UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

POSTGRADUATE PROGRAM ON ELECTRICAL ENGINEERING

LUCAS CÔGO LAMPIER

Supervisor: PhD. Teodiano Freire Bastos Filho

EVALUATION OF AN ONLINE REMOTE PHOTOPLETHYSMOGRAPHY METHODOLOGY FOR EMOTION RECOGNITION IN A CHILD-ROBOT INTERACTION

Vitória, ES
2020

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
POSTGRADUATE PROGRAM ON ELECTRICAL ENGINEERING

LUCAS CÔGO LAMPIER

EVALUATION OF AN ONLINE REMOTE PHOTOPLETHYSMOGRAPHY METHODOLOGY FOR EMOTION RECOGNITION IN A CHILD-ROBOT INTERACTION

Master Dissertation presented to the Electrical Engineering Postgraduate Program at Universidade Federal do Espírito Santo (UFES) as a partial requirement to obtain the Electrical Engineering Master Degree.

Supervisor: PhD. Teodiano Freire Bastos Filho

Vitória, ES
2020

Ficha catalográfica disponibilizada pelo Sistema Integrado de Bibliotecas - SIBI/UFES e elaborada pelo autor

---

Côgo Lampier, Lucas, 1993-
C676e Evaluation of an online remote photoplethysmography methodology for emotion recognition in a child-robot interaction / Lucas Côgo Lampier. - 2020.
107 f. : il.

Orientador: Phd. Teodiano Freire Bastos Filho.
Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Autismo. 2. Aprendizado do computador. 3. Processamento de imagens. 4. Emoções. I. Freire Bastos Filho, Phd. Teodiano. II. Universidade Federal do Espírito Santo. Centro Tecnológico. III. Título.

CDU: 621.3

Lucas Côgo Lampier

EVALUATION OF AN ONLINE REMOTE PHOTOPLETHYSMOGRAPHY METHODOLOGY FOR EMOTION RECOGNITION IN A CHILD-ROBOT INTERACTION

Dissertação de Mestrado apresentada ao Curso de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Espírito Santo (UFES) como parte dos requisitos necessários para a obtenção do grau de Mestrado em Engenharia Elétrica.

Apresentada em Vitória, 28 de Abril de 2020.

Prof. Dr. Teodiano Freire Bastos Filho
PPGEE - UFES
Orientador

Alan Silva da Paz Floriano

Dr. Alan Silva da Paz Floriano
Instituto Federal do Espírito Santo - IFES - São Mateus
Examinador

Prof. Dr. Patrick Marques Ciarelli
Universidade Federal do Espírito Santo - UFES
Examinador

This work is dedicated to my parents, Ana and Irismar, to my girlfriend Gabriela, to my brother Danilo, and all the others that I had the pleasure to meet during the last two years

Acknowledgements

Apesar to texto estar em inglês, tomei a liberdade de escrever os agradecimentos em português para facilitar a compreensão das pessoas citadas neste texto.

Primeiramente, agradeço à minha família, que me deu todo o suporte possível durante toda a minha caminhada acadêmica, desde os primeiros garranchos na pré-escola até o momento da entrega deste trabalho. Para realizar a pesquisa de mestrado foram necessárias algumas noites em claro, alguns finais de semana sem participar de reuniões de família e muita paciência, tanto minha como de vocês. Ana, Irismar e Danilo, obrigado por todo o carinho e apoio emocional que vocês me deram.

Gostaria de agradecer também a minha namorada Gabriela, que apareceu de surpresa no final do curso de mestrado, ainda assim, me ajudou diretamente participando da pesquisa descrita neste texto, e também me dando muito carinho e compreensão durante a parte mais difícil do curso de mestrado, que foi a escrita deste documento.

Agradeço aos meus amigos Ives, Yuri, Rafael, Evandro, Otávio e Smith; a companhia de vocês deixou esses dois anos bem mais agradáveis. Agradeço também ao Dr. Alan, que me deu as primeiras direções durante a pesquisa, ao Yves por me ajudar com *machine learning* e também aos amigos que fiz durante minha estadia no NTA.

Obrigado também aos professores Dr. Teodiano, Dr. Denis e Dra. Eliete por me orientarem durante a pesquisa, me ajudando a corrigir os erros e mostrando os melhores caminhos para seguir.

Agradeço à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pela bolsa concedida, e, por fim, agradeço também à FAPES (Fundação de Amparo à Pesquisa e Inovação do Espírito Santo), e ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) pelo suporte financeiro que me permitiu obter os equipamentos necessários ao desenvolvimento desta Dissertação de Mestrado.

A todos, muito obrigado!

“Nothing in life is to be feared.
It is only to be understood.
Now is the time to understand more,
so that we may fear less.”
Marie Curie, Chemist and Physicist

Abstract

New Mobile Autonomous Robot for Interaction with Autistics (N-MARIA) is a robot built at UFES to help therapy with children with Autistic Spectrum Disorder (ASD). It has audiovisual communication equipment, a system allowing the therapist to send commands to the robot, and also an algorithm allowing to follow the child at a safe distance. Aiming to improve the *N-MARIA*'s Child-Robot-Interaction (CRI), this work proposes an online remote photoplethysmography (rPPG) to extract the pulse rate signal using a webcam. The obtained cardiac information is then used to train a classifier and infer the child's emotion during CRI. The results are presented to the therapist using a Graphic-User-Interface (GUI). The algorithm is projected to work online using a low-cost webcam. Different rPPG techniques presented in the literature are evaluated by precision and processing time and compared with ground truth, Electrocardiography (ECG) and Photoplethysmography (PPG). The results show that the error for heart rate measurement, while the subject is still in front of the camera, is relatively low (with a median error of 3 bpm), but fails in situations of fast movements (with a median error of 15 bpm). For emotion recognition, the prediction accuracy for three outputs were evaluated: arousal; valence; and six discrete emotional states. The classification accuracy results were better when using ECG to separate arousal, and rPPG for valence and the six discrete emotions. They also indicate that some emotional states may be separable using cardiac signals, however, the results presented an low overall result, which indicates that, using only cardiac signals results in a poor classification results, at least, in the way that they were tested in this work.

Key-words: Remote-photoplethysmography. Autism. Emotion Classification. Heart Rate. Heart Rate Variability.

List of Figures

List of Tables

List of Abbreviations and Acronyms

AI	Artificial Intelligence
ANS	Autonomic Nervous System
ASD	Autism Spectrum Disorder
BSS	Blind Source Separation
CRI	Child-Robot Interaction
DB	Data Base
ECG	Electrocardiogram
EEG	Electroencephalogram
FFT	Fast Fourier Transform
FLM	Facial landmarks
FPS	Frames per Second
GUI	Graphic User Interface
HOG	Histogram of Orient Gradients
ICA	Independent Component Analysis
LSP	Lomb-Scargle periodogram
MARIA	Mobile Autonomous Robot for Interaction with Autistics
NN	Normal-to-Normal
NB	Narrow Band
N-MARIA	New Mobile Autonomous Robot for Interaction with Autistics
OA	Overlap-Adding
PCA	Principal Component Analysis
PDD	Pervasive Developmental Disorder
PNS	Parasympathetic Nervous System
POS	Plane Orthogonal to Skin

PPG	Photoplethysmography
PSD	Power Spectral Density
RGB	Red, Green and Blue
ROI	Region of interest
RMSE	Root squared mean error
SAM	Self-Assessment Manikins
SB	Sub-Band
SNR	Signal to Noise Ratio
SNS	Sympathetic Nervous System
SR	Sample Rate
STD	Standard deviation
VJ	Viola-Jones
WP	Welch Periodogram

Contents

1 Introduction

1.1 Motivation

Autism is one of the best known pervasive developmental disorder (PDD) which is a group of neural development clinical conditions characterized by early onset of delays and deviations of various skills, mainly social and communicative (KLIN, 2006).

The first known article to describe autism as a particular condition is the work of Kanner, (1943). This work presents the study of case of eleven children with similar patterns, also describing some characteristic behaviors for autistics, as monotonous repetitions, particular interests and difficulty to social communication and interaction (COHMER, 2014). A benchmark in the autism identification was set in 1978, when Michael Rutter defined it based on the following criteria: delay and deviation on social skills; difficulty to communicate; unusual behaviors, as repetitive movements and mannerism; whose symptoms appear before the age of three (RUTTER, 1978; KLIN, 2006).

According to the United Nations (UN), there are about 70 million autistics in the world (2018), mostly boys (STOCK, 2018). Also, according to data from 2014 collected by the Centers for Disease Control and Prevention, in the United States of America, the prevalence of Autism Spectrum Disorder (ASD) was 16.8 per 1000 children under 9 years old (REDFIELD et al., 2014). Brazil does not have official statistics, but it is speculated that the number is close to 2 million autistics (STOCK, 2018).

In order to stimulate autistic children's attention, social skills and ability to interact with the environment, the *New Mobile Autonomous Robot for Interaction with Autistics (N-MARIA)* is being developed at UFES. The robot is made up of a *PIONNER 3-DX* as a mobile platform, a playful design to attracts child's attention, a tablet for audiovisual interaction as dynamic face, infrared and color cameras to infer emotions by face-expressions and face-temperatures, touch sensors and laser sensor to detect contact, locate the child and maintain a safe distance from he/she (GOULART et al., 2019b). The robot is 141 cm tall, close to a 9 years old child height, to make easier a face to face interaction. In (GOULART et al., 2019b) is reported the interaction of 36 children with *N-MARIA*. Before the child saw the robot for the first time, the child stayed relaxed for approximately 10 min. Then the robot was uncovered and the child could see and interact with it. The experiment was recorded by RGB cameras and a test was performed to evaluate social skills of the child, as tactile interaction, shared engagement, eye gaze and proximity, using data from cameras and touch sensors.

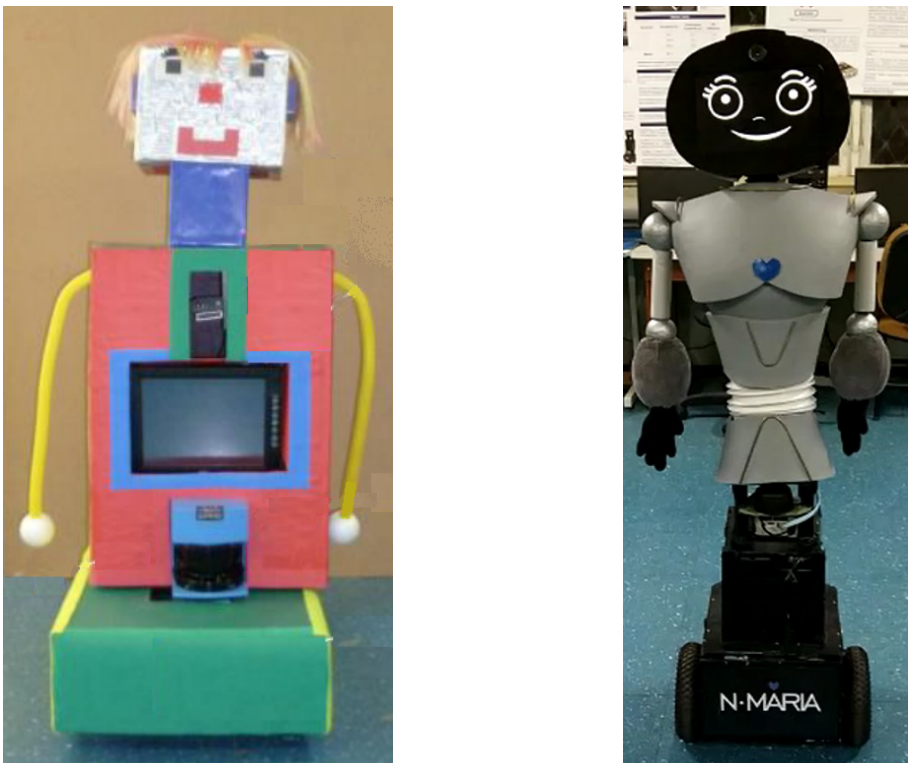
The test was divided into two parts. In the first, the robot presented itself through artificial voice and interacted with the child, asking about their name and hobbies. The second part was

the tactile interaction and shared engagement. During the experiments, the child was asked about the robot structure, how they liked the design and what they wanted to change in it. They were also asked to identify facial expressions played by the robot, which simulated emotional states (happiness, sadness, surprise, disgust, fear and anger). They achieved a success rate over 92% (GOULART et al., 2019a).

A previous version of the robot was also used to interact with autistic children. The experiment was conducted with four children, two autistics and two non-autistics, they interacted and played with the robot under the supervision of the mediator. The robot was programmed to follow the child at a safe distance, to play videos, or perform tricks, according to the mediator's command (GOULART et al., 2015).

The way of interacting with the robot changed from child to child. Some remained seated on the floor, others touched the robot and moved around it, and there were also those that showed some excitement. The results of the research conducted in (GOULART et al., 2015) show that the Child-Robot Interaction (CRI) was better than expectations. They touched and looked more closely at the robot and interacted with the mediator also more than expected. Only one autistic child was afraid to interact with the robot. The original (*MARIA*) and new (*N-MARIA*) versions of the robot are presented in Figure 1.1.

Figure 1.1 – Original and new version of the *MARIA* robot.



(a) Previous version of *MARIA*. (b) New *MARIA*.

Source – GOULART et al., 2015 and 2019a

From these previous studies, a feature that can be used to improve the CRI is the emotion

recognition, as the robot can identify a specific emotion in the child it may adapt its behavior to make the interaction more comfortable and enjoyable. In (GOULART et al., 2019a) was used a system for emotion recognition using thermal images. The system used two image sources: an RGB (Red, Green and Blue) camera to locate specific regions in the child's face, and an infrared thermal camera to capture the thermal values of these regions. In total, 28 children participated in the experiment, which consisted of children watching five different videos to excite different emotions (disgust, fear, happiness, sadness, surprise) and also a neutral video. The results were very promising, reaching an accuracy of 85.75% in recognizing these emotional states.

An advantage of the previous system for emotion recognizing during CRI is that it does not need any contact sensors to infer physiological information related to emotion, which makes the interaction much more comfortable. Another possible way to infer the child's emotional state while interacting with a robot is through heart rate analysis (APPELHANS; LUECKEN, 2006), which is the proposal of this dissertation.

1.2 Justification

According to Appelhans and Luecken (2006), emotions experienced by humans are related to changes in physiological arousal, with the autonomic nervous system (ANS) as an important actor at this regulation. The ANS is divided into the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS). The SNS is activated in stress situations, the heart rate frequency is set to increase, which prepares the individual for a challenge. The PNS has an opposite effect, it becomes dominant on relaxing and safe situations, and it sends signals to decrease and to stabilize the heart rate. Thus, the heart rate variation (HRV), according to the situation makes the cardiac beats an indicator of the person's emotional arousal level.

The traditional methods to detect heartbeats are electrocardiography (ECG), as the gold standard, and photoplethysmography (PPG). Both methods often use skin-contact sensors to perform the measurement (electrodes, in case of EEG, and optical devices to detect the blood volume variation under the skin, for PPG) (PASTORE et al., 2009; TASKFORCE, 1996; LU et al., 2008). However, getting cardiac information with traditional heart rate sensors is unsuitable to a CRI, as they are usually wired, which restrict the child's movement, and even the wireless options may get the children uncomfortable to use the device. However, in the last decade a number of methodologies to detect the heartbeats remotely using a common RGB camera were developed, this procedure is called remote photoplethysmography (rPPG) (KRANJEC et al., 2014).

The concept that most rPPG techniques use is very similar to the classical PPG: they measure the variation of blood volume under the skin. Among other factors, the optical properties of the skin are related to the blood volume under it. As the heartbeats lead to periodical blood volume variations, the optical properties of the skin also changes in the same frequency, which leads to small skin-color changes in the same frequency of the heartbeats. This color variation

can be recorded by an RGB camera and, then, the image frames can be processed to generate the pulse rate signal. Thus, this work proposes the implementation of an online rPPG sensing to be incorporated into the robot and used to get children's heart rate and, possibly, infer their emotions, without the need to attach any contact sensors to the child's skin. The system is conceived to work online, as the child interacts with the robot.

Some situations in a future CRI may decrease the accuracy of the system. The movement of the child difficulties the estimation of the rPPG, which decreases the quality of the signal, affecting the heart rate estimation and also the emotion estimation. Also, the movement increases the heart rate, which may confuse the emotion prediction system.

1.3 Objective

The main goal of this work is to build a system to estimate the heart rate using the rPPG and use the remote cardiac signal to estimate online the emotional state of the subject. To build the system, the quality and processing speed of the different steps of the rPPG methodologies are compared to select the most suitable ones. Also, it is tested if the rPPG signal is precise enough to identify the different emotions using the same protocol presented in (GOULART et al., 2019a).

1.4 Organization

This text is organized in six chapters:

- Chapter one - Introduction: this chapter introduces the motivation, justification and objective of the study.
- Chapter two - Theoretical Background: the second chapter provides the theoretical background of ASD, methodologies to extract the rPPG and emotion recognition based on cardiac parameters.
- Chapter three - Technique Evaluation: the third chapter presents different techniques for each step of the rPPG measurement and compares each of them to select the fastest and most accurate ones.
- Chapter four - Results: this chapter presents the errors of the heart rate estimation achieved using rPPG, also presents the accuracy of the emotion classifier using cardiac parameters extracted from both ECG and rPPG, comparing the accuracy of each one. At the end, it is shown the suitability of using the system online.
- Chapter five - Conclusion: The final chapter contains the final statements about the whole work and the author's personal considerations and suggestions for future researches.

2 Theoretical Background

2.1 Autism Spectrum Disorder

As autism may be manifested in a wide range of skills and abilities and also in different intensities, it is usually called Autism Spectrum Disorder (ASD). ASD is recognized by a loss in a range of social skills, as interpersonal relationship, communication difficulties, behavior and interests patterned and stereotyped. These conditions should manifest by the age of three. About 60% to 70% of the people with autism are considered with a mental disability, but this number is decreasing in recent studies, probably by earlier diagnostics and treatment, and better understanding this disorder (KLIN, 2006).

