.56 -0.47	5.90 52.71 -0.37	5.78 50.97 -0.46
Id-3	MAE MSE r	9.56 236.90 0.14	5.63 46.60 -0.21	4.64 26.40 0.06
Id-4	MAE MSE r	19.01 493.87 0.03	10.81 204.60 0.15	20.02 535.70 0.20
Overall	MAE MSE r	9.87 210.38 -0.09	7.43 102.38 -0.04	8.57 159.29 -0.12

Table 7.2: Computed evaluation metrics for candidates having their heads rotating.

In this scenario, we evaluated the ability of the selected face detectors to handle head rotation. The participants were instructed to rotate their heads to the side in two separate instances, each lasting around 8–9 seconds, while videos were being recorded.

As illustrated in Figures (7.4a, 7.4c, 7.4d), after a few seconds of relative correlation between predicted and GT HRs, there are significant sudden spikes in the predicted HRs, followed by a return to a moderately correlated state.

From our observations, it is clear that significant head rotations have a negative impact on the correlation between the predicted and GT HR. This is evident in the sudden spikes seen across all the Figures (7.4, 7.5, 7.6) of this segment, which occur shortly after the head rotation and last for the duration of the rotation.

The results of various face detectors were compared for each candidate, and it was discovered that the effectiveness of the face detection technique plays a crucial role in reducing the impact of head rotation significantly. For example, in the case of candidate ID3, the predicted HRs through the Mediapipe face detection approach show a better correlation with the GT HRs, as seen in Figures (7.4c, 7.5c, 7.6c).

The results for candidates ID1, ID2, and ID3 align with the impact of head rotation on the correlation between predicted and GT HR, but in the case of ID4, the results displayed significant deviation from the GT HR, as shown in Figures (7.4d, 7.5d, and 7.6d).

Despite the absence of other noticeable noises in the video, and aside from head movements, we were unable to determine the cause of these fluctuations. It is possible that the fluctuations in the results may be attributed to environmental illumination noise or an undisclosed underlying health issue.

7.2.3 Scenario - Resting

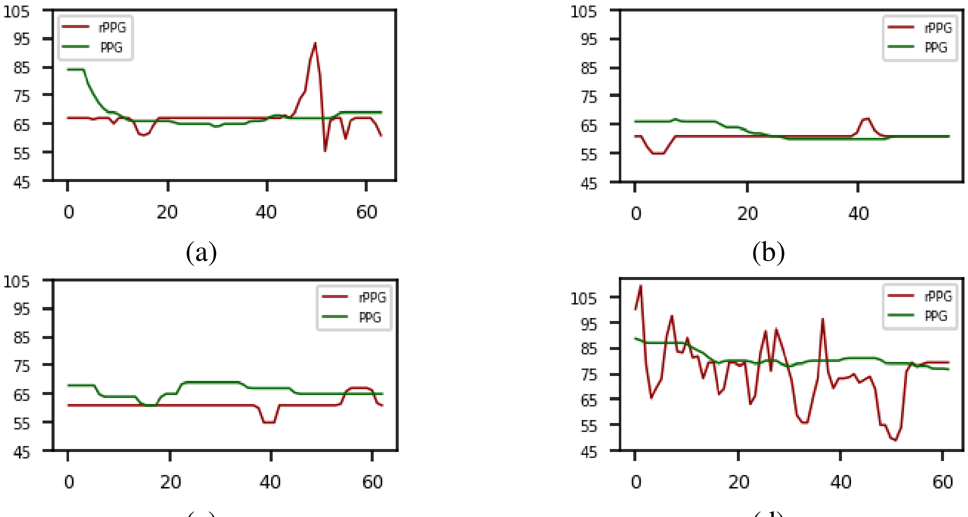
In the LGI-PPGI dataset, the resting scenario is one of the four scenarios recorded. Videos of this scenario are obtained while subjects are seated in a relaxed state.

In the following page, we classify and present our results through a series of plots. For comparison purposes, each plot bears the reference PPG HR (assigned with a green color) and the computed rPPG HR (assigned with a red color).

For instance, in the “Dlib” subsection, there are four plots, and each plot is labeled with an alphabet letter associated with the candidate’s name. As illustrated, we have applied the same classification concept for the results obtained using Haarcascade and MediaPipe.

Haarcascade

Figure 7.7 plots the computed HR (colored red - rPPG) against reference HR (colored green - PPG) for all the candidates with the Haarcascade face detector.



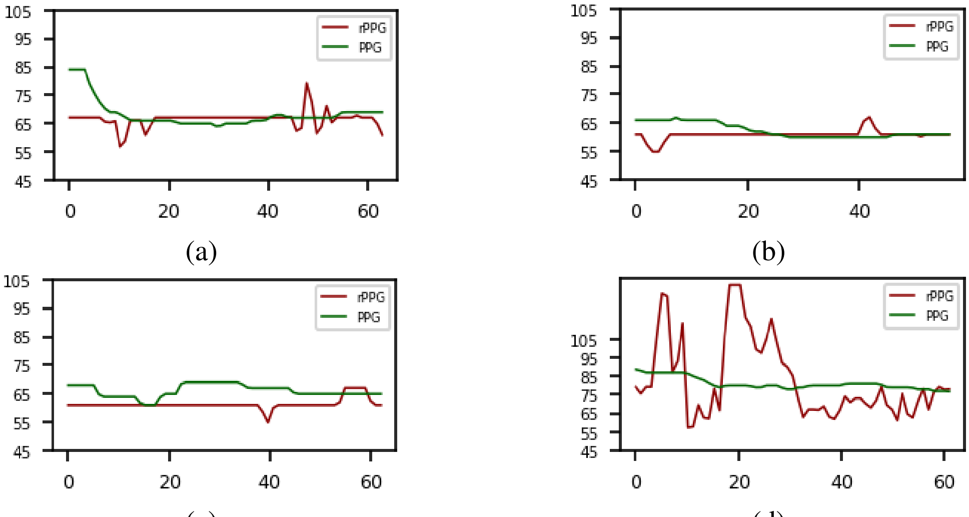
(a) (b)

(c) (d)

Figure 7.7: X-axis represents time (sec) and Y-axis represents HR (bpm). Plots labeled with a, b, c, and d are associated with candidate ID1, ID2, ID3, and ID4 respectively.

Dlib

Figure 7.8 plots the computed HR (colored red - rPPG) against reference HR (colored green - PPG) for all candidates with the Dlib face detector.