2.1.1 Historical Review

(KANNER, 1943) is the first work to describe autism as a particular condition, based on the case of eleven children (aged 2-8 years) with similar behaviors. This study described some features that are common for autistics, such as monotonous repetitions, particular interests, and communication and social interaction difficulties. It is also described conditions about the child's growing and living environment, It shows that the disorder appears in different economical and environment situations.

Rutter (1978) defined the identification protocol of autism as a disorder appearing before the age of 30 months, not necessarily associated with an intellectual disability or neurological dysfunction, marked by a particular social and language development, appearing with peculiar and repetitive activities, as stereotyped patterns, and resistance to change. The increasing number of works related to autism as a particular condition, together with Rutter's works, had an important role to add the disorder as a new class of conditions in the third edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-III): the Pervasive Developmental Disorder (KLIN, 2006).

2.1.2 Diagnosis

According to Klin (2006), the multiple behavior features characterizing the autism disorder are subdivided into three clusters: impaired social interaction; qualitative impairments in communication skills; and presence of restricted and repetitive patterns of behavior, interests, and activities. For an autistic diagnosis, the person must have at least six autistic behavior features, at least of one from each cluster (KLIN, 2006). The features from each cluster are presented below:

- The features inside the impaired social interaction cluster are: lack in the use of non-verbal

communication, as body and facial expressions or eye contact; not maintaining relationships with people and not sharing the own interests and experiences; lack of social and emotional reciprocity.

- The qualitative impairments in communication skills are composed of: difficulties at the verbal language development; difficulties to initiate and maintain a conversation with others, repetitive use of certain words or even peculiar sounds; lack of imaginative and imitative ability.

- The criteria of the last cluster, the presence of restricted and repetitive patterns of behavior, interests, and activities, are: intense and rigid preoccupations with restrict and stereotyped patterns of interests, inflexible personal routines, and rituals; gesture stereotypes (like jumping, or shake a body part); and particular interest on a specific part of objects.

2.1.3 Therapy

According to Lai et al. (2014), about 1% of the population is inside the autistic spectrum, mostly boys. Therapy, intervention, and support to autistics should be individualized, multidisciplinary and multidimensional. Therefore, behavioral and educational approaches maximize their quality of life, improving their communication, social skills, learning and independence. The therapy procedures include targeted skill-based intervention, structured teaching, parent-mediated intervention, and targeted behavioral intervention for anxiety and aggression (LAI et al., 2014). Drugs are rarely an approach, only considered when a patient presents associated conditions. For example, antipsychotics seem to reduce the repetitive behaviors in children with autism. However, they have potential side effects such as sedation, weight gain and involuntarily movements (LAI et al., 2014). Therefore, psychosocial and behavioral interventions are the main approaches used in autism, with the joint action of psychologists, neurologists, psychiatrists, pediatricians, and speech therapists.

2.1.4 Social Robots for ASD Therapy

The technology progress brought new tools at the autism intervention: the robots. Social robots may be a powerful support for ASD children, mainly for the ones with a special affection for technological systems. Robots has been already tested to assist ASD children in improving eye-contact, self-initiated interactions, imitations and emotion recognition (PENNISI et al., 2016).

The robots presented at the review work of Pennisi et al. (2016) covered a wide range of functions: they were used as a measurement tool to compare ASD and non-ASD children; some played a role as a playmate of the child; others were programmed to perform actions that were once played by the therapist, and also a tool that therapists could use to optimise their work. After analyzing the cases, Pennisi et al. (2016) claim that the use of robotics in therapy brought

positive results in most of the cases, showing that social robotics may increase the effectiveness of ASD therapy. Examples of robots for ASD therapy are presented bellow.

N-MARIA is a robot developed in UFES to be used as a therapist's toll in the ASD therapy. It is bigger than most robots used in this field, with 141 cm, the robot is as high as an 9 years old child, and it is assembled with a set of tools to interact with the child, as a tablet on the head to express emotions and also talk. A set of video, thermal and distance sensor are attached to the robot and used to analyze the child's behavior and control the robot behavior to follow he/she at a safe distance. The studies conducted with *N-MARIA* and its previous version showed that the children respond better to the mediator when the robot participates of the process (GOULART et al., 2015; GOULART et al., 2019b; GOULART et al., 2019a). *N-MARIA* is presented in Figure 1.1b.

According to Pennisi et al. (2016), the most used robot in the ASD therapy until 2016 was *NAO*. Tapus et al. (2012) investigated the engagement of 4 autistic children while they interacted with *NAO*, comparing it to a human partner. In the tests, the child starts an arm movement and the partner (human or robot) repeats it. Thus, the initiative of the child to initiate an interaction with the partner is analyzed. For two children there was no difference between the engagement to the human or robot partner. One child presented more focused eye gaze and smiled more while interacting with the robot. Only one child presented more movement initiations while interacting with the robot than with a human partner. *NAO* is presented in Figure 2.1a.

The robot *Kaspar* was used in (ZORCEC et al., 2018) as an intervention toll in the ASD therapy. A study of case with 2 severe autistic children evaluated interaction aspects, including the learning of basic emotions and social skills for a year in a hospital environment. Both children had a very fast and spontaneous interaction with *Kaspar*, which facilitated the improvement of their emotional and social skills. *Kaspar* is presented in Figure 2.1b.

CHARLIE is a low-cost robot proposed by Boccanfuso et al. (2017). In the study, 12 children participated, 8 children without communication problems, 3 autistic and 1 with speech deficiency. The improvement of spontaneous speech, communication and social skills were measured. The results showed that adding the robot as a tool in therapy improved the communication and speech skills of the children. *CHARLIE* is presented in Figure 2.1c.

2.2 Emotion Recognition Based on Heart Rate

Emotion recognition plays a fundamental role in person-to-person interaction. Inferring the partner's mental state provides a way to personalize actions and behavior according to the actual situation. To better interact with people, social robots should also personalize their actions according to the user's emotional and mental states, making the interaction more natural.

Figure 2.1 – Examples of robots used to ASD therapy in the literature.

(a) *NAO* robot.

(b) *Kaspar* robot.

(c) *CHARLIE* robot.

Source – TAPUS et al., 2012; ZORCEC et al., 2018; BOCCANFUSO et al., 2017

2.2.1 Emotion Models

The first challenge at emotion recognition is to quantify something subjective as emotions. In the literature, there are two ways to model the human emotions: the discrete emotion models and the multi-dimensional emotion space models.

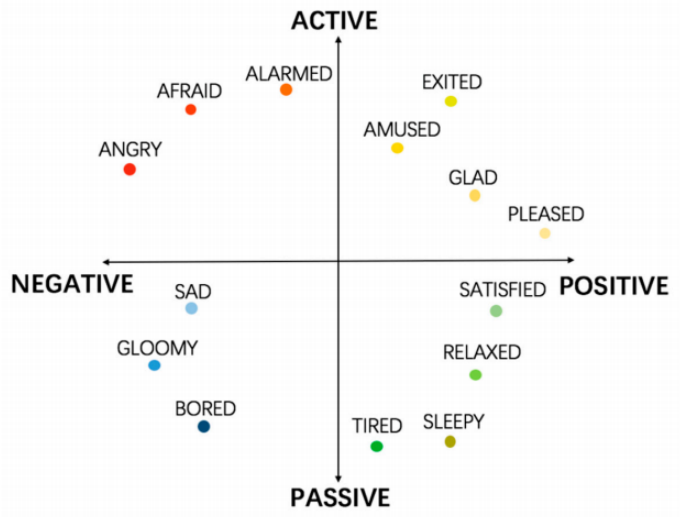
Ekman (1992) defends that there are a discrete number of emotions, where each one of these emotions displays different features. He summarized six primordial emotions: happiness, sadness, anger, fear, surprise, and disgust. According to him, every other emotion is a combination of these ones (EKMAN, 1992). Other authors stand for a different number of basic emotions. For instance, Izard (2007) defines ten basic emotions: interest, joy, surprise, sadness, fear, shyness, guilt, anger, disgust, and contempt.

The multi-dimensional emotion space is an effort to quantify the emotions and their intensities (SHU et al., 2018). Lang (1995) uses a bi-dimensional space to separate emotions in valence (positive or negative emotion) and arousal (passive to active). As some emotions overlap each other at the 2D model, Mehrabian (1997) added another dimension: the dominance. This new axis represents how much a person can control a specific emotion. Figure 2.2a and Figure 2.2b present the positioning of different discrete emotion in the 2D and 3D models.

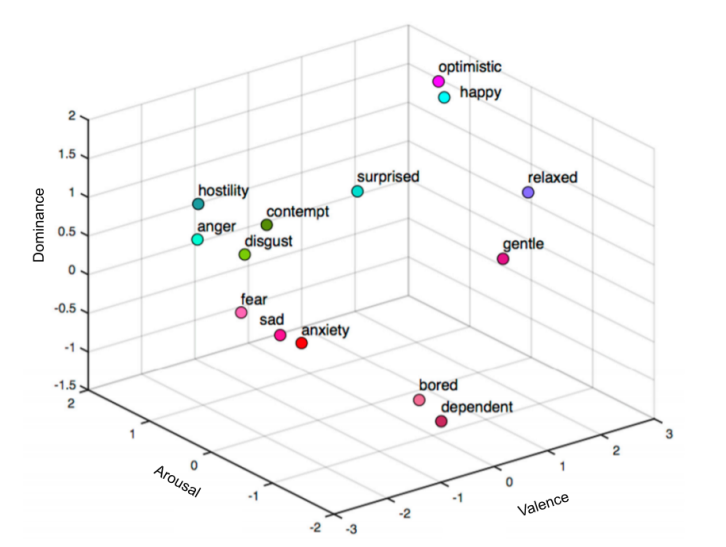
2.2.2 Sources for Emotion Recognition

According to Thomaz et al. (2016), three main sources of data are used for emotion measurement and recognition: facial expressions and features; body posture and motion; and physiological signals. Facial expression based recognition gets images from the user's face, extracts spatial and temporal features, and then, uses it to train a classifier to identify the most probable emotion expressed by the face (KO, 2018). Relating to body postures, Sanghvi et al. (2011) use the body posture gestures to detect the engagement of children playing chess with a robot, and Venture et al. (2014) had success at recognizing human emotion based on the person's gait, using an RGB-D camera. Different physiological signals have been used to detect emotions.

Figure 2.2 – Multi-dimensional emotion space axes.



(a) 2D emotional model.



(b) 3D emotional model.

Source – SHU et al.,2018

Goulart et al. (2019a) proposed a system for emotion recognition in a child-robot interaction using thermal images. Also, electrodermal, respiratory, electroencephalographic and cardiovascular signals have already been used to gather emotional information (SHU et al., 2018).

2.2.3 Cardiac Parameters as Emotional Indicators

The main key linking cardiac signals to different emotional states is the autonomic nervous system (ANS). It controls the different degrees of physiological arousal experienced by the person. The ANS is subdivided into an excitatory sympathetic nervous system (SNS), that is responsible for the heart rate frequency increasing, and an inhibitory parasympathetic nervous system (PNS),

that has an opposite effect. As a person experiences physical or psychological stress the SNS overcomes the PNS, accelerating the heart pulses and increasing physiological arousal. When experiencing a comfortable or relaxing state, the PNS becomes dominant, decreasing heart rate and arousal levels (APPELHANS; LUECKEN, 2006).

As the heart rate variation (HRV) reflects, together with other numerous physiological and environmental factors, SNS and PNS activities at ANS, HRV contains important information about physiological arousal, which has an important role on the individual emotion regulation (APPELHANS; LUECKEN, 2006). The traditional way to measure HRV is the electrocardiography (ECG), which should have a minimum sample rate of 250 Hz (KWON et al., 2018) to locate the R peaks with precision in the signal. The mostly used method to extract the interbeat differences signal (also knows as Normal-to-Normal (NN) intervals) from the ECG is by calculating the temporal distance between the consecutive R-peaks (APPELHANS; LUECKEN, 2006). Figure 2.3a presents the R peaks at the PQRS complex in an ECG model, and Figure 2.3b presents a real ECG sample captured at 400 Hz using *BrainNet*[1]. The features for HRV analysis may be calculated in time, frequency and geometrical domains.

2.2.3.1 Heart Rate Variation Features

Time-domain features are normally based on statistical computation of variation between NN intervals. The most common time-domain features used to describe the HRV are (SCHAAFF; ADAM, 2013; APPELHANS; LUECKEN, 2006; BRENNAN et al., 2001; TASKFORCE, 1996):

- $meanNN$: mean value of NN intervals[2].
- $SDNN$: standard deviation[3] of NN intervals.
- $SDSD$: standard deviation of successive differences of the NN intervals.
- $RMSSD$: square root of the mean of the sum of the squared differences between adjacent NN intervals. RMSSD is presented in Equation 2.1.

$$RMSSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(NN_{i+1} - NN_i)^2} \tag{2.1}$$

- $pNNx$: number of adjacent NN intervals whose difference is superior than x ms divided by the total number of NN intervals. The most used value for x is 50 ms. $pNNx$ is described in Equation 2.2.

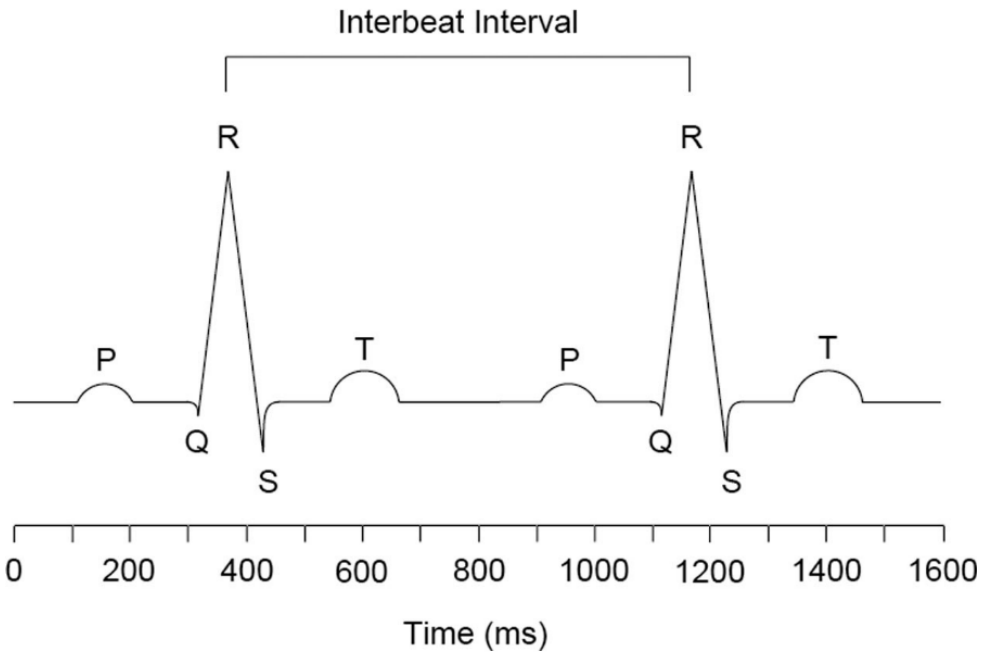
$$pNNx = \frac{1}{N-1}\sum_{i=1}^{N-1}(NN_{i+1} - NN_i) > x\,ms \tag{2.2}$$

[1] BrainNet (from EMSA/Brazil) is a signal acquirement equipment used to capture multiple electric-based physiological signals.

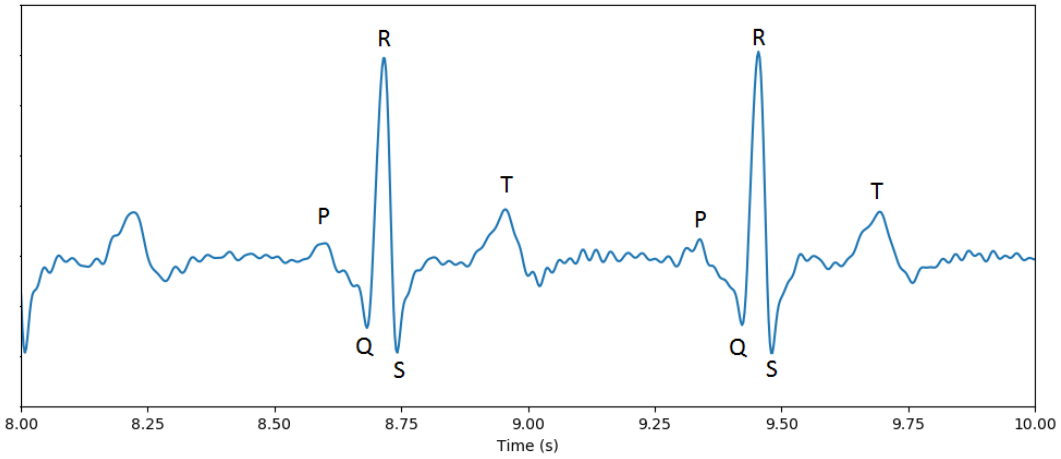
[2] The mean formula is defined in Appendix A.1.

[3] The standard deviation operations is defined in Appendix A.1.

Figure 2.3 – Exemplification of ECG's PQRS complex.



(a) Location of the PQRS waves at the ECG model.



(b) PQRS complex at a real ECG sample.

Source – APPELHANS; LUECKEN, 2006 and Author's database, 2020

There is also a tool called Poincaré plot that represents the correlation between consecutive intervals in a 2D plot. The two numerical features extracted from the plot are the standard deviation of the axes $X1$ and $X2$ presented in Figure 2.4. The standard deviation of $X1$ axis is called $SD1$, and the second is called $SD2$ (BRENNAN et al., 2001). The equations for $SD1$ and $SD2$ are presented in Equation 2.3 and Equation 2.4:

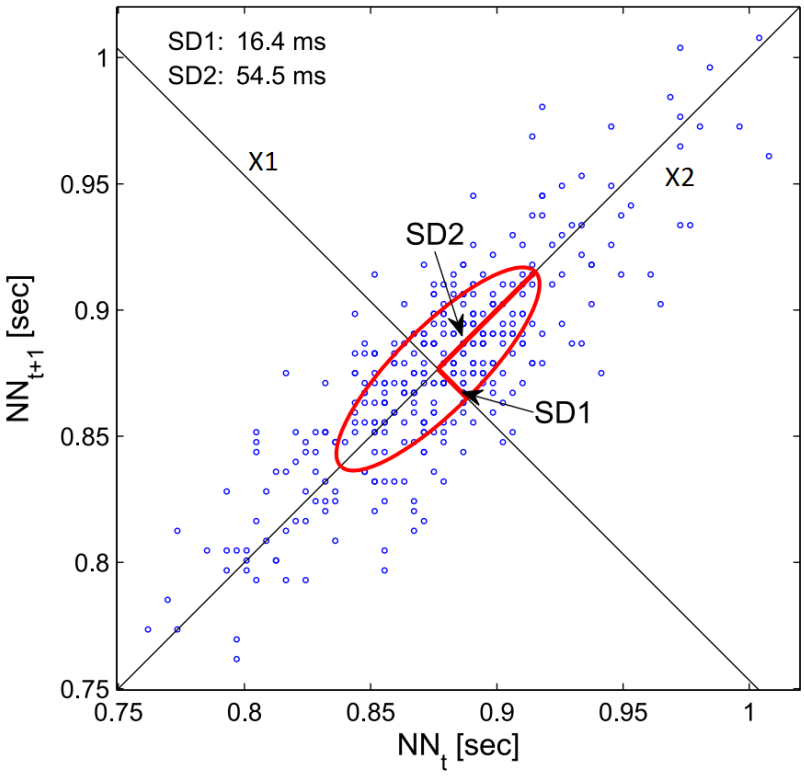
$$SD1 = \sqrt{\frac{1}{2}(SDSD)^2} \tag{2.3}$$

$$SD2 = \sqrt{2(SDNN)^2 - \frac{1}{2}(SDSD)^2} \tag{2.4}$$

where:

- $SDSD$ is the standard deviation of successive differences of NN intervals;
- $SDNN$ is the standard deviation of NN intervals;

Figure 2.4 – Correlation of each method.



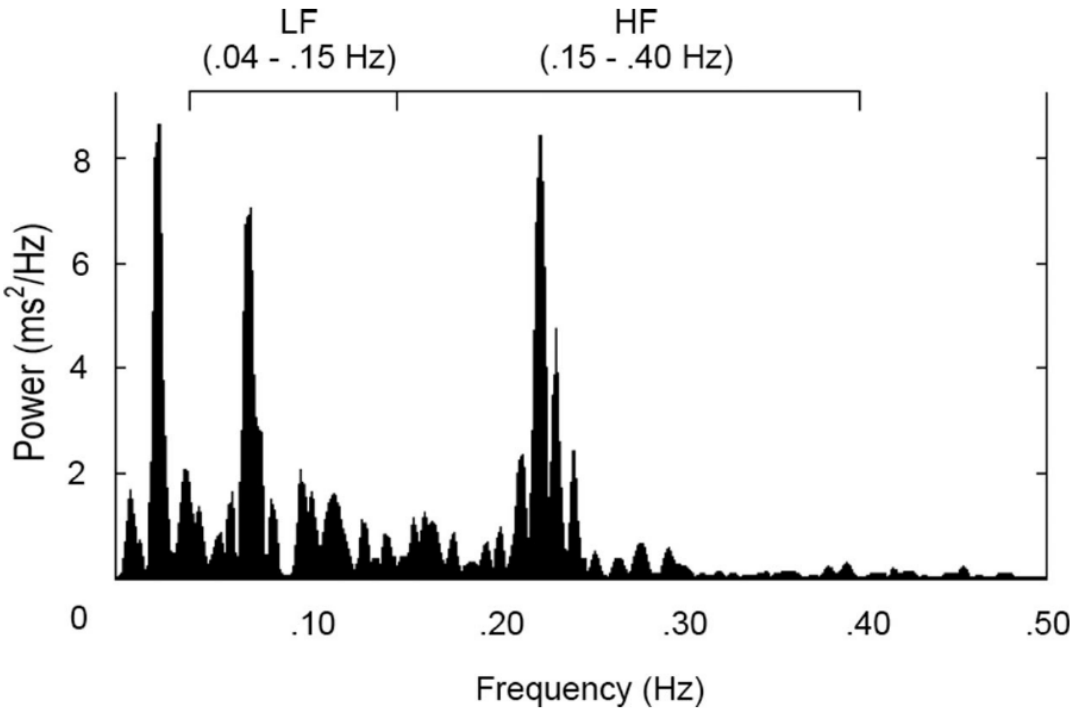
Source – Adapted from (SCHAAFF; ADAM, 2013)

The geometrical analysis bases itself on geometric properties of the probability distribution of NN intervals and differences between consecutive NN intervals. They are less affected by outliers but require a greater number of samples and are less precise (APPELHANS; LUECKEN, 2006).

To calculate the frequency domain features, the interbeat series is transformed to the frequency domain, and its Power Spectrum Density (PSD) is calculated. Then the PSD is divided into the sum of Low Frequencies (LF) power region and High Frequencies (HF) power region. The LF range is from 0.04 Hz to 0.15 Hz, and the HF range starts at 0.15 Hz and goes to 0.4 Hz. The features used to analyze HRV are the sum of each region's power and the ratio between them (LF/HF). According to (APPELHANS; LUECKEN, 2006), as the SNS ans PNS have opposite functions, the first accelerating the heartbeats and the seconds decelerating them, relative shifts

or biases toward sympathetic or parasympathetic dominance over cardiac function reflect directly at the LF/HF ratio value. The LF and HF regions are presented in Figure 2.5.

Figure 2.5 – LF and HF regions at the NN interbeats PSD.



Source – (APPELHANS; LUECKEN, 2006)

2.3 Remote Heart Rate Estimation

The first work to demonstrate that the blood volume variation caused by heartbeats produces enough skin-color variation that can be captured by a camera using ambient light was (VERKRUYSSE et al., 2008). The authors recorded an RGB video of a volunteer's face after exercising. The regions of frames containing the skin were manually selected and the raw Pulse Volume (PV) signal was extracted by getting the RGB values of all skin-pixels at each frame and averaging them for each color-channel. Then a digital 4^{th} order Butterworth filter was used to remove noise outside the range 0.8 Hz to 6 Hz. Analyzing each color-channel separately, they concluded that the green channel has the strongest cardiac signal (which agrees with the fact that hemoglobin absorbs better the green light), however, red and blue channels contain complementary pulse information.

Since then, the number of works related to rPPG has increased significantly, and different parts of the original rPPG methodology have been improved (ROUAST et al., 2018). This section explains in detail the state of art of each part of the rPPG methodology.

2.3.1 General Methodology for Remote Photoplethysmography Based on Color Cameras

According to Rouast et al. (2018), the general methodology to estimate the heart rate using an rPPG signal may be divided into three steps: signal extraction, signal estimation and heart rate estimation. The signal extraction part is where the image processing techniques are applied to identify the skin-region at the video, track it along the frames and extract the RGB temporal signal[4]. The rPPG signal estimation is where signal processing techniques are used to estimate the pulse volume signal from an RGB temporal signal. The last part is the heart rate estimation, which may follow two branches, a power spectrum analysis to find the most predominant frequency to represent the pulse rate or identify the rPPG peaks and get the interbeat intervals.

2.3.1.1 Remote PPG Signal Extraction

In general, the first step to estimate the rPPG signal is to define an ROI to extract the RGB signal. It can be set manually, or automatically. The most used region to extract the raw signal is the head, more precisely the cheeks and forehead regions. To find a face, the most commonly used algorithm is the Viola-Jones (VJ) technique (VIOLA; JONES, 2001), which is a machine learning algorithm that locates faces using simple features. There are also techniques that directly extract the skin-region, without the face detecting part. However, these techniques are highly influenced by objects in the background whose color is similar to the human skin (ROUAST et al., 2018).

With the ROI selected, a raw rPPG signal is generated using two approaches. The first and most common is based on skin-color variation: at each beat, the heart pumps blood to the whole body. This pumping process variates the skin-tissue blood volume and also the optical properties of the skin, which are connected to the blood volume under it. These small color changes may be captured by an RGB camera and processed to infer the rPPG (ROUAST et al., 2018).

The second approach is by analyzing head movements. The same way that pumped blood changes optical skin properties, it has a mechanical impact on the head, neck and trunk. The Newtonian reaction of the body to blood pressure changes generates a head displacement of approximately 5 mm that also may be recorded in video (BALAKRISHNAN et al., 2013).

The color-based methodologies have higher robustness to the subject movement and allow measurements at greater distances. The head movement-based works even when the skin is occluded, and is less affected by the subject skin's tone or the environment's light variations, but is more affected by movement and has a lesser range (AL-NAJI et al., 2017). As one of the objectives is to embed the system in a robot with free movement, and, most of the time, at least

[4] There are also methodologies where the head movement caused by the blood pressure is used (ROUAST et al., 2018).

some skin region of the child will be visible, the color-based methods will be used in this work.

2.3.1.2 Remote PPG Signal Estimation

There are several ways in the literature to separate the pulse signal from the noise at the RGB time series extracted from the ROIs. They normally follow one of these two strategies: the first is to directly weigh the RGB channels combining them to enhance the pulse signal, the second is to use a Blind Source Separation (BSS) algorithm, as the Principal Component Analysis (PCA), to separate pulse signal from noise (UNAKAFOV, 2018; WANG et al., 2017a).

Wang et al. (2017a) directly compare the Signal to Noise Ratio (SNR) of eight different methodologies over four different rPPG challenges: different skin-tones, changes in luminance, subject recovering after exercise and during the exercise. The eight methodologies are: the green channel, green-red channels difference, PCA, independent component analysis (ICA), standardized skin-color chrominance-based rPPG (CHROM), pulse blood volume (PBV) vector, spatial subspace rotation (2SR) (that exploits the skin pixel's distribution), and plane-orthogonal-to-skin (POS) method. The achieved result shows that, for most situations, the POS method got the highest SNR, and for fitness videos, POS and CHROM scored higher.

Unakafov (2018) compared six methods, green channel, green-red difference, an adaptive green-red difference (the channels are weighted according to first frame values), ICA, CHROM and POS technique. The results also pointed out POS as the best methodology, followed by CHROM.

2.3.1.3 Heart Rate Estimation

There are two ways to infer the heart rate using the rPPG signal, analyzing its PSD or detecting the number of peaks inside a time interval. The PSD approach consists of using a time-frequency transform to get the most powerful frequency inside the human heart rate band. The most common transform is the Fourier method, calculated using the fast Fourier transform (FFT) algorithm, however, there are also alternatives as Welch Periodogram (WP) (ROUAST et al., 2018). The peak detection approach consists of splitting the signal and finding the local maximums representing each beat at each segment. The estimated heart rate is calculated by counting the number of peaks at the interval and dividing by its time length.

2.3.1.4 State of Art of the rPPG sensing

As Rouast et al. (2018) present, the different rPPG methodologies are difficult to compare. Most of the authors use particular datasets with different recording hardware, under different light conditions and subject with different skin tones. However, the state of art results gives an idea of how good rPPG can get cardiac information. The methodologies presented in (ROUAST et al., 2018) achieved a heart rate Root Squared Mean Error (RMSE) that ranges from 0.11 bpm

to 7.73 bpm; this can be set as a reference line to the error that the implemented method may have.

Wang et al. (2017a) compared the Signal to Noise Ratio (SNR) from their technique (the Plane Orthogonal to Skin (POS)) against other 7 rPPG techniques. They used a particular dataset containing face-video recordings in multiple situations like different luminance intensities, subject still or in movement, and different skin tones. According to them, their method presented a better signal in most of the cases. It is also shown that the best situation to do the rPPG sensing is with a static well-illuminated subject. Movements, poor illumination and dark skin-tones decrease the quality of the signal, and the movement situations are the most difficult.

An interesting comparison of the method is also done in (UNAKAFOV, 2018), which uses the DEAP dataset (a public database with multiple physiological signals, including the PPG, and face videos provided by (KOELSTRA et al., 2012)) and compares methodologies from other works without proposing a new one. Therefore, their result may be less biased, in which six techniques are compared. The simplest method is using only the green color channel to estimate the heart rate, getting a median RMSE of 6.78 bpm. The second method is using the difference between the green and red channels, which reached a median RMSE of 4.96 bpm. The third is an adaptive green-red difference, which got a median RMSE of 5.55 bpm. The fourth is applying the Independent Component Analysis (ICA) on the color signal, getting a median RMSE of 4.77 bpm. The fifth is a method that transforms the RGB signal to a chrominance plane and then estimates the rPPG, which got a median RMSE of 3.46 bpm. The last is the POS that reached a median RMSE of 3.25 bpm. This result presents a reference of the error that this work should reach at the end.

3 Technique Evaluation

The rPPG methodologies generally are divided into three steps: the face tracking and skin segmentation part; the rPPG signal estimation; and the heart rate and heart rate variation estimation. In each of these parts the techniques presented at the literature will be tested for speed (to be able to work online the algorithm should be fast) and quality (the more similar the rPPG estimation is from the true PPG signal, better the algorithm). Then, all the sections in this chapter will have a part of development, explaining the algorithms, and a part of the results, comparing the processing time and the quality of the techniques.

The overall methodology proposed to extract the cardiac information from the videos and classify it by their emotional information is divided into six sections, as presented in Figure 3.1:

1. Face-videos and reference cardiac signals recording: two databases were created during the research, one focused on the user's movement, where the subjects are asked to perform specific actions to test the robustness of the rPPG even when the user is not still, and other focused on record the subject's cardiac activity while they watch a set of small videos chosen to trigger a specific emotional state.

2. Face tracking and skin segmentation: this part is focused on evaluating the algorithms presented in the literature to find a face and keep following it through the video. It also presents the methodologies to extract pixels containing the skin-region at the face.

3. Remote PPG signal estimation and filtering: at this part are evaluated different methodologies to estimate the pulse signal from the RGB time series.

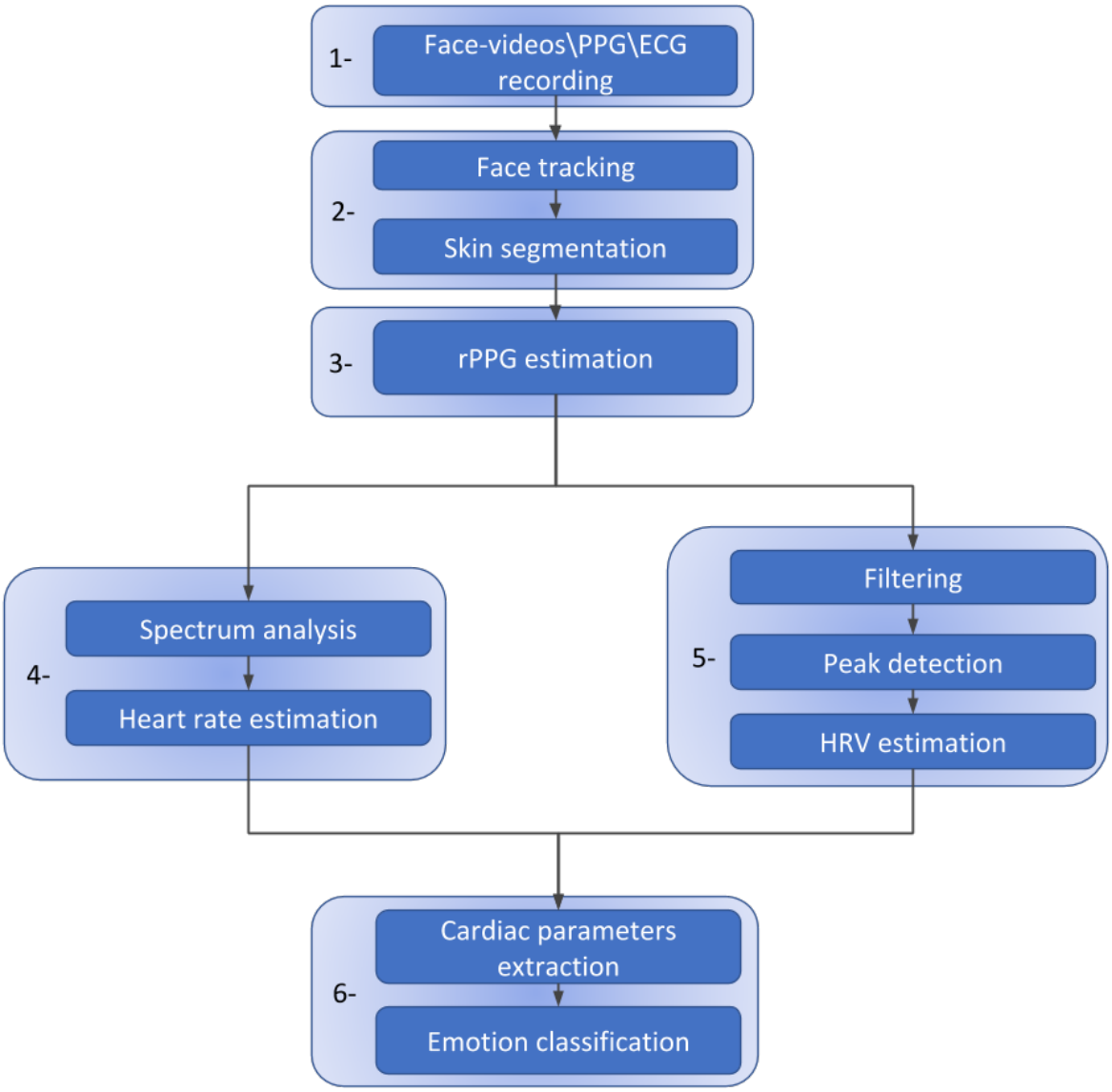
4. Heart rate estimation: the main methodologies based on PSD analysis to infer the heart rate from rPPG are tested.

5. To infer HRV parameters from the rPPG, the interbeat peaks should be identified at the signal first. In this part, the peaks generated by each algorithm are compared with peaks from the ECG.

6. After getting the cardiac signal from the input images, the last step is to evaluate the possibility of differentiating emotions based on cardiac information from rPPG. The parameters used in the literature to quantify emotions using ECG and PPG are also used for rPPG and, in the end, a classifier is proposed for testing the three signals.