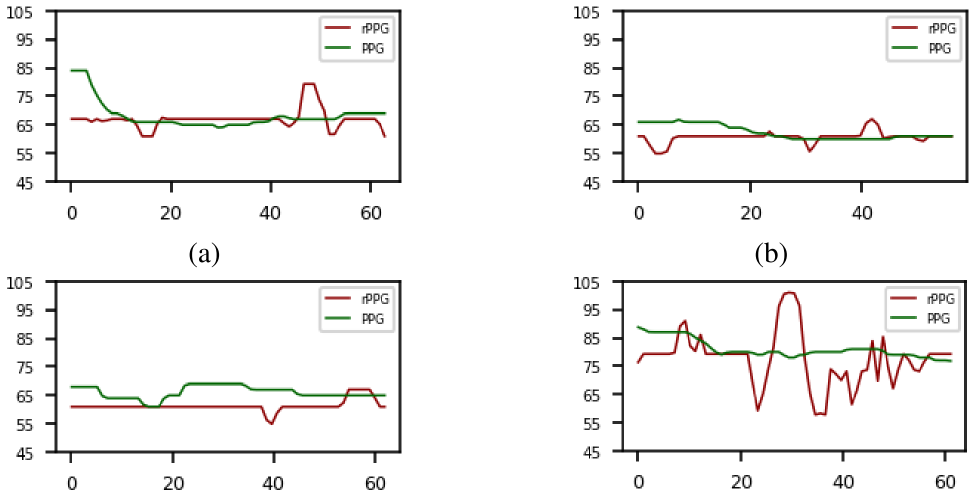
(a) (b)

(c) (d)

Figure 7.8: X-axis represents time (sec) and Y-axis represents HR (bpm). Plots labeled with a, b, c, and d are associated with candidate ID1, ID2, ID3, and ID4 respectively.

MediaPipe

Figure 7.9 plots the computed HR (colored red - rPPG) against the reference HR (colored green - PPG) for all candidates with the MediaPipe face detector.



(a)

(b)

(c)

(d)

Figure 7.9: X-axis represents time (sec) and Y-axis represents HR (bpm). Plots labeled with a, b, c, and d are associated with candidate ID1, ID2, ID3, and ID4 respectively.

Performance summary

Table 7.3 summarizes the metrics obtained for all candidates in the resting scenario. Overall, the lowest MAE and MSE values are achieved when the face detector is set to MediaPipe. With an average MAE of 4.92 and an MSE of 48.16, MediaPipe is slightly ahead of Haarcascade, but far ahead of Dlib in extracting quality data.

Resting				
Candidate	Metric	Haarcascade	Dlib	MediaPipe
Id-1	MAE	4.60	3.83	4.07
	MSE	53.67	33.76	36.37
	r	-0.03	0.02	-0.01
Id-2	MAE	2.76	2.61	2.80
	MSE	16.96	14.84	16.43
	r	-0.50	-0.41	-0.3
Id-3	MAE	5.12	5.01	5.03
	MSE	34.36	31.53	32.37
	r	-0.18	-0.17	-0.18
Id-4	MAE	9.39	16.29	7.79
	MSE	153.93	438.60	107.47
	r	0.39	0.16	0.10
Overall	MAE	5.46	6.93	4.92
	MSE	64.73	129.68	48.16
	r	-0.08	-0.1	-0.1

Table 7.3: Computed evaluation metrics for candidates in resting mode.

Given the relaxed state of candidates, we had the initial assumption of seeing a closer connection between the predicted and GT HRs. Soon after comparing the results, it became clear that our initial expectation was more of a misconception, mainly due to candidate-induced interventions during data acquisition.

For instance, in the case of candidate ID1, we expected the candidate to remain still and have no facial motion [74, 75]. However, 48 seconds into data acquisition, the candidate clearly induces head movement accompanied by continuous facial expression.

As shown in Figures (7.7a, 7.8a, and 7.9a), during the last 20 seconds, the predicted HR deviates from the reference HR. We attributed this deviation to the subject's sudden movement, which is evident as sharp spikes in the HR measurements.

As the illumination was not identical for all candidates, the results obtained for candidates ID2 and ID3 are more in line with real observations, Figures (7.7b, 7.7c, 7.8b, 7.8c, 7.9b, and 7.9c). However, the same cannot be said for candidate ID4, Figures (7.7d, 7.8d, and 7.9d). Despite the subject's compliance (ID4) with the instructions, the predicted HR does not match the GT HR.

Considering that, all candidates were in the same environment and in a resting state, we expect this discrepancy to be due to moderate variations in the illumination of the environment and/or an indication of underlying health issues such as poor circulation.

7.2.4 Scenario - Talk

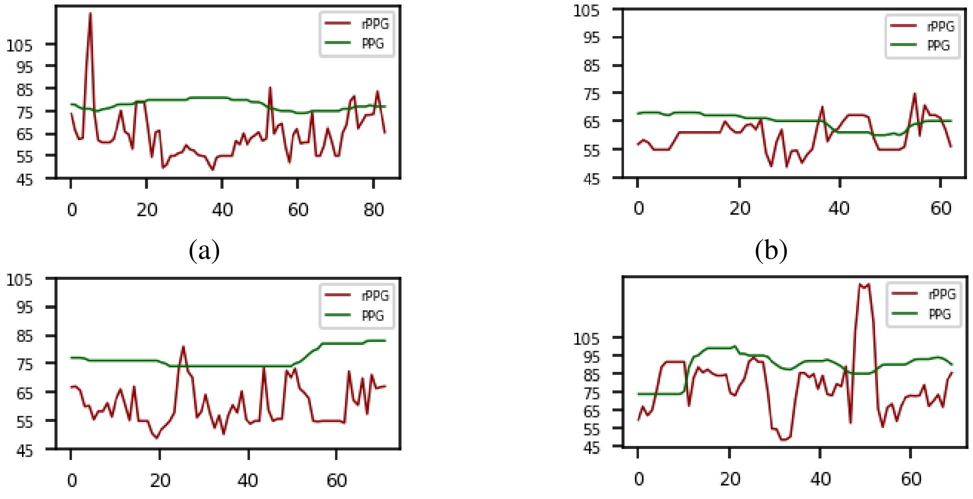
In the LGI-PPGI dataset, the talk scenario is one of the four existing scenarios. Videos of this scenario are recorded while subjects engage in a conversion while being outdoor.

In the following, we classify and present our results through a series of plots. For comparison purposes, each plot bears the reference PPG HR (assigned with a green color) and the computed rPPG HR (assigned with a red color).

For instance, in the "Haarcascade" subsection, there are four plots, and each plot is labeled with an alphabet letter associated with the candidate's name. As illustrated, we have applied the same classification concept for the results obtained using Dlib and MediaPipe.

Haarcascade

Figure 7.10 plots the computed HR (colored red - rPPG) against reference HR (colored green - PPG) for all the candidates with the Haarcascade face detector.



(a)

(b)

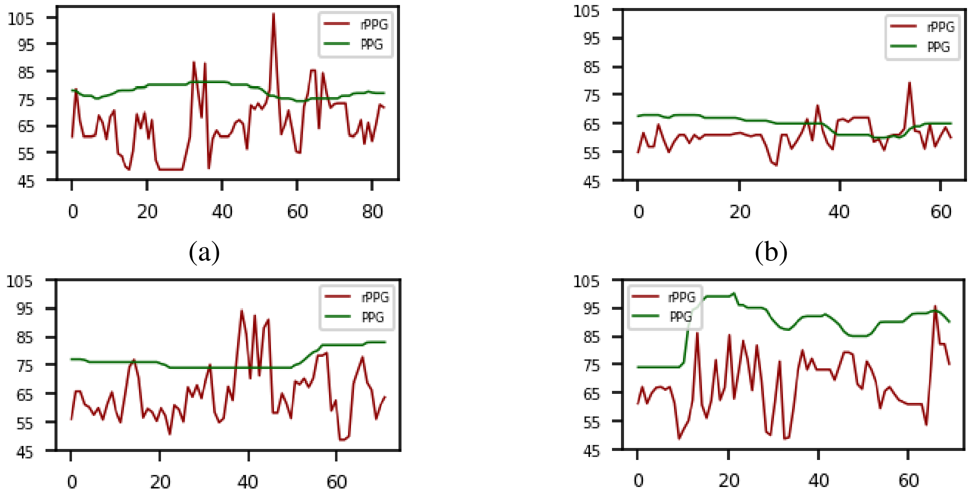
(c)

(d)

Figure 7.10: X-axis represents time (sec) and Y-axis represents HR (bpm). Plots labeled with a, b, c, and d are associated with candidate ID1, ID2, ID3, and ID4 respectively.