The main problem may be simplified as an online emotion classifier based on rPPG. An online system has limited computational time, which sets a boundary on the algorithm's complexity, so, for each part, the processing time of each technique is measured together with its

Figure 3.1 – Methodology to extract the cardiac information.



Source – Author's database, 2020

accuracy. The techniques that most approximate the cardiac signal obtained by the RGB camera to the one got with ECG and PPG are chosen. And finally, features are extracted from the rPPG, ECG and PPG, and a classifier is used to identify the emotions on each signal.

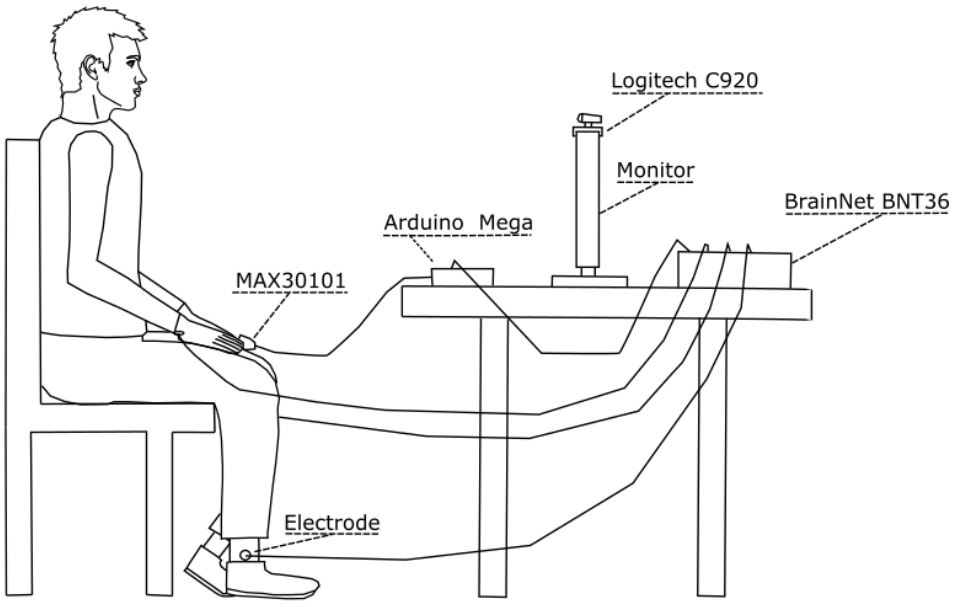
3.1 Databases

Two databases were created to test the techniques. The first one was made aiming to compare the robustness of rPPG estimation methodologies to the user's movement, so the cardiac information was recorded together with face-videos while the subject was performing a set of movements in front of the camera. The second dataset was created aiming to analyze the cardiac

changes while the user was experiencing different emotions. To record the emotional state of the volunteers, they were asked to watch six short videos to excite emotions while the RGB and thermal images were recorded. For both databases, the PPG and ECG were also recorded as a reference signal.

The schematic for capturing data at both databases is presented in Figure 3.2. The webcam used to record the video was a *Logitech C920*[1]. The images were recorded at 30 fps, and the distance between the camera and the subject was about 70 cm. The PPG data were acquired using the pulse oximeter sensor *MAX30101* attached with an *Arduino Mega* board using a sample rate of 50 Hz. The ECG signal was recorded using a sampling rate of 400 Hz with the *BrainNet* equipment [2]. The D1 derivation was used to get the ECG (electrodes at the left and right wrists, and the ground at the right ankle, as presented in Figure 3.3), according to Pastore et al. (2009). The *Arduino* board was also used for recording synchronization: at the beginning of each video, a command was given the *Arduino* board to start to take data from the PPG sensor and also to set a port to a high state. When the video was finished, a command was given to *Arduino* to stop taking data from the PPG and set the port to a low state. This port was connected to the *BainNet* so it was possible to split the video and take only the part that the *Arduino*'s port was with a high level.

Figure 3.2 – Schematic to capture face-videos, PPG and ECG signals.

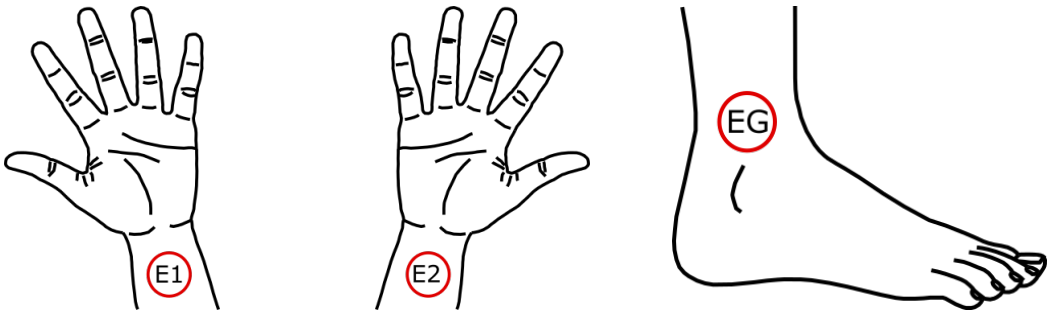


Source – Author's database, 2020

This research was approved by the ethics committee from UFES, under the number CAAE: 44899015.0.0000.5060.

[1] <https://www.logitech.com/pt-br/product/hd-pro-webcam-c920>.
[2] <http://www.emsamed.com.br/pt-br/brainnet-bnt-36>.

Figure 3.3 – Electrodes position.



(a) Left hand electrode. (b) Right hand electrode. (c) Right foot electrode.

Source – Author's database, 2020

3.1.1 Movement Database

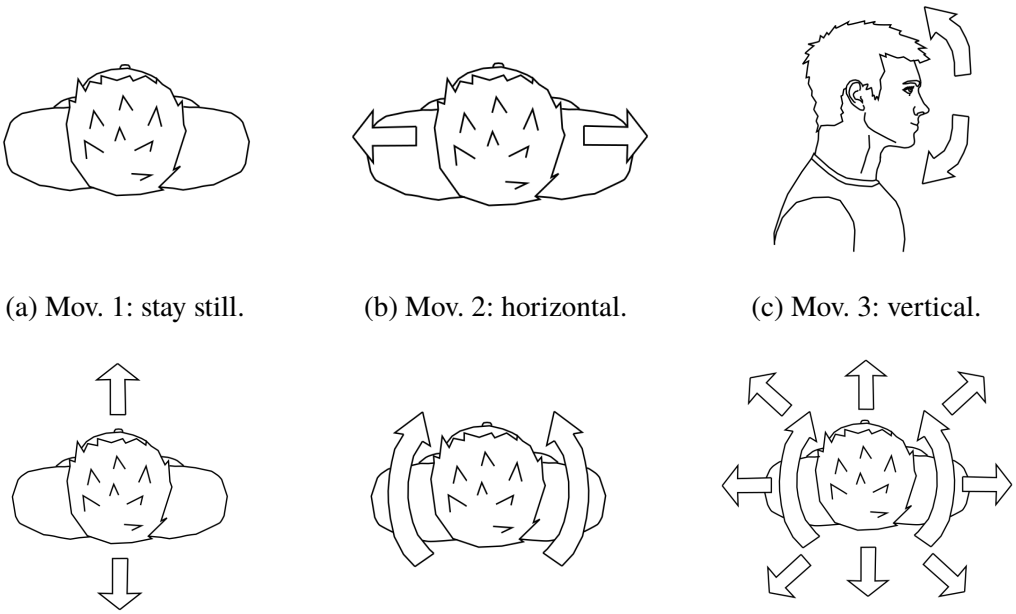
At this database, the subjects were asked to perform six movements, each for one minute. The whole movements were done with the volunteers sat on a chair at a distance of approximately 70 cm from the camera. Together with the face-videos, the PPG and the ECG from the users were also recorded. In total, 15 people participated in the tests, aged from 20 to 37 years. They were 3 women and 12 men, with a wide range of skin tones. The frames were recorded at a resolution of 320x240 pixels. The movements performed are specified below:

- Movement 1 - Stay still: in the first video, the subjects are asked to not perform movements at all. They are asked to stay still in front of the camera. See Figure 3.4a.
- Movement 2 - Horizontal: the second movement is to move the body to the sides. See Figure 3.4b.
- Movement 3 - Vertical: the third movement is to move the head up and down. See Figure 3.4c.
- Movement 4 - Back and forward: the fourth movement is to approximate the body to the camera and then back to the original position. See Figure 3.4d.
- Movement 5 - Head rotation: in the fifth movement, subjects are asked to rotate the head to both directions. See Figure 3.4e.
- Movement 6 - Mixed movements: in the last part, the subjects could mix all other movements and perform another one if they wanted. See Figure 3.4f.

3.1.2 Emotion Database

At this database the subjects were asked to watch the same videos used by Goulart et al. (2019a): the first one is a neutral video, a video clip of an infant movie; the second one is

Figure 3.4 – Movements performed by the subject at database one.



(a) Mov. 1: stay still. (b) Mov. 2: horizontal. (c) Mov. 3: vertical.

(d) Mov. 4: back and forward. (e) Mov. 5: head rotation. (f) Mov. 6: mixed.

Source – Author's database, 2020

supposed to trigger disgust, a part of a survival program where a person is eating worms; the third is a part of a scary movie, with a ghost; the fourth is used to excite happiness with videos of babies laughing and playing with dogs; the fifth is to trigger sadness with a montage of abandoned dogs; and the last one is a surprising video, a commercial with a mouse falling in a rat-trap but escaping in a funny way. Before each video, a four seconds countdown was shown on the screen, and the record started at the beginning of the countdown. A frame of each video is presented in Figure 3.5.

The videos used in (GOULART et al., 2019a) were selected by a physiologist. The acquisition protocol had the following steps: first, it was explained to the volunteer the experiment's objective, that is to identify emotions through rPPG, and also, a general idea of the whole steps of the research. The volunteer was then informed that he/she could stop the test at any time he/she wanted. After the explanation, the subject was asked to sign the consent form[3] and sit in a chair comfortably. The ECG electrodes were attached to the subjects' wrists and ankle, and the PPG sensor was attached to the right hand's middle finger.

The videos were played and the data were automatically recorded: color RGB frames (640 x 480 pixels)[4] at 30 fps, thermal images (384 x 288 pixels) at 8.7 fps[5], ECG at 400 Hz and PPG at 50 Hz. At the end of each video, the volunteer was asked about the arousal and valence

[3] A copy of the consent form is presented at Annex A.

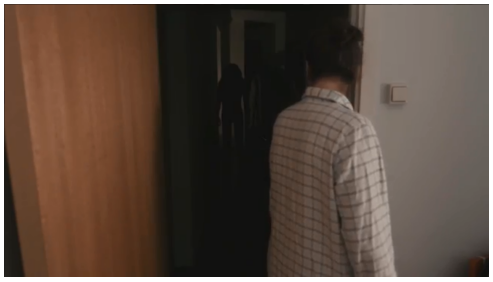
[4] The videos were recorded at a higher resolution than the movement database to make easier to use it for different emotion detection methodologies in future works.

[5] The thermal images were recorded for further emotion analysis with other methodologies.

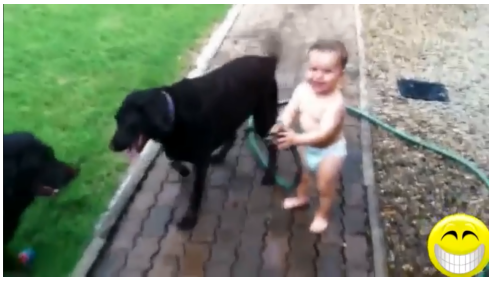
Figure 3.5 – Frames of the videos used to excite emotion.

(a) Neutral video.

(b) Disgusting video.



(c) Scary video.



(d) Happy video.

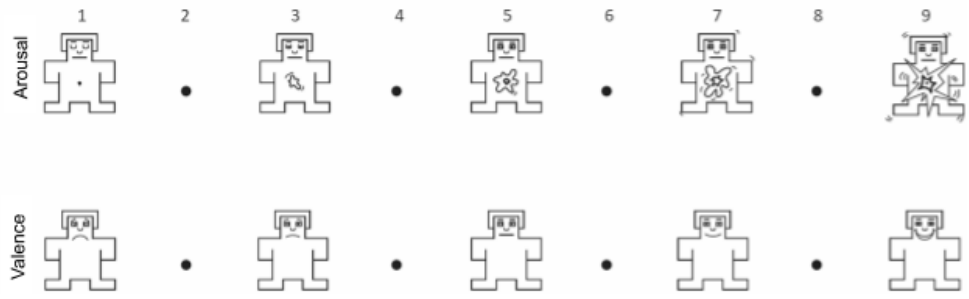
(e) Sad video.

(f) Surprise video.

Source – Author's database, 2020

that the videos had triggered using the Self-Assessment Manikin (SAM) (MORRIS, 1995). SAM quantifies the arousal and valence on a scale of 1 (very low for arousal and very bad for valence) to 9 (very high for arousal and very good for valence). The volunteer was also asked if he/she had already seen the videos before. The scales are presented in Figure 3.6.

Figure 3.6 – SAMs questionnaire for arousal and valence.



Source – Adapted from MORRIS, 1995

In total, 30 subjects participated in the tests (12 women and 18 men), aged between 19 to 30 years, with multiple skin tones. The only restriction criterion was to exclude subjects that had psychological traumas or phobias (as in the protocol of Goulart et al. (2019a)).

3.2 Face Detection and Tracking

In this part of the text the evaluation methodology and results of the face-detecting and object-tracking algorithms using two image processing libraries *OpenCV* and *dlib* are presented.The processing time of the codes and their success rate were calculated and evaluated. For the face-detectors, as the videos were recorded with only one person each time, in a white background[6], it was supposed that, whenever only one face was found by them on a frame, it was a success. For the object trackers, the evaluation process was based on the number of frames that it took to lose the target, and the precision was calculated comparing the face region located by the object tracker, with the region from the face detector as a reference. As one of the main targets of the final algorithm is to work online, besides accuracy, the processing time is crucial.

It was chosen to use both, the face detector and the object tracker instead of only the face-detector for two reasons: the speed and to deal with different face positions. Object trackers are usually faster than face-detectors, which is a crucial point for an online system. The second reason is that face-detectors have difficulties to identify faces turned to the sides, and the object trackers handle better with these situations, since it follow the changes at the face region frame by frame as it is moving. The detectors look for specific features that defines a face, and do not use information in the previous frames to optimize the search. So, when a face is turned to the sides, some features used by the face detectors may be hidden, decreasing the success chances.

3.2.1 Face Detection

Most of the works use Viola-Jones (VJ) to identify faces in input images (ROUAST et al., 2018), as it is a fast algorithm and has really good precision. Also, it is already implemented in open source libraries as *OpenCV* (OPENCV, 2018), which makes the code easy to use. In this work, VJ was compared to another face detection algorithm, also available in an open-source library, the face locator provided by *dlib* (KING, 2009). It is based on the histogram of oriented gradients (HOG) algorithm (SURASAK et al., 2018) together with a linear Support Vector Machine (SVM) classifier, an image pyramid of upsampled versions of the image and a sliding window to detect features. For both detectors, the recommended configuration given by their respective documentation was used.

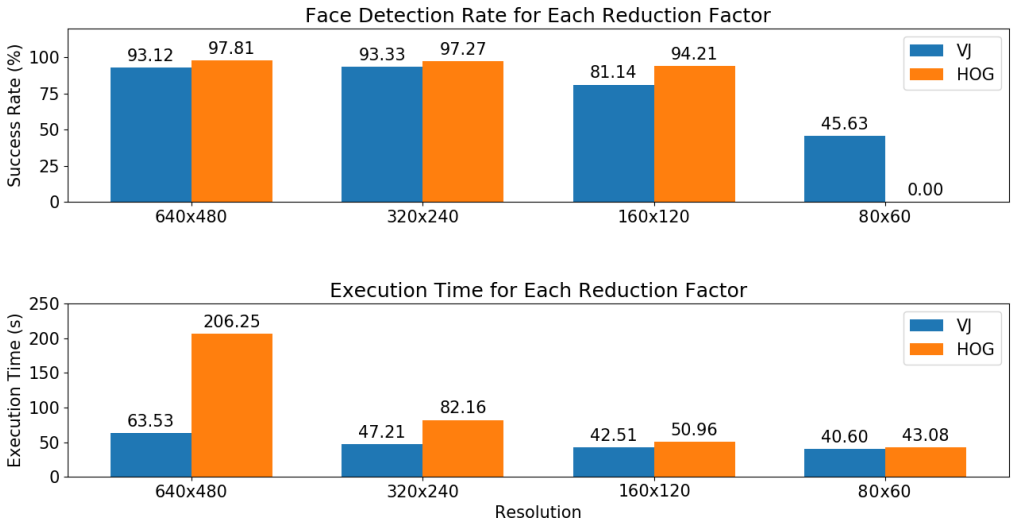
Both face-detectors were compared by the capability to find faces in different frame resolutions, 640x480 (original), 320x240, 160x120, 80x60 pixels and processing time. So, both

[6] It was chosen to a white background to decrease the noise sources of the video, however it may be be considered adapt the algorithm in the future to deal with the background where the child will interact with the robot

detectors were set to find a face at the same videos for multiple subjects at the Emotion Database. The Emotion Database was used because this dataset was recorded at a higher resolution. The processing time and the number of frames where a face was found were recorded. The face-videos have about 1315 frames, recorded at 30 fps (totaling about 44 seconds). The reason that the frame resolution was tested is because one of the most time-consuming parts of the code is the face location and tracking. So, a smaller frame means a smaller processing time, and, as only the mean RGB value of the whole skin region is used for the rPPG estimation, by reducing the frame resolution, it should not have a big impact on the mean value.

To use the VJ algorithm, first the cascade file with the face features was loaded, then, for each frame-size (640x480 (original), 320x240, 160x120, 80x60), the original image was passed from RGB to gray scale. Then, the image was equalized, as recommended in the *OpenCV* web-page. Finally, the image was given to the VJ to find the face on it. The HOG was much simpler: the image was just loaded and given as input to the algorithm. The execution time and the number of frames where a face was found was recorded. The result of the tests are presented in Figure 3.7.

Figure 3.7 – Success rate and execution time of VJ and HOG.



Source – Author's database, 2020

Analyzing the results, the HOG algorithm, for higher resolutions, had a higher success rate than VJ, but both algorithms scored higher than 90% for 320x240 and 640x 480 resolutions. When the frame size is lower, the algorithms run faster, but the success rates are lower. For example, for 160x120 the VJ found the face in 81% of the videos and the HOG in 94%. At the 80x60 resolution, VJ was capable to identify a face in almost half of the frames and HOG did not find anything. Looking at the spent time, VJ was much faster than HOG for bigger resolutions. An oddly result is that VJ did not show any great decreasing of time for smaller resolutions. For HOG, the processing time decreased considerably, from 206 s in the 640x480 resolution to 43 s

at the 80x60 images. The VJ got 63.53 s at the 640x480 resolution and 41 s for the 80x60.

Considering the success rate and the processing time, the VJ algorithm was selected for our system. The success rate was just a little lower than for HOG at the 320x240 resolution, but the mean execution time was almost half of HOG. The 320x240 resolution was also chosen because, as will be shown at the next section, to extract the skin, another algorithm was used to detect specific points of the face. So, a higher resolution may increase its accuracy. Thus, for the proposed system, the VJ face-detector was chosen, and the images were set to a frame resolution of 320x240 pixels.

3.2.2 Object tracking

Instead of using only the face-detector to locate the user's face, it's interesting to use a faster tracking algorithm to follow the face through the videos after it is firstly located by the face-detector. Tracker algorithms follow a specific region of an image at the posterior frames. *OpenCV* (versions 3.4.2+) provides a set of eight trackers[7]. A small explanation and the references about each of them are given as follows:

- Boosting - A real-time object-tracking technique based on the AdaBoost algorithm[8];
- MIL - This algorithm separates the tracked object from the background by training a classifier online[9];
- CSRT - Algorithm based on the work "*Discriminative Correlation Filter with Channel and Spatial Reliability, an Algorithm Implemented*"[10] (LUKEŽIČ et al., 2018);
- MedianFlow - Implementation of the paper "*Forward-Backward Error: Automatic Detection of Tracking Failures*" (KALAL et al., 2010). This tracker is very fast, suitable for videos with simple movements, where the object is not occluded, and, according to the authors, it's quite accurate for those situations[11];
- TLD - This algorithm improves the accuracy of other trackers by splitting the tracking task into three parts: tracking, learning and detecting. During the tracking process, the detection part analyzes every frame and performs full scanning of the image to localize the object and, if necessary, the algorithm corrects itself. When an error is detected (basically, when there is a difference between detection and tracking outputs), the code tries to learn how to avoid it in the following frames. The algorithm is based on the paper "*Tracking-Learning-Detection*" (KALAL et al., 2012). This methodology works as error detection and correction for other tracking methodologies. To use it, first, another tracker must be

[7] <https://docs.opencv.org/3.4/d0/d0a/classcv_1_1Tracker.html>.
[8] <https://docs.opencv.org/3.4/d1/d1a/classcv_1_1TrackerBoosting.html>.
[9] <https://docs.opencv.org/3.4/d0/d26/classcv_1_1TrackerMIL.html>.
[10] <https://docs.opencv.org/3.4/d2/da2/classcv_1_1TrackerCSRT.html>.
[11] <https://docs.opencv.org/3.4/d7/d86/classcv_1_1TrackerMedianFlow.html>.

chosen. The MedianFlow algorithm is chosen as the tracker in the *OpenCV* implementation. It is supposed to be able to handle harder situations as occlusions and rapid movements[12];

- MOSSE - Minimum Output Sum of Squared Error is a tracker implemented based on the work "*Visual Object Tracking using Adaptive Correlation Filters*"[13] (BOLME et al., 2010);
- GOTURN - Generic Object Tracking Using Regression Networks (HELD et al., 2016) is the only tracker implemented in *OpenCV* that uses Convolutional Neural Networks (CNN) (until version 3.4.2). The weights of the network must be downloaded separately from a *GitHub* repository before use. This algorithm does not handle occlusions, but it works well with changes of viewpoint, deformations and lighting changes[14].

An ideal tracker must be fast, robust to occlusion, illumination and viewpoint changes. However, real algorithms hardly reach all these features, so it must be chosen the optimum method to fulfill the main requirements of this research, that are: 1- the system is an online application, so it must run fast; 2- as the robot is supposed to follow the child in the room, and keep a small distance from him/her, it is not common to have objects occluding the child from the robot's camera; 3- the child will be most of the time with the face turned to the robot, so there is not a problem if the tracker loses the target some times, because VJ can be used again to reset the face's location.

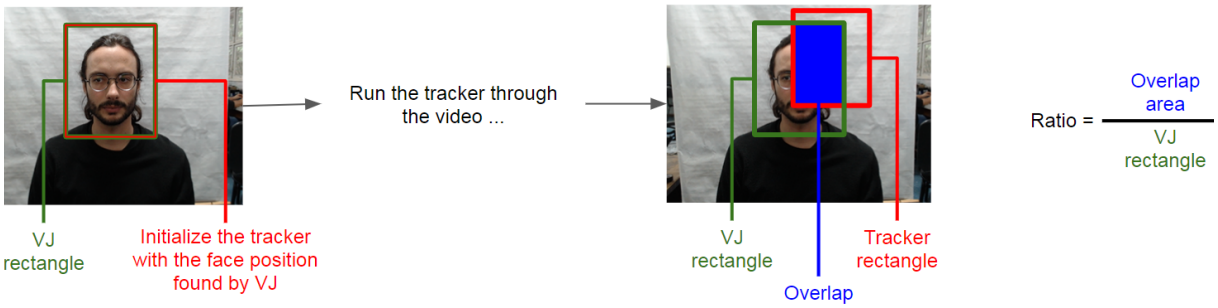
For the tests, the videos from the movement database were used, as they provide more challenging situations for the algorithms. The tests were performed as follows: first, the VJ algorithm was run for each video to find the face location in all frames as the ground truth. Next, the trackers were initialized with the face's position of the first frame, and then, they tracked it in the following frames. To evaluate the algorithms, the time spent to process the videos was recorded, as well as the number of times that each tracker lost the face. Also, the overlapped area between the face-rectangles returned by the trackers and VJ in the last frame was calculated. So, it was possible to compare if the tracker was still covering the whole face region in the end of the video. the procedure to calculate the overlap ratio is presented in Figure 3.8.

The trackers were set to run the horizontal movement videos of each volunteer (movement 2 from the Movement Database[15], see Figure 3.4b). Each video was recorded at, approximately, 30 fps, totaling about 1800 frames per video. There is no occlusion or camera-movement in the dataset, and, in the whole videos, the distance of the subjects from the camera did not change too much, so it is a relatively easy task for the trackers. The time spent by each algorithm is presented in Figure 3.9.

Looking at Figure 3.9, it's possible to see that some of the trackers are much faster than VJ. To process a 44 s, 320x240 video, VJ spent 47 s, which is about 28 fps; MOSSE tracker spent

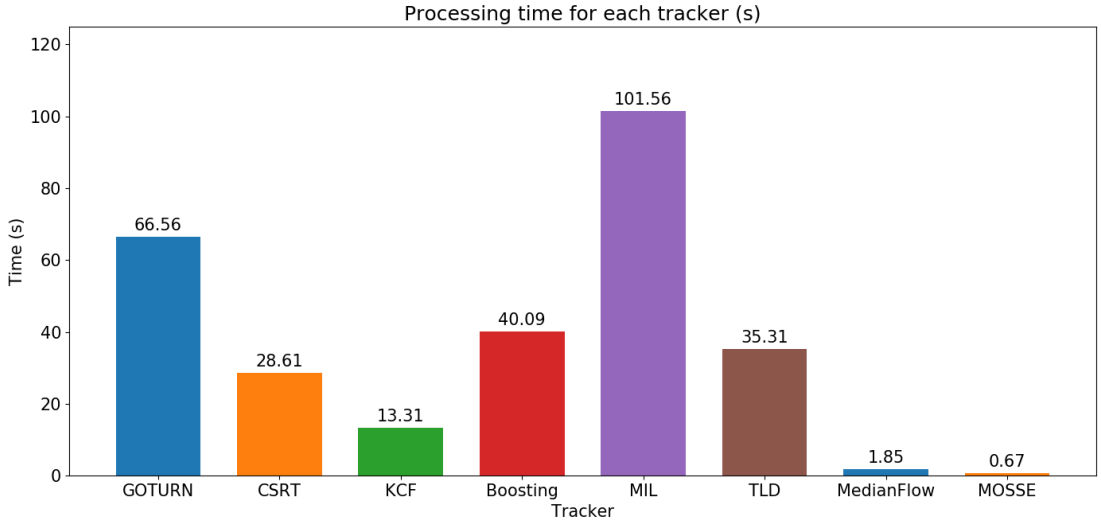
[12] <https://docs.opencv.org/3.4/dc/d1c/classcv_1_1TrackerTLD.html>.
[13] <https://docs.opencv.org/3.4/d0/d02/classcv_1_1TrackerMOSSE.html>.
[14] <https://docs.opencv.org/3.4/d7/d4c/classcv_1_1TrackerGOTURN.html>.
[15] It was used the movement 2 videos because it has the greatest head displacement

Figure 3.8 – Procedure to calculate the face's overlap between the trackers and VJ.



Source – Author's database, 2020

Figure 3.9 – Execution time of each object-tracker.



Source – Author's database, 2020

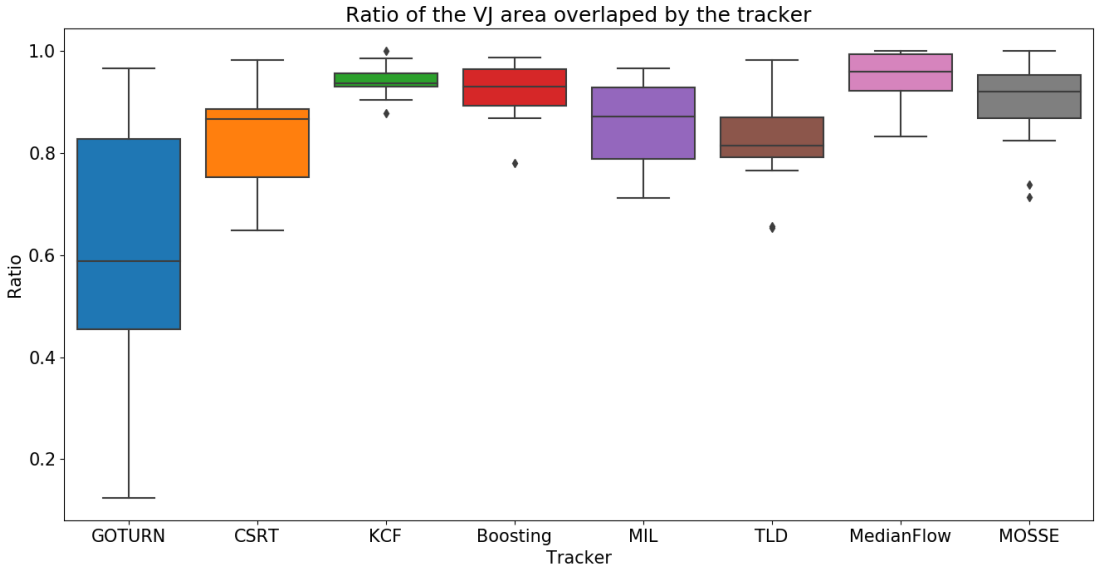
0.67 seconds to process a 60 s 320x240 video, which is about 2687 fps; MedianFlow got 972 fps; and the KCF got 135 fps. These are the fastest ones, but all trackers that processed the video under 60 s may be suitable. So, to help to choose the tracker[16], the overlapped VJ area ratio at the last frame of the videos is presented in Figure 3.10.

To improve the robustness of the evaluation, the overlap test was repeated for the whole videos of the movement database[17]. This test has a more challenging scenario, where the faces are not just in one position but there is also head rotation. Which changes the aspect of the tracked face. The result is presented in Figure 3.11

[16] The number of times that each tracker lost the target was also counted, however, as it was an easy task, none of them have lost the target at the test

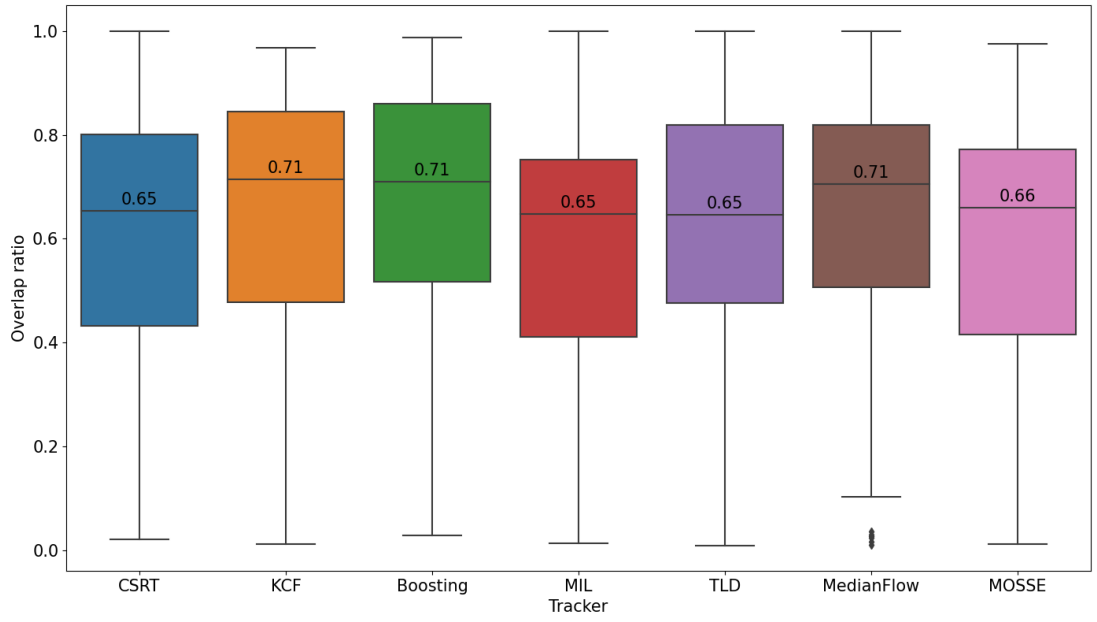
[17] This test was performed after the defense of this work. By this time, an actualization on the *OpenCV* software made the GOTURN tracker unavailable, so, it was removed from the test. And, as this tracker had already a poor result in the before tests, remove it should not affect the choose of the optimum tracker

Figure 3.10 – Ratio of the VJ's area overlapped by the trackers at the last frame (horizontal movement).



Source – Author's database, 2020

Figure 3.11 – Ratio of the VJ's area overlapped by the trackers at the last frame (All movements).



Source – Author's database, 2020

In an ideal situation, the best tracker is the one whose ratio equals one, which means that the whole area selected by VJ as a face region was still covered by the tracker in the last frame. The best median values were reached by MedianFlow, KCF and Boosting, but the one with less variation was KCF. The worst median results were reached by GOTURN, MIL and

TLD. As MedianFlow was the second faster tracker and got the best median overlapping ratio, it was chosen as the system's face-detector.

3.3 Skin Segmentation

To generate the rPPG signal, a part of the subject's skin must be extracted from the image. Two possibilities to get the skin-pixels were tested, the first is to find Regions of Interest (ROIs) at specific parts of the face (specific ROIs), and the second is to select the skin pixels in the frame according to their color.

3.3.1 Specific ROI Approach

The advantage of this method is that it can select the regions of the skin with the best view of blood flow, as the cheeks and forehead (GOULART et al., 2019a). The disadvantage of this approach is that it may not be robust to partial face occlusions and an auxiliary algorithm is necessary to locate the ROIs. There is an Artificial Intelligence (AI) algorithm trained to find specific points of the face, called facial landmarks (FLM) (KAZEMI; SULLIVAN, 2014), available at the *dlib* library [18]. The AI's result with the ROIs selected are presented in Figure 3.12. The red dots at the image are the 64 points returned by the FLM algorithm and the blue region is the extracted skin-pixels. To make the selection of the ROIs, first, lines linking specific points are drawn, then the intersection of these lines are taken to form a polygon. The pixels inside this polygon are then selected to represent the skin. This process is illustrated in Figure 3.13.

One of the biggest problems of this approach is the right positioning of the FLM mask. The face rotation or the use of glasses set difficulties for the AI to identify the right point location, as presented in Figure 3.14.

3.3.2 Color Approach

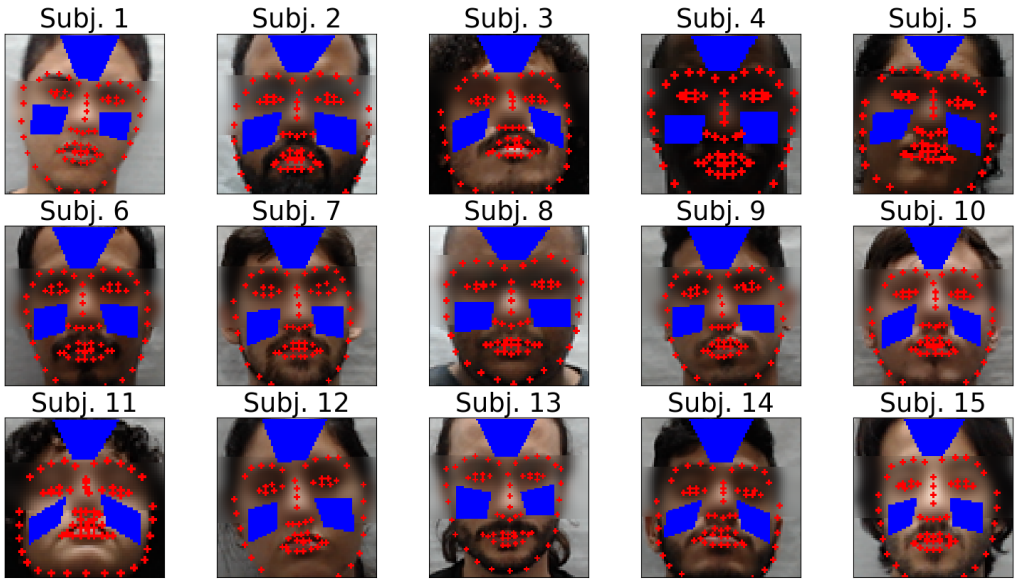
The extraction by color has the advantage of being robust to partial occlusions. As long as a piece of skin is visible in the video, it can be extracted and the signal may be generated. The disadvantages are that some noisy regions are also included in the signal.