Dlib

Figure 7.11 plots the computed HR (colored red - rPPG) against the reference HR (colored green - PPG) for all candidates with the Dlib face detector.



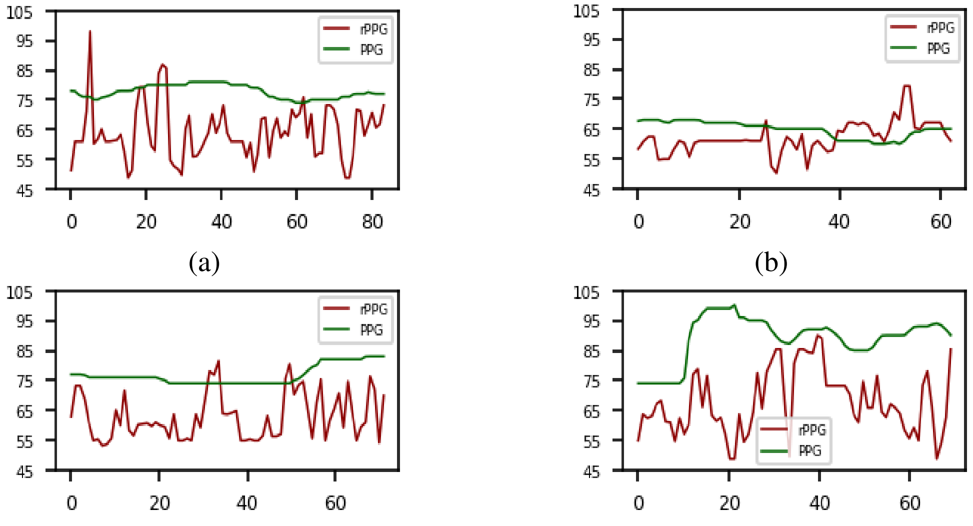
(a) (b)

(c) (d)

Figure 7.11: X-axis represents time (sec) and Y-axis represents HR (bpm). Plots labeled with a, b, c, and d are associated with candidate ID1, ID2, ID3, and ID4 respectively.

MediaPipe

Figure 7.12 plots the computed HR (colored red - rPPG) against the reference HR (colored green - PPG) for all candidates with the MediaPipe face detector.



(a) (b)

(c) (d)

Figure 7.12: X-axis represents time (sec) and Y-axis represents HR (bpm). Plots labeled with a, b, c, and d are associated with candidate ID1, ID2, ID3, and ID4 respectively.

Performance summary

Table 7.4 summarizes the metrics obtained for all candidates in the talking scenario. Overall, the lowest MAE and MSE values are achieved when the face detector is set to Dlib. With an average MAE of 13.84 and MSE of 287.25, Dlib outperforms the other face detectors and contributes to a better signal extraction.

Talk				
Candidate	Metric	Haarcascade	Dlib	MediaPipe
Id-1	MAE	15.29	14.37	14.62
	MSE	310.32	286.91	276.99
	r	-0.31	-0.31	-0.08
Id-2	MAE	6.65	5.97	6.30
	MSE	59.81	48.08	54.71
	r	-0.09	-0.29	-0.49
Id-3	MAE	16.43	13.73	14.37
	MSE	320.02	243.44	250.46
	r	0.04	-0.08	0.15
Id-4	MAE	18.47	21.31	21.50
	MSE	466.46	570.60	612.76
	r	-0.05	0.23	0.11
Overall	MAE	14.21	13.84	14.19
	MSE	289.15	287.25	298.73
	r	-0.10	-0.11	-0.08

Table 7.4: Computed evaluation metrics for candidates in talking mode.

In summary, our findings indicate that this scenario poses significant challenges due to the high levels of noise present in the videos.

Despite these difficulties, we found that the predicted HR for candidate ID2 is relatively closer to the reference HR. This can be attributed to the minimal movement of the candidate and less shaky video footage. However, this trend cannot be extended to other candidates due to the presence of various issues such as shaky videos, unstable lighting conditions, and excessive noise caused by the candidates themselves. These factors make it challenging to make further assumptions about the results.

7.2.5 Overall

To evaluate our results, we utilized the average values for both predicted and GT HR, as well as their corresponding MAE, as benchmarking metrics.

Averaged HRs

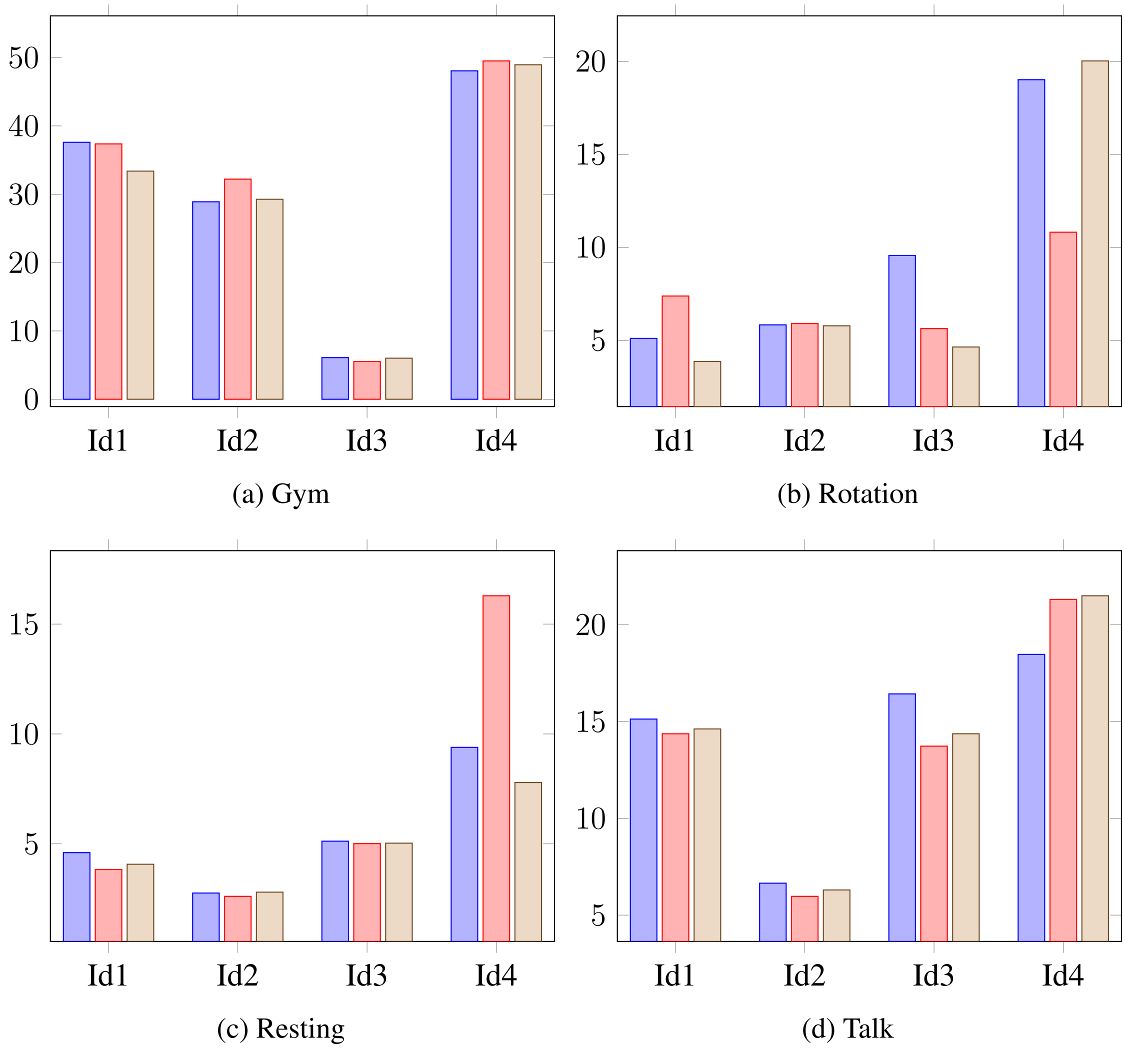
Table 7.5 articulates the average HRs predicted by the Equation 6.2, for all the candidates in all the segments of the dataset. The averaged values for each segment are only representative of the duration that measurements took place. The threshold column indicates the difference between the lower and upper averaged HR range for all face detectors. The lower threshold is obtained by subtracting the lowest HR(rPPG) from the HR(PPG), and the upper threshold is obtained by subtracting the highest HR(rPPG) from the HR(PPG).