Color thresholds must be defined to identify skin-pixels. The rPPG method proposed by Bousefsaf et al. (2013) uses the values presented in (MAHMOUD, 2008). At the work, the face is located using VJ, then the face-images are converted to the YCbCr[19] color plane, due to its efficiency in skin detection (PHUNG et al., 2002). To convert an RGB image to YCbCr, one can use Equation 3.1. Then, the limits presented in Equation 3.2 are used to select the pixels

[18] <http://dlib.net/files/shape_predictor_68_face_landmarks.dat.bz2>.

[19] Y describes brightness, and the other two components describe the blue-Y difference and the red-Y difference.

Figure 3.12 – Cheecks and forehead skin extracted from images using the FLM.



Source – Author's database, 2020

containing skin. Figure 3.15 shows the skin selection technique applied to the subjects from the movement database.

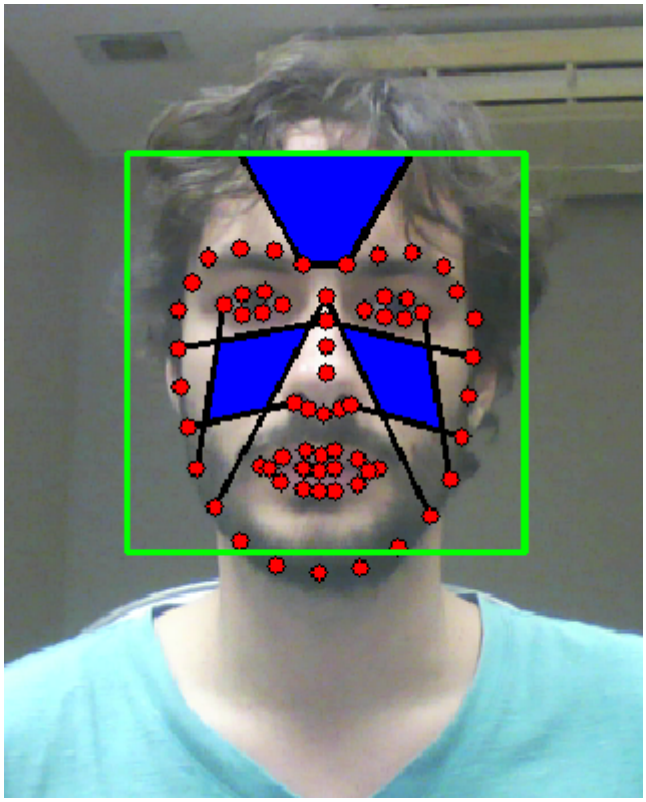
$$\begin{cases} Y = 0.299R + 0.587G + 0.114B \\ C_r = R - Y \\ C_b = R - Y \end{cases} \tag{3.1}$$

$$\begin{cases} Y > 80 \\ 85 < C_b < 135 \\ 135 < C_r < 180 \end{cases} ; \quad \text{where} \quad Y, C_b, C_r = [0, 255] \tag{3.2}$$

The faces in Figure 3.15 show that the limits from Equation 3.2 worked well for most of the subjects, extracting better the well-illuminated part of the skin. For subjects with darker skin tones as Subject 4 (that has the darkest skin-tone), the limits failed to extract the skin.

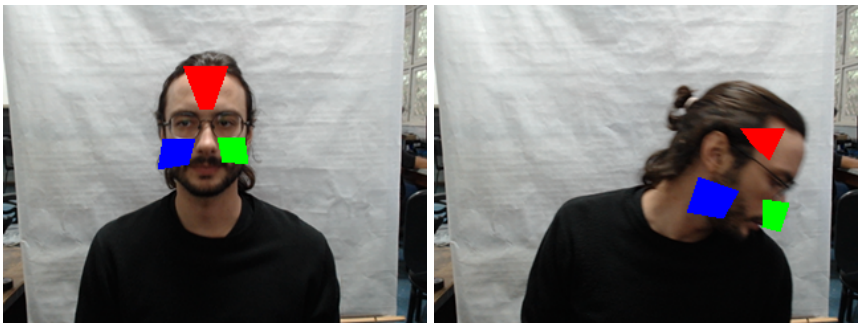
To specify the skin color selection to the user's skin-tone and decrease the possibility of including non-skin pixels in the selection, a new approach was developed. This approach is a mixture of specific ROIs and color-based techniques. In the first frame of the video, after the face is located, the FLM algorithm is applied to find the cheeks of the user, as presented in Figure 3.13. The forehead ROI is excluded because the user may have a hair fringe. A Gaussian Blur filter with a mask-size of 5x5 is then applied at the image to eliminate high-frequency noise, and the face-image is converted to the YCbCr color space. The pixel's color-intensities of the cheeks are then used to define new skin thresholds. To avoid including colors from facial hair or some

Figure 3.13 – Facial-landmarks ROI selection.



Source – Author's database, 2020

Figure 3.14 – Facial-landmarks errors.

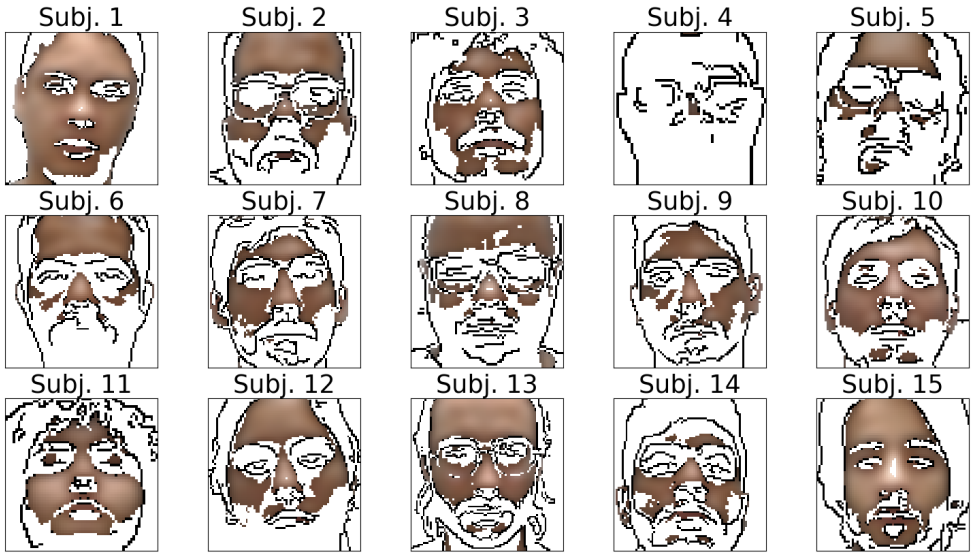


Source – Author's database, 2020

objects as glasses, part of the pixels are pruned by excluding the pixels whose values are too different from the median color values of the selected pixels at each channel. The lower limit is defined as the median value minus n times the standard deviation of the channel, and the higher is the median plus n times the standard deviation[20]. This procedure is defined in Equation 3.3. After calculated at the first frame, the same limits are used for the following frames at the same

[20] The median and the standard deviation (Equation A.2.) are defined at Appendix A.1

Figure 3.15 – Skin extracted using the limits of Equation 3.2.



Source – Author's database, 2020

video, they do not need to be calculated again[21].

$$\begin{cases} Y_{minor} < Y < Y_{major} \\ C_{bminor} < C_b < C_{bmajor} \\ C_{rminor} < C_r < C_{rmajor} \end{cases} \tag{3.3}$$

where:

$$\begin{cases} Y_{minor} = Median(P_y) - n * Std(P_y) \\ Y_{major} = Median(P_y) + n * Std(P_y) \\ C_{bminor} = Median(P_{C_b}) - n * Std(P_{C_b}) \\ C_{bmajor} = Median(P_{C_b}) + n * Std(P_{C_b}) \\ C_{rminor} = Median(P_{C_r}) - n * Std(P_{C_r}) \\ C_{rmajor} = Median(P_{C_r}) + n * Std(P_{C_r}) \end{cases} \tag{3.4}$$

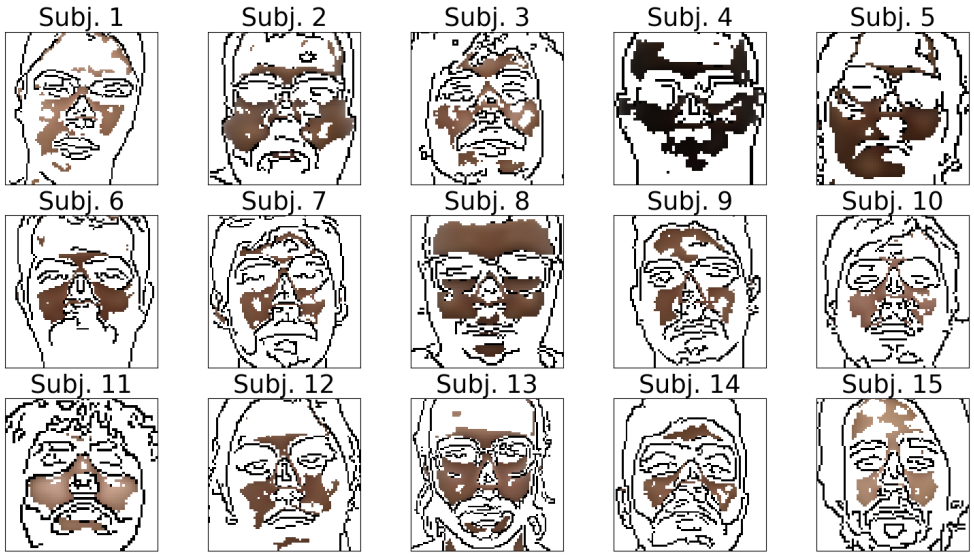
where:

[21] As in a real application the illumination conditions can change, it is reasonable to repeat this procedure after some frames to also adapt the algorithm to the environment change conditions.

- $minor$ represents the lower limit for a pixel to be selected;
- $major$ represents the upper limit for a pixel to be selected;
- Y, C_b, C_r = [0,255];
- $Median()$ is the median operation;
- $Std()$ is the standard deviation operation;
- P_Y is the value of the pixels at the Y channel;
- P_{C_b} is the value of the pixels at the Cb channel;
- P_{C_r} is the value of the pixels at the Cr channel;
- n is an input parameter to increase or decrease the limits.

Figure 3.16 presents the result of the adaptive skin segmentation with $n = 1$. For most of the subjects, the cheeks region are well selected as well as regions of the face at the same illumination as the cheeks.

Figure 3.16 – Adaptive skin segmentation with excluding using n = 1 on Equation 3.3.



Source – Author's database, 2019

Figures 3.17 and 3.18 illustrate the effect of changing the value of n. In Figure 3.17, n is increased to 2, allowing a wider skin-region to be selected, and also ignoring most of the glasses and facial hair. Increasing even more the value of n, setting it to 3, almost the full skin-face is extracted, and most of the non-skin region is still being excluded, as presented in Figure 3.18.

To choose a valuer for n, it should be checked if it is better to take a small or a big sample of skin. A bigger region will probably include more non-skin pixels, and also have regions with different illuminations, however, bigger the amount of skin samples, more diffuse is the impact of non-skin pixels. For a small sample, the situation is the opposite: as the samples are the most similar to the cheeks in color, they will be probably well-vascularized regions of the face, allowing to get a better signal. However, smaller the number of samples, bigger is the noise-impact if a non-skin pixel is selected.

Figure 3.17 – Adaptive skin segmentation with excluding using n = 2 on Equation 3.3.

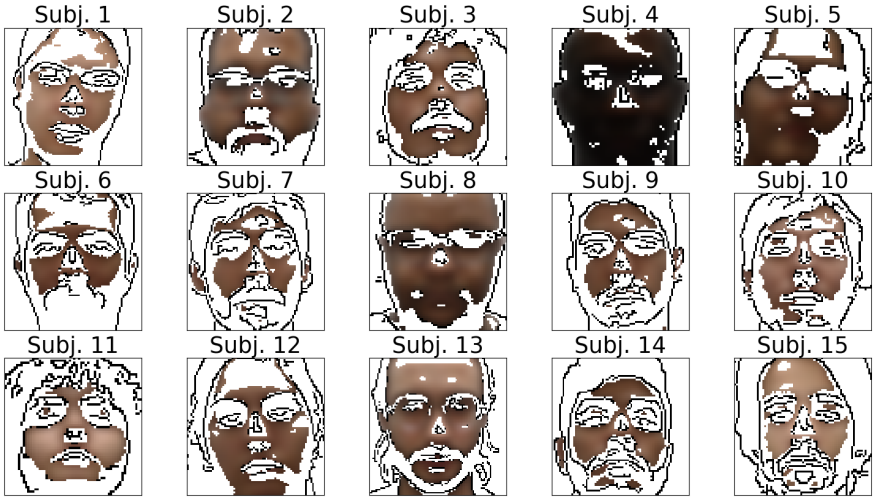
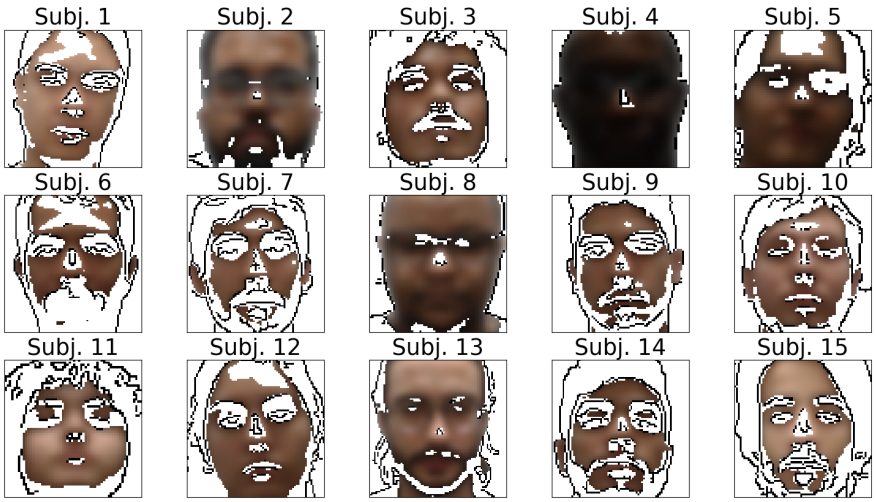


Figure 3.18 – Adaptive skin segmentation with excluding using n = 3 on Equation 3.3.



3.3.3 Evaluation of the skin extraction techniques

To evaluate the specific ROIs and the adaptive color skin-color techniques, two metrics were used: the time spent for each one, and the cross-correlation of their respective rPPG and PPG signals (with PPG as ground truth).

To estimate the rPPG signal, the POS methodology was used as it got the best results in (WANG et al., 2017a) and (UNAKAFOV, 2018). The data used were from the movement database, because it has the challenging environments for rPPG. The pseudo-algorithm to get the RGB temporal signal that was used for each skin extraction method is presented in Algorithm 1.

The test was done based on the Pseudo-Algorithm 1. The adaptive skin color approach from Equation 3.3 was tested for four different values of n: 0.5, 1.5, 2.5, 3.5. That way, it was possible to analyze the impact of the amount of skin extracted from the result. The method of

Algorithm 1: Pseudo-algorithm for temporal skin-RGB series extraction.

```
Result: Temporal RGB series
load the videos in the database;
select the skin-extraction method;
for video in Database do
    start counting time;
    for frame in video do
        find the face at the video;
        if face found then
            find the skin-region with the selected method;
            calculate the mean RGB values of the skin-pixels;
            add the frame index and the RGB mean values to the result pile;
        else
            pass to the next frame;
        end
    end
    stop counting time;
    save the result for the video;
    save the spent time;
end
```

specific ROIs was also tested.

The face-detector used was a combination of VJ and MedianFlow (selected based on the results of Section 3.2.1). The time spent was calculated as a sum of face-detector and skin-extraction processing times (as the first one was the same for every skin-extraction methodology, it did not affect the results).

The rPPG signals were compared to the reference PPG signal, as the rPPG is a PPG in its essence. To compare the similarity between both signals, the cross-correlation method was used[22] [23], and the method that got the highest absolute correlation value was chosen. The result is shown in Figure 3.19.

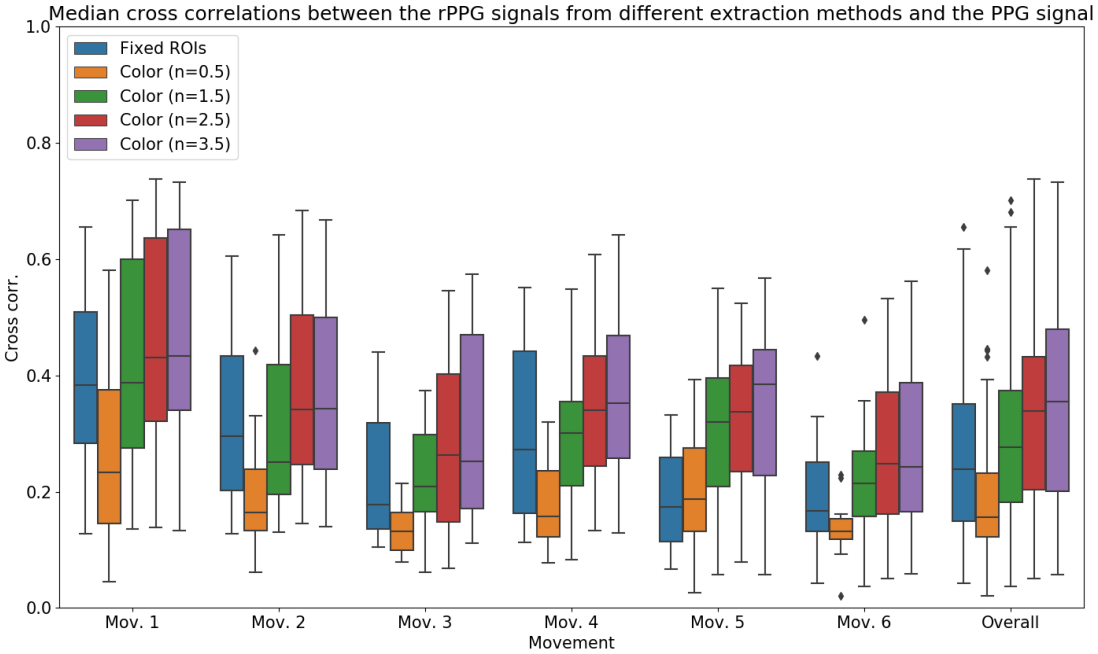
Figure 3.19 presents the correlations between the rPPG signals and the PPG, for each video of the movement database. The blue color represents the correlations using the specific ROIs skin extraction method, the orange is the correlation using $n = 0.5$ in Equation 3.3; $n = 1.5$ for the green boxes; $n = 2.5$ for the red ones; and $n = 3.5$ for the purple ones. The correlations show that for most cases, bigger the skin region extracted, more correlated was the rPPG to the PPG. So, the skin-extraction methodology chosen was the adaptive with $n = 3$, because the best values for n were 2.5 and 3.5. Thus, the mean value between them was taken.

The execution time of each setup was also evaluated. The time took by each setup to

[22] The Cross-correlation is explained at Appendix A.2.

[23] The correlation functions is already implemented at the library *numpy*(OLIPHANT, 2006). Documentation: <https://docs.scipy.org/doc/numpy/reference/generated/numpy.correlate.html>.

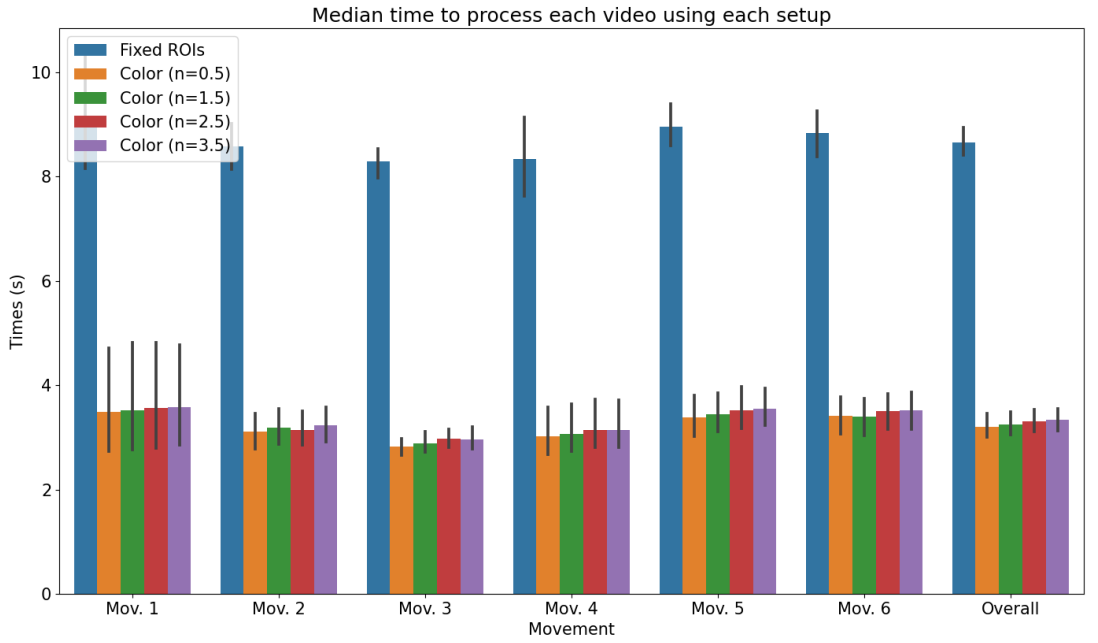
Figure 3.19 – Correlation between rPPG and PPG signals for each extraction setup.



Source – Author's database, 2020

process the videos was saved, which it is showed in Figure 3.20. It can be seen in the results that the slowest setup is the one with specific ROIs.

Figure 3.20 – Median time spent by each setup.



Source – Author's database, 2020

As expected, the specific ROIs had the slowest result because they need an AI algorithm to

run at each frame, and selecting the desired region. Looking at the spent time of the adaptive-color methods, the processing time tends to increase a bit for bigger n values (more samples to calculate the mean RGB values), however this time increasing is very small.

3.4 Remote PPG Signal Estimation

This section compares the result of the most well-rated rPPG estimation using Blind Source Separation (BSS) and linear combination based techniques found in the literature (WANG et al., 2017a; UNAKAFOV, 2018). The features measured from each technique are the computational cost and the correlation with the PPG signal. The two BSS techniques evaluated are PCA and ICA, and the two based on linear color combination are POS and CHROM.

3.4.1 Principal Component Analysis

Principal Component Analysis (PCA)[24] is a BSS method widely used for pattern recognition and dimension reduction. The main idea of the algorithm is to decompose the signal in N components that explain the maximum variance possible of the N dimensions of the signal (LEWANDOWSKA et al., 2011). The components are ordered from the highest variance to the lowest and they are orthogonal to each other (TIPPING; BISHOP, 1999).

The objective of using PCA in rPPG is to separate the pulse signal from the noise at the RGB temporal series. The method proposed by Lewandowska et al. (2011) applies the PCA algorithm to the RGB temporal samples, transforms the principal components to the frequency domain, and gets the component with the highest peak in the power spectral density (PSD) function. It supposes that for a short time the pulse information has periodic variations. Figure 3.21 exemplifies the process.

In Figure 3.21(a) are presented the temporal signals with the skin-pixels RGB intensities at each frame summed. Figure 3.21(b) shows the three components after applying PCA to the RGB series. Figure 3.21(c) presents the spectra of the three principal components. All of them have a peak in 1 Hz; the first and third components are the ones with the highest peaks, but the third one is slightly higher. So, this component should be taken to represent the pulse signal. The PCA algorithm is already implemented in the library provided by Pedregosa et al. (2011)[25].

3.4.2 Independent Component Analysis

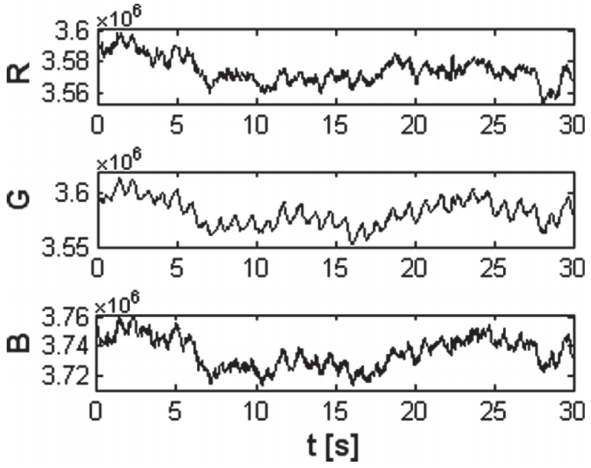
The main idea of the Independent Component Analysis (ICA)[26] is to suppose that a complex signal can be split into the weighted components of its sources. ICA is a convergence

[24] For details see Appendix A.4.

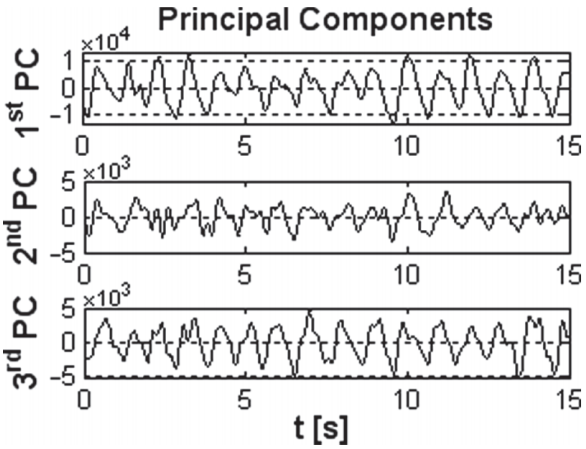
[25] Documentation about Sklearn's PCA: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>.

[26] Appendix A.5 presents more information about the ICA technique.

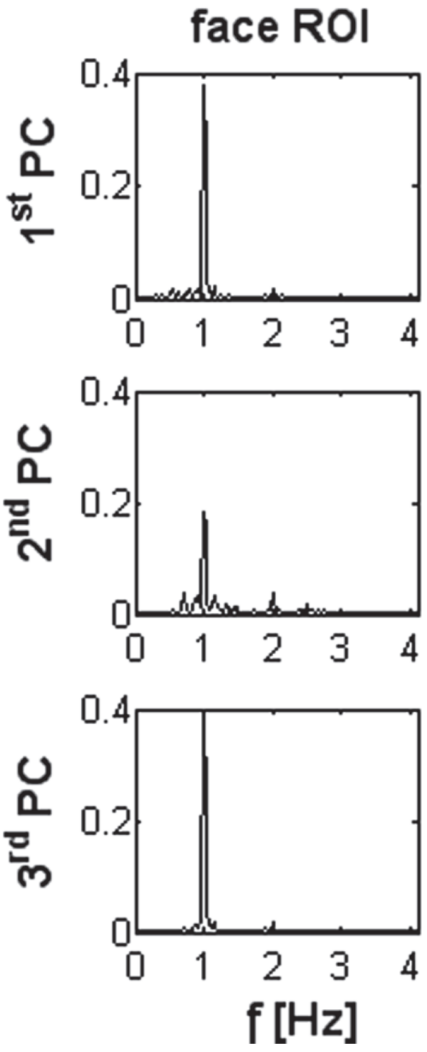
Figure 3.21 – rPPG signal generating using PCA. Adapted from (LEWANDOWSKA et al., 2011).



(a) RGB temporal signal.



(b) PCA components of the RGB signal.



(c) Spectra of the PCA components.

Source – LEWANDOWSKA et al., 2011

algorithm. It tries to find the optimum weights for the unknown sources. This makes ICA slower than PCA, but normally it can separate the signals better.

The methodology to separate the pulse signal from noise is the same using ICA or PCA: the PSD of the components is calculated through FFT, and the one with the highest peak is kept chosen to represent the rPPG[27].

3.4.3 Chrominance-Based rPPG

The Chrominance-Based rPPG (CHROM) was proposed by Haan and Jeanne (2013). Its main objective is to increase the robustness of the rPPG to the user motion. CHROM assumes

[27] See Appendix A.3 for more information about the Fourier transform.

that the light reflected by the standardized-skin over time is represented in Equation 3.5.

$$C_i = I_{C_i}(\rho_{C_{dc}} + \rho_{C_i} + s_i), \tag{3.5}$$

where C_i is one of the RGB channels, $C \in \{R, G, B\}$, I_{C_i} is the intensity of the light source, $\rho_{C_{dc}}$ is the stationary part of the reflection coefficient of the skin at the respective channel, ρ_{C_i} is the zero-mean time-varying signal representing the pulse signal and s_i is the additive specular reflection contribution, which is identical for all the channels when a white color is illuminating the skin.

Assuming that I_{C_i} affects proportionally the three channels, the series can be normalized to generate a signal independent of the intensity of the light source. The temporal signal C_i is then normalized, dividing every channel by its mean value and subtracting a unity (the mean operation is represented by the character μ). C_{ni} is the normalized signal, presented in Equation 3.6.

$$C_{ni} = \frac{C_i}{\mu(C_i)} - 1 \tag{3.6}$$

To suppress the motion artifacts of the signal, and keep the variations due to blood pulse, Haan and Jeanne (2013) define two chrominance signals, standardized for different skin-tones[28], X_s and Y_s, as presented in Equation 3.7.

$$X_s = 3R_n - 2G_n;\ Y_s = 1.5R_n + G_n - 1.5B_n \tag{3.7}$$

The signals are then filtered with a pass-band filter using the common human heart rate range as limits (0.67 - 4 Hz). The result is represented by X_f and Y_f. To obtain the pulse signal, P, and deal with the in-band disturbances, the authors propose Equation 3.8.

$$P = X_f - \alpha Y_f \quad \text{where:} \quad \alpha = \frac{\rho(X_f)}{\rho(Y_f)}; \quad \rho(.) \text{ is the standard deviation operation.} \tag{3.8}$$

The rPPG signal is the result from Equation 3.8.

3.4.4 Plane Orthogonal to Skin

In (WANG et al., 2017a), a detailed study about the reflectance of the skin related to the rPPG is done to evaluate different parameters that impact the changes in the skin-color besides the blood volume, as light intensity variation, motion and skin reflection[29]. Based on this skin-model, the authors suggest a Plane Orthogonal to Skin (POS), an alternative to CHROM, where the plane direction enhances the strength of the blood-pulse signal. The equation to get the rPPG signal is very similar to CHROM, presented before. The difference is in the calculus of signals X_s and Y_s. The new weights of them are shown in Equation 3.9.

$$X_s = G_n - B_n;\ Y_s = -2R_n + G_n + B_n \tag{3.9}$$

[28] Details about the procedure can be found in (HAAN; JEANNE, 2013).
[29] Details of the study can be seen in (WANG et al., 2017a).

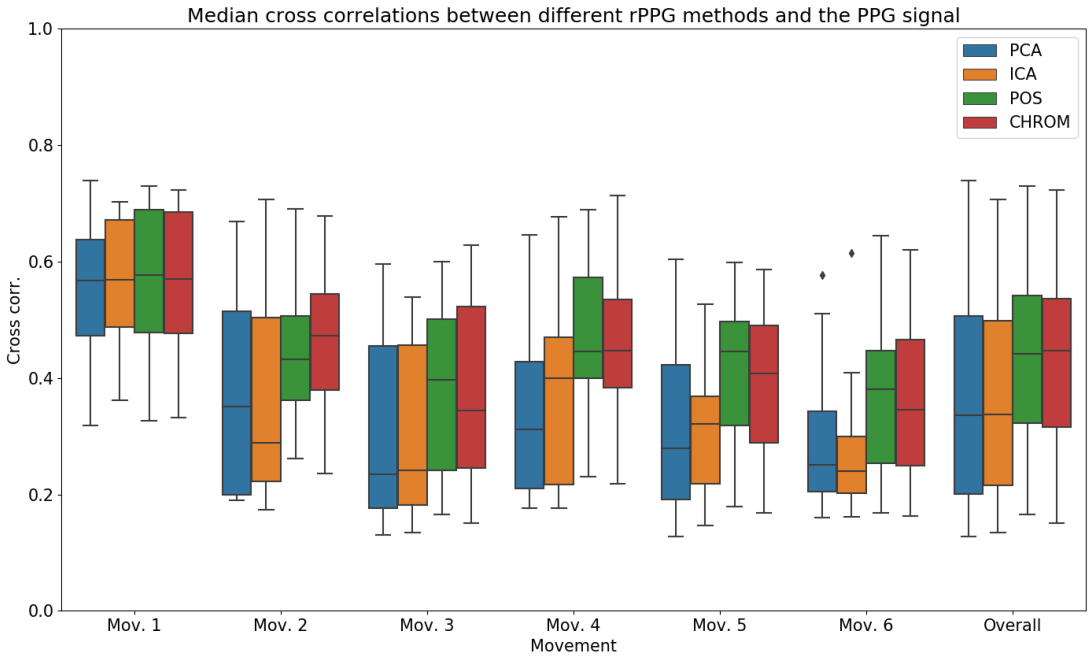
The procedure to get rPPG is almost the same as in CHROM: first, the signal is normalized with Equation 3.6. Next, the orthogonal signals are calculated with Equation 3.9, and then the pulse signal is generated using Equation 3.8[30].

3.4.5 Remote PPG methods Evaluation

As in the skin extraction part, the pulse signal was estimated using four different rPPG methodologies, and the one that generated the most correlated signal to PPG was chosen. There was also the condition of the processing time, in which the algorithm should not take too long to process the signal to be able to work online.

The movement database was used in this test. VJ and MedianFlow were combined to find and track the face. Equation 3.3 with $n = 3$ was applied to find the skin-pixels and then the mean RGB values were extracted to get the raw signal. The initial idea was to generate rPPG using the whole RGB series. However, because the ICA and the PCA work better for periodic signals, and in a long RGB series the heart rate frequency may have considerable changes, the whole signal was split. Then, to generate the rPPG signal, the 1-min raw RGB signals was split into six 10 s segments. The rPPG is calculated with the segments and compared to its respective PPG signal. The correlation of the signals are presented in Figure 3.22.

Figure 3.22 – Correlation of each method.