Gym						
Candidate	HR(PPG)	**HR(rPPG)**			**Threshold(bpm)**	
		Haarcascade	**Dlib**	**MediaPipe**	Lower	Upper
Id-1	113.32	78.74	78.79	83.13	-34.58	-30.19
Id-2	106.90	81.79	78.06	81.04	-28.84	-25.11
Id-3	113.80	113.54	114.48	113.83	-0.26	+0.68
Id-4	117.22	70.83	69.33	70.18	-47.89	-46.30
Resting						
Candidate	HR(PPG)	**HR(rPPG)**			**Threshold(bpm)**	
		Haarcascade	**Dlib**	**MediaPipe**	Lower	Upper
Id-1	68.39	67.51	66.44	66.99	-1.95	-0.88
Id-2	62.33	60.81	60.86	60.65	-1.68	-1.47
Id-3	65.94	61.24	61.35	61.33	-4.7	-4.59
Id-4	80.40	75.39	82.95	77.52	-5.01	+2.55
Rotation						
Candidate	HR(PPG)	**HR(rPPG)**			**Threshold(bpm)**	
		Haarcascade	**Dlib**	**MediaPipe**	Lower	Upper
Id-1	69.02	66.08	68.49	67.53	-2.94	-0.12
Id-2	59.57	57.51	57.59	57.53	-2.06	-1.98
Id-3	66.14	68.15	60.49	61.48	-5.65	+2.01
Id-4	85.84	70.45	76.59	69.10	-16.74	-9.25
Talk						
Candidate	HR(PPG)	**HR(rPPG)**			**Threshold(bpm)**	
		Haarcascade	**Dlib**	**MediaPipe**	Lower	Upper
Id-1	77.49	64.54	65.47	64.07	-13.42	-12.02
Id-2	64.91	60.21	61.01	62.07	-4.7	-2.84
Id-3	77.56	60.47	65.47	62.96	-17.09	-12.09
Id-4	88.55	79.40	67.86	67.63	-20.92	-9.15

Table 7.5: The highlighted cells point out the supremacy of the face detector in the modular architecture, each segment, and for each candidate.

Averaged MAE

In this section, we used the MAE values from Tables 7.1, 7.2, 7.3, and 7.4 to visualize MAE for all the candidates and utilized face detectors in each scenario. Figure 7.13 represents a comparative overview of the averaged MAE of the utilized face detectors (Haarcascade, Dlib, and Mediapipe) for gym, rotation, resting, and talk scenarios. The X-axis of every chart is labeled with the dataset candidate's ID number and hosts three columns. Each column encompasses three bars for each of the face detectors. The height of each bar represents the MAE of the predicted HR for the corresponding face detector and candidate.



(a) Gym

(b) Rotation

(c) Resting

(d) Talk

Figure 7.13: The bars in purple are associated with the results obtained from the Haarcascade face detector. The bars in red are associated with the Dlib face detector. The bars in beige are associated with the MediaPipe face detector.

Figure 7.13 provides a clear visual representation of the performance of the different face detectors in terms of MAE between the predicted and GT HRs for each candidate and in each scenario.

Table 7.6 summarizes the overall MAE of the rPPG modules with various face detectors. By averaging the MAE values in each segment and for all candidates, we tried to highlight the average parity between each predicted unit against its observed peer.

Overall-MAE			
Scenario	Haarcascade	Dlib	MediaPipe
Gym	30.16	31.15	29.04
Resting	5.46	6.93	4.92
Rotation	9.87	7.43	8.57
Talk	14.21	13.84	14.19

Table 7.6: Overall MAE table for all the noisy scenarios.

As demonstrated, for the gym and resting scenarios, Figures (7.13a and 7.13c), the MediaPipe face detector leads to the lowest overall MAE, followed by Haarcascade and Dlib. For the rotation and talk scenarios, Figures (7.13b and 7.13d), the Dlib face detector leads to the lowest overall MAE, followed by MediaPipe and Haarcascade.

Final assessment

As a final assessment, we examined the number of highlighted cells for the selected face detectors in Table 7.5. We found that the rPPG method using the Dlib face detector is the most successful approach. In 8 out of 16 cases, or 50% of all categories and candidates, the average HR predicted through the rPPG-Dlib architecture is either in line or very close to the average reference HR(PPG). The rPPG modules equipped with the Haarcascade and MediaPipe face detectors make up 31.25% and 18.75% of plausible results, respectively. From this perspective, the dominance of the Viola-Jones algorithm over MediaPipe's state-of-the-art architecture is noteworthy.

The above assessments are aligned with our initial motivations and research objectives. The face detection technique that a rPPG study opts for, can be the solution to curb the effect of disturbances during the extraction of the rPPG signal

7.3 Limitations

Our study provides valuable information on rPPG and camera-based HR extraction in noisy scenarios. We also consider the importance of acknowledging certain limitations of the study. These limitations are crucial considerations when interpreting the findings, as they may have an impact on the overall assumption.

These limitations are as follows:

- The small size of the dataset used in the study means that the results are only applicable to the four candidates and four noisy scenarios that were analyzed.

- Using only the RGB channel to extract raw signals from the videos in the dataset means that the results obtained through other color channels are still unknown.

- During our evaluations, we only measured the performance of three face detectors, Haarcascade, Dlib, and Mediapipe, leaving many other face detection approaches untested.

- Another limitation of our study relates to the environment in which the videos were captured. Given the fact that the environments were not homogeneous, we had difficulty in generalizing our results to other scenarios and candidates. Therefore, our results are only applicable to the specific candidate and the scenario in question.

- The evaluation metrics that assisted us in the process of assessing the performance of the proposed architecture are limited to MAE, MSE, and Pearson correlation. Other metrics, such as the Root Mean Square of Successive Differences (RMSSD), were not included in the evaluation of the results.

- Considering that we only acquired the results from the forehead of the candidates, the performance of our proposed architecture against other facial regions remains unknown.

Overall, due to the specific nature of the dataset, we observed a lack of generalizability of the results, and the findings of this study may not be applicable to other datasets or populations.

Chapter 8

Conclusion and Future Work

In this chapter, we present the conclusions that we have drawn from the analysis and experimentation conducted throughout the thesis. In the following, we will elaborate and summarize our conclusions. Furthermore, we will also identify areas for future research that are prompted by the limitations of the current study.

8.1 Conclusion

In this thesis, our primary objective was to investigate the impact of noisy environments on the accuracy of camera-based HR extraction. To do this, we used a publicly available dataset, the LGI-PPGI-DB dataset, in order to adhere to ethical considerations.

Through a literature review, we answered the first RQ of our study. We identified the most commonly reported factors that impede or challenge accurate rPPG estimation in both controlled and uncontrolled settings, such as video compression algorithms, participant skin pigmentation, digitization, scene illumination, motion, and camera parameters.

In response to the second RQ, and to further examine the impact of the above-mentioned challenges, we developed a modular rPPG signal extraction pipeline that can be adapted to work with any face detection algorithm. We then evaluated our pipeline using three different face detection algorithms on the LGI dataset and compared our results with the PPG reference data provided.

In chapter 6.9.3 we presented our insights concerning the measures that improve the quality of extracted data (the second RQ), and the extent of disturbance's influence on estimated HR (the third RQ).

We also evaluated the overall performance of the selected face detectors in all four scenarios, using metrics such as MAE, MSE, and Pearson correlation. We found that Dlib face detection has the potential to be utilized as the ultimate signal extraction technique.

The concluding remarks of this thesis are as follows:

- Retrofitting a face detector that features facial landmarks proved to be an effective solution in curbing the effect of environment disturbances. As demonstrated, a RoI derived from a fixed facial landmark instantly enhances the quality of the extracted raw signal, especially if the predefined RoIs are located around moving facial attributes such as cheeks, eyes, and forehead.

- The subject's acute motion was found to be against the RoI tracking feature at certain times, thus disrupting the acquisition of the rPPG signal. The resilience of the Dlib and MediaPipe face detectors against acute head motion and rotation resulted in better and more consistent rPPG signal extraction. The predicted HRs obtained through these face detectors, fared better in comparison to the Haarcascade.

- The analysis of results exposed the extent by which subject motion influences the rPPG signal. Small and consistent subject movements were not as troubling as sudden and sharp head movements, but were noticeable in the predicted HRs.

- Unstable environment illumination also proved to be a consistent deterioration factor as it continued to manipulate the collected RGB traces, but interestingly, it was not the only illumination-related irritating factor. The illumination color temperature seemed to have an extensive impact on the relevancy of predicted HRs. In scenarios with warm color temperatures, predicted HRs were closer to the reference HR compared to those with a cooler illumination color temperature.

8.2 Future work

Given the sensitivity of rPPG measurements, working with far too noisy or corrupted datasets jeopardizes any possible breakthroughs. To overcome such obstacles, it would be better to develop a lab-controlled realistic dataset whose fluctuating variables do not constantly intervene with each other.

The examination of individual factors that undermine rPPG estimation provides a solid basis for introducing offsetting measures. Once this set of measures is found, they can be utilized in a scenario with a mixture of challenging factors to examine their effectiveness.

In this study, we observed a disparity between the discovered and predicted HRs with candidates whose environment illumination color temperature was tilted toward shades of cool white. Unfortunately, given the size of the dataset, we could not draw any conclusions. Interestingly, we were also unable to find any study on the effect of illumination color temperature in rPPG estimation. In that sense, a sole study on the effect of various color temperatures on rPPG approaches can be another possible future opportunity.

While the majority of rPPG studies are too focused on obtaining accurate HR, studying the effect of motion artifacts may not be a bad idea. An extension of this thesis would be the extraction of signals stemming from the subject's motion and finding a correlation between motion and predicted HRs to neutralize or minimize the effect of motion artifacts.

In recent years, transformers have become a very popular research area in the field of AI. A vision transformer is a variant of the transformer architecture that is adapted for CV tasks, such as image classification, object detection, and segmentation. A state-of-the-art rPPG study would be the evaluation of a vision transformer for remote camera-based HR estimation in noisy environments.

Aside from the above-mentioned possibilities, there are also future research prospects that can address the shortcomings of current study. Considering that our sample size was very small, a suitable further study would be using a larger dataset that enables the generalization of the results and also provides a wider range of conditions.

The exploration of alternative color channels in rPPG could potentially enhance the precision of pulse rate estimation. For example, a comprehensive investigation into the impact of color channels on rPPG estimation in noisy environments could be conducted as a separate study.

Another appropriate study would be using additional face detectors to validate the results and identify the most suitable face detector for rPPG. Further research into the potential of other facial regions for rPPG is also necessary to expand our understanding of this non-contact method for HR estimation in noisy environments.

Acknowledgments

Embarking on the biomedical imaging master's program was a dream come true, but the journey was not without its challenges. There were moments when I felt overwhelmed by the hours spent staring at a computer screen and the isolation of remote learning. However, the unwavering support and encouragement of those around me helped me to persevere and push through the difficult times. I am forever grateful for their belief in me and my abilities. Despite the challenges, this journey has been an incredibly rewarding and fulfilling experience that I am proud to have completed.

First, I would like to express my sincerest gratitude to my first supervisor, Professor Iman Azimi for providing me the opportunity to conduct my thesis on such an interesting topic and mentoring me throughout the process.

I would also like to thank Professor Antti Airola for accepting to be the second supervisor months passed into the thesis and his support upon the completion of thesis.

I am also grateful for my parent's unconditional love and support throughout the years.

Last but not least, I wish to express the deepest of appreciations to my wife, Nastaran Agharokh Shahenayati, for her unconditional love, encouragement, and support throughout my life and education. During rough times, she cheered me, motivated me and sparkled brightness onto my life. She believed in me when I was in the most fragile periods, and for that, I am forever in her debt.

Turku December 2022

Amin Rezaei

Appendix A

First Appendix

Python programming language versions 3.7.0 and 3.10.4 were used to implement the code.

Below is the list of the most important libraries used for the experiment:

Numpy 1.22.3
Scipy 1.8.0
Pandas 1.4.2
Matplotlib 3.5.2
Scikit-learn 1.0.2
Opencv-python 4.5.5.64
Dlib 19.23.1
Mediapipe 0.8.9.1

Bibliography

[1] Mairbäurl, Heimo, and Roy E. Weber. "Oxygen transport by hemoglobin." Comprehensive Physiology 2.2 (2011): 1463-1489.

[2] Castaneda, Denisse, et al. "A review on wearable photoplethysmography sensors and their potential future applications in health care." International journal of biosensors & bioelectronics 4.4 (2018): 195.

[3] Heikenfeld, Jajack, et al. "Wearable sensors: modalities, challenges, and prospects." Lab on a Chip 18.2 (2018): 217-248.

[4] Boonya-Ananta, Tananant, et al. "Monte Carlo analysis of optical heart rate sensors in commercial wearables: the effect of skin tone and obesity on the photoplethysmography (PPG) signal." Biomedical Optics Express 12.12 (2021): 7445-7457.

[5] Elgendi, Mohamed. "On the analysis of fingertip photoplethysmogram signals." Current cardiology reviews 8.1 (2012): 14-25.

[6] Verkruysse, Wim, Lars O. Svaasand, and J. Stuart Nelson. "Remote plethysmographic imaging using ambient light." Optics express 16.26 (2008): 21434-21445.

[7] Lyon, Richard F. "A brief history of'pixel'." Digital Photography II. Vol. 6069. SPIE, 2006.

[8] SUNDARARAJAN, D. “Color Image Processing.” Digital Image Processing: A Signal Processing and Algorithmic Approach, SPRINGER, 2017, pp. 1-158.

[9] SUNDARARAJAN, D. “Color Image Processing.” Digital Image Processing: A Signal Processing and Algorithmic Approach, SPRINGER, 2017, pp. 407–434.

[10] Nixon, Mark, and Alberto Aguado. Feature extraction and image processing for computer vision. Academic press, 2019.

[11] Ibraheem, Noor A., et al. "Understanding color models: a review." ARPN Journal of science and technology 2.3 (2012): 265-275.

[12] Ford, Adrian, and Alan Roberts. "Colour space conversions." Westminster University, London 1998 (1998): 1-31.

[13] Puthusserypady, Sadasivan. Applied signal processing. Now Publishers, 2021.

[14] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.

[15] Gustafsson, Fredrik. "Determining the initial states in forward-backward filtering." IEEE Transactions on signal processing 44.4 (1996): 988-992.

[16] Kaniusas, Eugenijus. "Biomedical signals and sensors II." Biological and Medical 2015, pp, 1-25.

[17] Downey, Allen. Think DSP: digital signal processing in Python. " O'Reilly Media, Inc.", 2016, pp. 37-90.

[18] Proakis, John G. Digital signal processing: principles algorithms and applications. Pearson Education India, 2001, pp, 330-352.

[19] Lewis-Beck, Michael, Alan E. Bryman, and Tim Futing Liao. The Sage encyclopedia of social science research methods. Sage Publications, 2003, pp, 259-259.

[20] Smith, Steven W. "The scientist and engineer's guide to digital signal processing." (1997).

[21] Smith, Steven W. "The scientist and engineer's guide to digital signal processing." (1997), pp, 331-350.

[22] Subasi, Abdulhamit. Practical Machine Learning for Data Analysis Using Python. Academic Press, 2020, pp, 1-203.

[23] Jo, Taeho. Machine Learning Foundations. Springer International Publishing. `https://doi.org/10.1007/978-3-030-65900-4`, 2021. 3-22.

[24] Yu, Wei, et al. "Visualizing and comparing AlexNet and VGG using deconvolutional layers." Proceedings of the 33 rd International Conference on Machine Learning. 2016.

[25] Liu, Yiming, et al. "Motion-robust multimodal heart rate estimation using BCG fused remote-PPG with deep facial ROI tracker and pose constrained Kalman filter." IEEE Transactions on Instrumentation and Measurement 70 (2021): 1-15.

[26] Lee, Hyunwoo, Ayoung Cho, and Mincheol Whang. "Fusion method to estimate heart rate from facial videos based on RPPG and RBCG." Sensors 21.20 (2021): 6764.

[27] Poh, Ming-Zher, Daniel J. McDuff, and Rosalind W. Picard. "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation." Optics express 18.10 (2010): 10762-10774.

[28] Lugaresi, Camillo, et al. "Mediapipe: A framework for building perception pipelines. " arXiv preprint arXiv:1906.08172 (2019).

[29] Grishchenko, Ivan, and Valentin Bazarevsky. "Mediapipe holistic—simultaneous face, hand and pose prediction, on device." Retrieved June 15 (2020): 2021.

[30] Zhang, Fan, et al. "Mediapipe hands: On-device real-time hand tracking." arXiv preprint arXiv:2006.10214 (2020).

[31] Grishchenko, Ivan, et al. "Attention mesh: High-fidelity face mesh prediction in real-time." arXiv preprint arXiv:2006.10962 (2020).

[32] Viola, Paul, and Michael J. Jones. "Robust real-time face detection." International journal of computer vision 57.2 (2004): 137-154.

[33] Wikipedia contributors. "Kernel method." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 9 Oct. 2022. Web. 09 Jun. 2022.

[34] Byeong-Ho, K. A. N. G. "A Review on Image and Video processing." International Journal of Multimedia and Ubiquitous Engineering 2.2 (2007): 49.

[35] King, Davis E. "Dlib-ml: A machine learning toolkit." The Journal of Machine Learning Research 10 (2009): 1755-1758.

[36] 81 facial landmarks shape predictor. https://github.com/codeniko/shape_predictor_81_face_landmarks.git

[37] Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[38] King, Davis E. "Max-margin object detection." arXiv preprint arXiv:1502.00046 (2015).

[39] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). Vol. 1. Ieee, 2005.

[40] Felzenszwalb, Pedro F., et al. "Object detection with discriminatively trained part-based models." IEEE transactions on pattern analysis and machine intelligence 32.9 (2010): 1627-1645.

[41] Bazarevsky, Valentin, et al. "Blazeface: Sub-millisecond neural face detection on mobile gpus." arXiv preprint arXiv:1907.05047 (2019).

[42] Shafer, Steven A. "Using color to separate reflection components." Color Research & Application 10.4 (1985): 210-218.

[43] Lv, Jiangjing, et al. "A deep regression architecture with two-stage re-initialization for high performance facial landmark detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[44] Kartynnik, Yury, et al. "Real-time facial surface geometry from monocular video on mobile GPUs." arXiv preprint arXiv:1907.06724 (2019).

[45] https://github.com/google/mediapipe/blob/master/mediapipe/modules/face_geometry/data/canonical_face_model_uv_visualization.png

[46] Da Costa, German. "Optical remote sensing of heartbeats." Optics communications 117.5-6 (1995): 395-398.

[47] Puri, Colin, et al. "StressCam: non-contact measurement of users' emotional states through thermal imaging." CHI'05 extended abstracts on Human factors in computing systems. 2005.

[48] Fei, Jin, and Ioannis Pavlidis. "Thermistor at a distance: unobtrusive measurement of breathing." IEEE transactions on biomedical engineering 57.4 (2009): 988-998.

[49] Lewandowska, Magdalena, et al. "Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity." 2011 federated conference on computer science and information systems (FedCSIS). IEEE, 2011.

[50] Takano, Chihiro, and Yuji Ohta. "Heart rate measurement based on a time-lapse image." Medical engineering & physics 29.8 (2007): 853-857.

[51] Garbey, Marc, et al. "Contact-free measurement of cardiac pulse based on the analysis of thermal imagery." IEEE transactions on Biomedical Engineering 54.8 (2007): 1418-1426.

[52] Humphreys, Kenneth, Tomas Ward, and Charles Markham. "Noncontact simultaneous dual wavelength photoplethysmography: a further step toward noncontact pulse oximetry." Review of scientific instruments 78.4 (2007): 044304.

[53] Gush, R. J., and T. A. King. "Discrimination of capillary and arterio-venular blood flow in skin by laser Doppler flowmetry." Medical and Biological Engineering and Computing 29.4 (1991): 387-392.

[54] Poh, Ming-Zher, Daniel J. McDuff, and Rosalind W. Picard. "Advancements in non-contact, multiparameter physiological measurements using a webcam." IEEE transactions on biomedical engineering 58.1 (2010): 7-11.

[55] De Haan, Gerard, and Vincent Jeanne. "Robust pulse rate from chrominance-based rPPG." IEEE Transactions on Biomedical Engineering 60.10 (2013): 2878-2886.

[56] De Haan, Gerard, and Arno Van Leest. "Improved motion robustness of remote-PPG by using the blood volume pulse signature." Physiological measurement 35.9 (2014): 1913.

[57] Suzuki, Satoshi, et al. "Development of non-contact monitoring system of heart rate variability (hrv)-an approach of remote sensing for ubiquitous technology." International Conference on Ergonomics and Health Aspects of Work with Computers. Springer, Berlin, Heidelberg, 2009.

[58] Lu, Guohua, et al. "Contact-free measurement of heart rate variability via a microwave sensor." Sensors 9.12 (2009): 9572-9581.

[59] Poh, Ming-Zher, Daniel J. McDuff, and Rosalind W. Picard. "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation." Optics express 18.10 (2010): 10762-10774.

[60] Purtov, Konstantin, et al. "Remote photoplethysmography application to the analysis of time-frequency changes of human heart rate variability." 2016 18th Conference of Open Innovations Association and Seminar on Information Security and Protection of Information Technology (FRUCT-ISPIT). IEEE, 2016.

[61] McDuff, Daniel, Sarah Gontarek, and Rosalind W. Picard. "Remote detection of photoplethysmographic systolic and diastolic peaks using a digital camera." IEEE Transactions on Biomedical Engineering 61.12 (2014): 2948-2954.

[62] Haque, Mohammad A., et al. "Heartbeat rate measurement from facial video." IEEE Intelligent Systems 31.3 (2016): 40-48.

[63] Banitsas, K., et al. "A simple algorithm to monitor hr for real time treatment applications." 2009 9th International Conference on Information Technology and Applications in Biomedicine. IEEE, 2009.

[64] Scully, Christopher G., et al. "Physiological parameter monitoring from optical recordings with a mobile phone." IEEE Transactions on Biomedical Engineering 59.2 (2011): 303-306.

[65] Kwon, Sungjun, Hyunseok Kim, and Kwang Suk Park. "Validation of heart rate extraction using video imaging on a built-in camera system of a smartphone." 2012 annual international conference of the IEEE engineering in medicine and biology society. IEEE, 2012.

[66] Jonathan, E., and Martin Leahy. "Investigating a smartphone imaging unit for photoplethysmography." Physiological measurement 31.11 (2010): N79.

[67] Wang, Wenjin, et al. "Algorithmic principles of remote PPG." IEEE Transactions on Biomedical Engineering 64.7 (2016): 1479-1491.

[68] Soleymani, Mohammad, et al. "A multimodal database for affect recognition and implicit tagging." IEEE transactions on affective computing 3.1 (2011): 42-55.

[69] MAHNOB-HCI dataset user manual link `https://mahnob-db.eu/hci-tagging/media/uploads/manual.pdf`

[70] MAHNOB-HCI dataset access link `https://mahnob-db.eu/`

[71] Heusch, Guillaume, André Anjos, and Sébastien Marcel. "A reproducible study on remote heart rate measurement." arXiv preprint arXiv:1709.00962 (2017).

[72] COHFACE dataset access link `https://zenodo.org/record/4081054#.Y71-0nZBxD8`

[73] Boccignone, Giuseppe, et al. "An open framework for remote-PPG methods and their assessment." IEEE Access 8 (2020): 216083-216103.

[74] Pilz, Christian S., et al. "Local group invariance for heart rate estimation from face videos in the wild." Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2018.

[75] LGI-PPGI-DB dataset access link `https://github.com/partofthestars/LGI-PPGI-DB`

[76] Bobbia, Serge, et al. "Unsupervised skin tissue segmentation for remote photoplethysmography." Pattern Recognition Letters 124 (2019): 82-90.

[77] dataset access link https://sites.google.com/view/ybenezeth/ubfcrppg

[78] Stricker, Ronny, Steffen Müller, and Horst-Michael Gross. "Non-contact video-based pulse rate measurement on a mobile service robot." The 23rd IEEE International Symposium on Robot and Human Interactive Communication. IEEE, 2014.

[79] PURE dataset access link `https://www.tu-ilmenau.de/universitaet/fakultaeten/fakultaet-informatik-und-automatisierung/profil/institute-und-fachgebiete/institut-fuer-technische-informatik-und-ingenieurinformatik/fachgebiet-neuroinformatik-und-kognitive-robotik/data-sets-code`

[80] Li, Xiaobai, et al. "The OBF database: A large face video database for remote physiological signal measurement and atrial fibrillation detection." 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, 2018.

[81] Liu, Siqi, et al. "3D mask face anti-spoofing with remote photoplethysmography." European Conference on Computer Vision. Springer, Cham, 2016.

[82] Liu, Siqi, et al. "A 3D mask face anti-spoofing database with real world variations." Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2016.

[83] Niu, Xuesong, et al. "VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video." Asian conference on computer vision. Springer, Cham, 2018.

[84] Kopeliovich, M., M. Petrushan, and D. Shaposhnikov. "Approximation-based transformation of color signal for heart rate estimation with a webcam." Pattern Recognition and Image Analysis 28.4 (2018): 646-651.

[85] Zhang, Zheng, et al. "Multimodal spontaneous emotion corpus for human behavior analysis." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[86] Dasari, Ananyananda, et al. "Evaluation of biases in remote photoplethysmography methods." NPJ digital medicine 4.1 (2021): 1-13.

[87] Kim, Dae-Yeol, Kwangkee Lee, and Chae-Bong Sohn. "Assessment of ROI Selection for Facial Video-Based rPPG." Sensors 21.23 (2021): 7923.

[88] Wang, Wenjin, Sander Stuijk, and Gerard De Haan. "A novel algorithm for remote photoplethysmography: Spatial subspace rotation." IEEE transactions on biomedical engineering 63.9 (2015): 1974-1984.

[89] Martinez, Luis F. Corral, Gonzalo Paez, and Marija Strojnik. "Optimal wavelength selection for noncontact reflection photoplethysmography." 22nd Congress of the International Commission for Optics: Light for the Development of the World. Vol. 8011. SPIE, 2011.

[90] Tarassenko, Lionel, et al. "Non-contact video-based vital sign monitoring using ambient light and auto-regressive models." Physiological measurement 35.5 (2014): 807.

[91] Clifford, G. D. "Blind source separation: principal & independent component analysis." Biomedical Signal and Image Processing (2008): 1-47.

[92] Lewandowska, Magdalena, et al. "Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity." 2011 federated conference on computer science and information systems (FedCSIS). IEEE, 2011.

[93] Hyvärinen, Aapo, Juha Karhunen, and Erkki Oja. Independent Component Analysis. New York: Wiley, 2001. Print.

[94] Hyvärinen, Aapo, and Erkki Oja. "Independent component analysis: algorithms and applications." Neural networks 13.4-5 (2000): 411-430.

[95] Clifford, Gari D.. “Chapter 15 - BLIND SOURCE SEPARATION: Principal & Independent Component Analysis.” (2005).

[96] Cardoso, Jean-François, and Antoine Souloumiac. "Blind beamforming for non-Gaussian signals." IEE proceedings F (radar and signal processing). Vol. 140. No. 6. IET Digital Library, 1993.

[97] Hyvärinen, Aapo, and Erkki Oja. "A fast fixed-point algorithm for independent component analysis." Neural computation 9.7 (1997): 1483-1492.

[98] Chen, W., McDuff, D.: Deepphys: Video-based physiological measurement using convolutional attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 349–365 (2018)

[99] Yu, Zitong, et al. "Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[100] Špetlík, Radim, Vojtech Franc, and Jirí Matas. "Visual heart rate estimation with convolutional neural network." Proceedings of the british machine vision conference, Newcastle, UK. 2018.

[101] Bousefsaf, Frédéric, Alain Pruski, and Choubeila Maaoui. "3D convolutional neural networks for remote pulse rate measurement and mapping from facial video." Applied Sciences 9.20 (2019): 4364.

[102] Lee, Eugene, Evan Chen, and Chen-Yi Lee. "Meta-rppg: Remote heart rate estimation using a transductive meta-learner." European Conference on Computer Vision. Springer, Cham, 2020.

[103] Yu, Zitong, Xiaobai Li, and Guoying Zhao. "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks." arXiv preprint arXiv:1905.02419 (2019).

[104] Liu, Si-Qi, and Pong C. Yuen. "A general remote photoplethysmography estimator with spatiotemporal convolutional network." 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). IEEE, 2020.

[105] Reiss, Attila, et al. "Deep PPG: Large-scale heart rate estimation with convolutional neural networks." Sensors 19.14 (2019): 3079.

[106] Luguev, Timur, Dominik Seuß, and Jens-Uwe Garbas. "Deep learning based affective sensing with remote photoplethysmography." 2020 54th Annual Conference on Information Sciences and Systems (CISS). IEEE, 2020.

[107] Paracchini, Marco, et al. "Biometric signals estimation using single photon camera and deep learning." Sensors 20.21 (2020): 6102.

[108] Casado, Constantino Alvarez, and Miguel Bordallo López. "Face2PPG: An unsupervised pipeline for blood volume pulse extraction from faces." arXiv preprint arXiv:2202.04101 (2022).

[109] McDuff, Daniel J., Ethan B. Blackford, and Justin R. Estepp. "The impact of video compression on remote cardiac pulse measurement using imaging photoplethysmography." 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017.

[110] Song, Rencheng, et al. "New insights on super-high resolution for video-based heart rate estimation with a semi-blind source separation method." Computers in biology and medicine 116 (2020): 103535.

[111] Kim, So-Eui, et al. "Restoration of remote PPG signal through correspondence with contact sensor signal." Sensors 21.17 (2021): 5910.

[112] Sun, Yu, et al. "Use of ambient light in remote photoplethysmographic systems: comparison between a high-performance camera and a low-cost webcam." Journal of biomedical optics 17.3 (2012): 037005.

[113] Wikipedia contributors. "Window function." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 20 Oct. 2022. Web. 22 Oct. 2022.

[114] Van Gastel, Mark, Sander Stuijk, and Gerard de Haan. "Motion robust remote-PPG in infrared." IEEE Transactions on Biomedical Engineering 62.5 (2015): 1425-1433.

[115] Spigulis, Janis. "Multispectral, fluorescent and photoplethysmographic imaging for remote skin assessment." Sensors 17.5 (2017): 1165.

[116] Li, Xiaobai, et al. "Remote heart rate measurement from face videos under realistic situations." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[117] Cheng, Chun-Hong, et al. "Deep learning methods for remote heart rate measurement: a review and future research agenda." Sensors 21.18 (2021): 6296.

[118] Sejdic, Ervin, and Tiago H. Falk, eds. Signal Processing and Machine Learning for Biomedical Big Data. CRC press, 2018. 3-7.

[119] Bland, J. Martin, and Douglas G. Altman. "Statistics notes: measurement error." Bmj 312.7047 (1996): 1654.

[120] Cacciatore, M., & Yeo, S. (2017). Standard score. In M. Allen (Ed.), The sage encyclopedia of communication research methods (pp. 1673-1675). SAGE Publications, Inc, https://dx.doi.org/10.4135/9781483381411.n589

[121] Jacobs, Harold R. Mathematics: A human endeavor. Macmillan, 1994.

[122] Feng, Yuantao, et al. "Detect Faces Efficiently: A Survey and Evaluations." IEEE Transactions on Biometrics, Behavior, and Identity Science 4.1 (2021): 1-18.

[123] Deng, Yunbin, and Arya Kumar. "Standoff heart rate estimation from video: a review." Mobile Multimedia/Image Processing, Security, and Applications 2020 11399 (2020): 16-29.

[124] Ahonen, Timo, Abdenour Hadid, and Matti Pietikäinen. "Face recognition with local binary patterns." European conference on computer vision. Springer, Berlin, Heidelberg, 2004.

[125] Jia, Yangqing, et al. "Caffe: Convolutional architecture for fast feature embedding." Proceedings of the 22nd ACM international conference on Multimedia. 2014.

[126] Zhang, Kaipeng, et al. "Joint face detection and alignment using multitask cascaded convolutional networks." IEEE signal processing letters 23.10 (2016): 1499-1503.

[127] Lempe, Georg, et al. "ROI selection for remote photoplethysmography." Bildverarbeitung für die Medizin 2013. Springer, Berlin, Heidelberg, 2013. 99-103.

[128] Face vectors Image by Vectonauta on Freepik.com `https://www.freepik.com/free-vector/beauty-face-chart-with-man-face-drawing_20903388.htm#query=face&position=34&from_view=search&track=sph`

[129] Horton, Michael, Mike Cameron-Jones, and Raymond Williams. "Multiple classifier object detection with confidence measures." Australasian Joint Conference on Artificial Intelligence. Springer, Berlin, Heidelberg, 2007. 578-587.

[130] Rosebrock, Adrian. “OpenCV Haar Cascades.” PyImageSearch, 17 Apr. 2021, pyimagesearch.com/2021/04/12/opencv-haar-cascades.

[131] Rahmad, C., et al. "Comparison of Viola-Jones Haar Cascade classifier and histogram of oriented gradients (HOG) for face detection." IOP conference series: materials science and engineering. Vol. 732. No. 1. IOP Publishing, 2020.