Source – Author's database, 2020

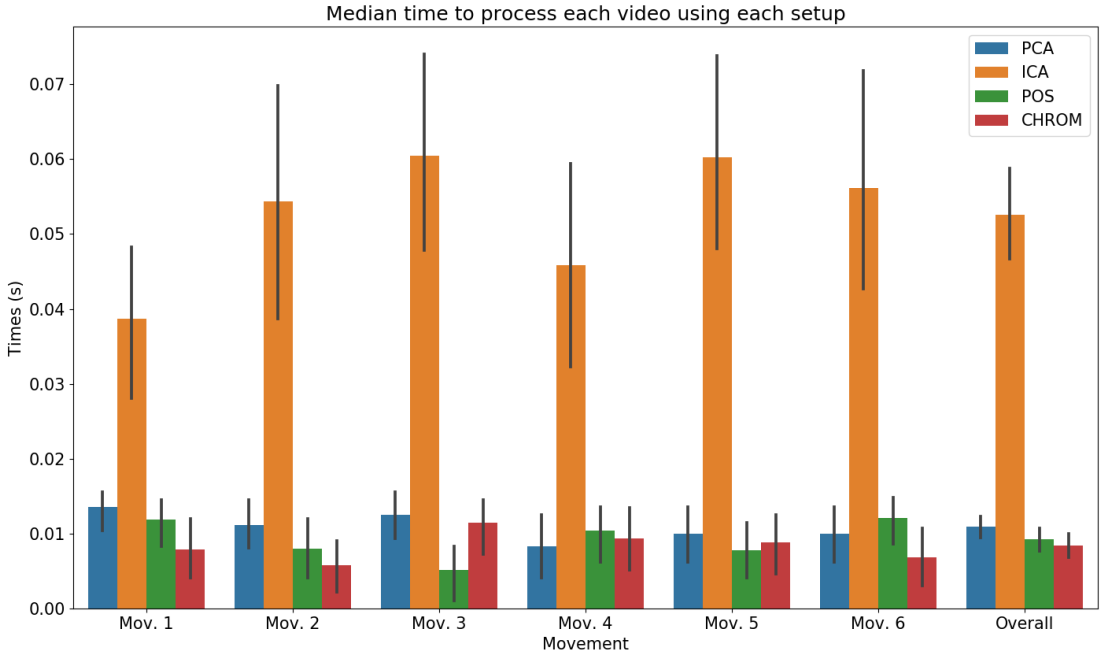
At the graph, the blue boxes represent the PCA correlation values; the orange ones represent the ICA; the green ones the POS; and the red ones the CHROM. For all movements

[30] The full algorithm cam be seen in (WANG et al., 2017a) page 1485, as *Algorithm 1*.

the best median correlations were achieved using POS or CHROM, agreeing with (WANG et al., 2017a) and (UNAKAFOV, 2018). Both got very similar results. The median correlation of CHROM was slightly higher than POS, but the maximum and minimum values of POS were a bit higher when the overall result is analyzed[31]. For movement 1, with the static volunteers, the four methods presented very similar correlations, but the BSS methods were worst when the subjects were moving.

Figure 3.23 shows the time that each method spent to process the whole RGB signals. POS, CHROM and PCA had a similar processing time, spending about 0.01 s to process each video. ICA took the longest processing time, about 0.05 s.

Figure 3.23 – Total time spent by each setup.



Source – Author's database, 2020

POS and CHROM got the best correlation results (agreeing with the results from Wang et al. (2017a) and Unakafov (2018)) and also short processing time. Both had practically the same correlations, so both are suitable. As the overall maximum and minimal values and also the limits of the standard deviation of POS were a bit higher than CHROM's, and it was better evaluated at the above papers, POS was the chosen method to be used at the final system.

3.4.6 Filtering

The literature presents some techniques to improve the rPPG signal. In this section, some of them will be explained and evaluated in relation to processing time and correlation of the

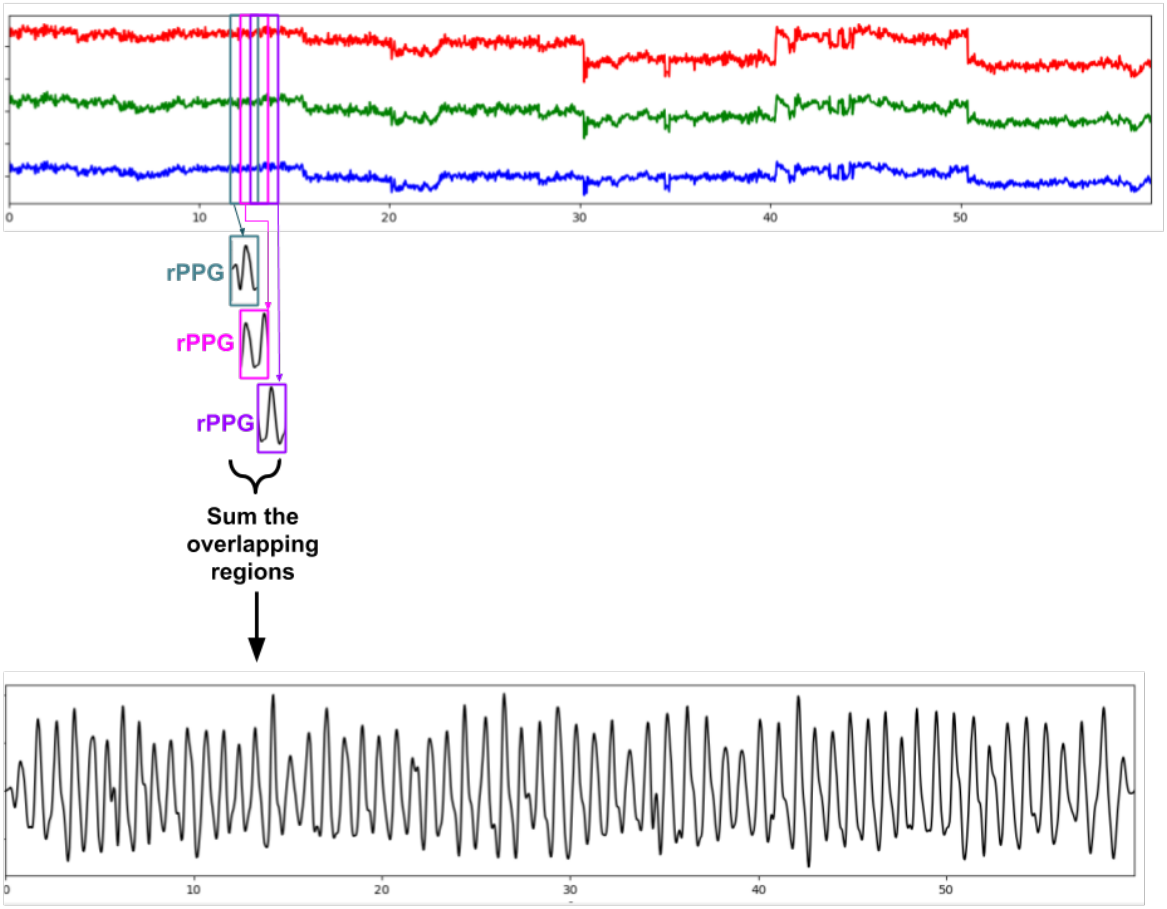
[31] As the selection of the face tracking and skin-pixels were done using the POS technique, there is a probability of it has a better result because of it. However, due to the high complexity of testing all the combinations (and also the better results achieved by POS in the literature), this possibility was not considered.

generated signal with PPG.

3.4.6.1 Overlap-Adding (OA)

At the methodology from Haan and Jeanne (2013), the rPPG signal is calculated using a sliding window, the overlapped regions of consecutive windows are added after rPPG signal is calculated using the RGB signal inside the window, and them, normalized by subtracting its mean value divided by its standard deviation. According to the authors, a short window size is preferable because it can suppress instantaneous distortions and avoid the influence of unwanted low-frequency noise, as the respiration signal. The minimum window size suggested by the authors is the length of one cardiac cycle, and the window should slide one frame each time. The common human HR range is from 40 bpm to 240 bpm. So the longest period for a cycle is 60/40, which is about 1.5 s. The sample rate (SR) of the camera is 30 fps, which means that in 1.5 s about 45 frames are recorded, so the chosen length for the overlap-adding window is 45 samples. The process is illustrated in Figure 3.24.

Figure 3.24 – Illustration of the overlap-adding technique.



Source – Author's database, 2020

The RGB temporal series is represented by red, green and blue lines on the upper graph.

The colored rectangles represent the RGB raw signal inside the window and the rPPG estimated using it. On the other hand, the bottom black graph represents the rPPG signal after finished the window procedure for the whole signal.

3.4.6.2 Sub-band processing (SB)

Presented in (WANG et al., 2017b), this procedure aims to improve the robustness of the rPPG signal to powerful noises inside the human heart rate frequency-band, e.g. the noise from the user movement. The authors claim that with three color channels, only two independent distortion sources can be eliminated. However, in real situations, there are more than two sources of noise to affect the signal, as multiple light conditions or different movements, where each disturbance source may add noise in different frequencies at the signal.

The proposed method exploits the fact that the pulse and motion components have the same frequency in different color channels. So the principle is to separate the frequency components inside the color channels, calculate the rPPG separately at each frequency bin (or frequency sub-band) and then combine them into a final signal. The methodology to calculate the rPPG should be a linear combination of the channels. Using this processing, the features of the pulse and noise signal may be exploited here to enhance the pulse influence and suppress the noise.

As the pulse signal has a small amplitude, the color variation caused by light condition changes or user motion should be much greater, which will increase the total amplitude of the signal. So, weighting each segment of rPPG calculated using each window before combining them in the overlap-adding process may reduce the impact of the noise. The weight is calculated according to Equation 3.10.

$$w = \frac{\sum \left(abs \left(fft \left(P \right) \right) \right)}{\sum \left(abs \left(fft \left(RGB \right) \right) \right)} \tag{3.10}$$

where:

- w is the weight calculated;
- $abs(.)$ represents the absolute operation;
- fft represents the FFT algorithm;
- RGB is the RGB series segment;
- P is the pulse signal estimated for the RGB segment;

The process is exemplified in Figure 3.25, and the algorithm adapted from (WANG et al., 2017b) is presented in Algorithm 2.The procedure is done in the frequency domain to save processing time, since matrix multiplication can be used to apply Equation 3.9, Equation 3.8 and Equation 3.10 in the frequency domain, instead of using a for loop to calculate it at each

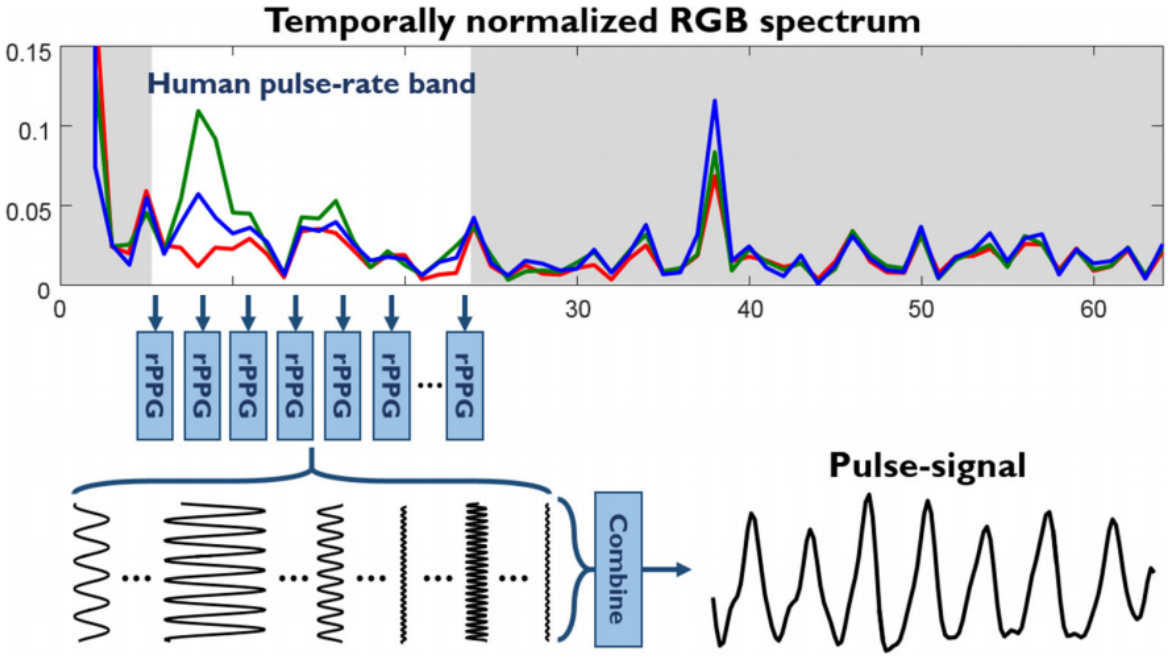
frequency bin, and the signal is normalized according to Equation 3.11 in the overlap-adding process.

$$\mathbf{X_{norm}} = \frac{\mathbf{X} - \mu(\mathbf{X})}{\sigma(\mathbf{X})} \tag{3.11}$$

where:

- $\mu(.)$ represents the mean operation;
- $\rho(.)$ represents the standard deviation operation

Figure 3.25 – Sub-band exemplification.



Source – (WANG et al., 2017b)

3.4.6.3 Narrow Band filtering (NB)

Based on the method from (GUDI et al., 2019), the idea of this technique is to compute the rPPG only in the frequency band with the most periodic signal. Similar to the Sub-band method, the RGB signal is converted to frequency domain using FFT. It is supposed that, after calculating the rPPG on each bin, the bin containing the signal with the highest power is the one representing the pulse rate. So, the rPPG is calculated separately for each frequency bin and the amplitude of all frequencies, except to the one with the greatest power, are set to zero. This will return a sinusoidal signal, making the peak detection an easier task, ideal for HRV signal extraction. It works as a narrow ideal pass-band filter in the signal (GUDI et al., 2019). The authors suggest a bin bandwidth of 0.47 Hz.

Algorithm 2: Algorithm for the Sub-band Filtering.

Input: RGB signal $3 \times N$ (N is the number of samples);
Input parameters: l (length of the segment), B = $[b_{min}, b_{max}]$ (b_{min} is the minimum frequency bin representing the heart rate, b_{max} is the maximum);
Initialize output: **P** = zeros(N) (output pulse signal);
$b = l/2 \longrightarrow$ frequency bins of the FFT;
for $n = 1,2, ...N - 1 + l$ **do**
  C = RGB(:n:n+l-1) $\longrightarrow$ Segment a part of the signal;
  C = **C**/mean(**C**) - 1 $\longrightarrow$ Normalize the raw RGB segment;
  F = fft(**C**) $\longrightarrow$ Calculate the FFT of the three color channels;
  S = [0,1,-1;-2,1,1]***F** $\longrightarrow$ Apply Equation 3.9 in the Freq. domain;
  Z = **S**(1,:)+**abs**(**S**(1,:)./**S**(2,:))*.**S**(2,:) $\longrightarrow$ Apply Equation 3.8 in the Freq. domain;
  Z = **Z**.***abs**(**Z**)./**abs**(**sum**(**F**)) $\longrightarrow$ Apply the weights;
  Z(:,1:**B**(1)-1)=0;**Z**(:,**B**(2)+1:end)=0 $\longrightarrow$ Remove the frequencies outside the human heart rate range;
  P = **real**(**ifft**(**Z**)) $\longrightarrow$ return to the time domain;
  P(1,n:n+l-1) = **P**(1,n:n+l-1) + (**P** - mean(P))/**std**(P) $\longrightarrow$ Normalize the signal and apply the overlap-adding procedure;
end

To avoid the influence of the noise from movement that may appear in specific points of the spectra, the signal is overlap-added, so the influence of high power frequencies that appear for a small time are mitigated.

The trade-off of this technique is that for a small length window at the sliding process it will be very influenced by noise, and with a big window it loses the sensibility to small heart beat frequency variations.

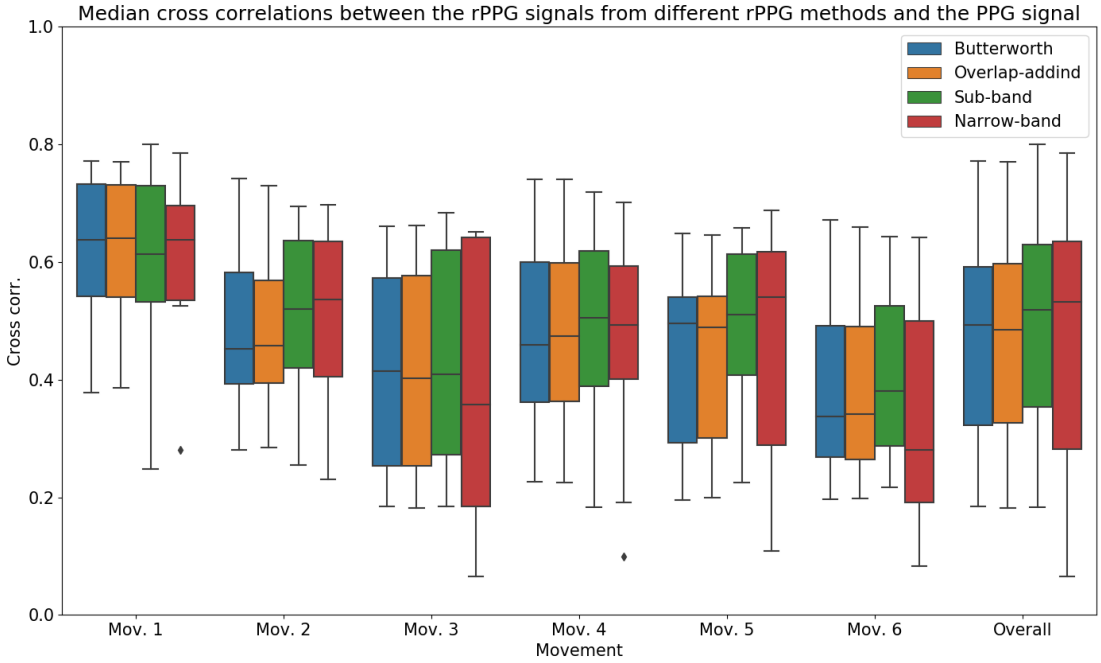
3.4.6.4 Filtering Methods Evaluation

To evaluate the suitability of the filtering techniques, like in the other tests, an rPPG signal was generated using each filtering approach and also using a 6^{th} order Butterworth pass-band filter with cuttoff frequencies at 0.66 and 4 Hz. The cross-correlation of each signal and the corresponding PPG is presented in Figure 3.26.

In Figure 3.26 the blue box represents the signal filtered only with the Butterworth filter; the orange box are the results from the signal processed with the overlap-adding technique and the Butterworth filter; the green bar is the Sub-band method; and the red the Narrow-band. The correlations between the filtering techniques are very similar. The median correlation of the Sub-band and the Narrow-band methods presented higher median values, but the variation of the Narrow-band method is higher.

Looking at Figure 3.27, that presents the time that the techniques spent to process each video. There is a big difference between the methods. The Butterworth filtering is the fastest, as it took 0.02 s to process a 60 s signal; the Overlap-adding got almost 1 s; the Sub-band is faster, with

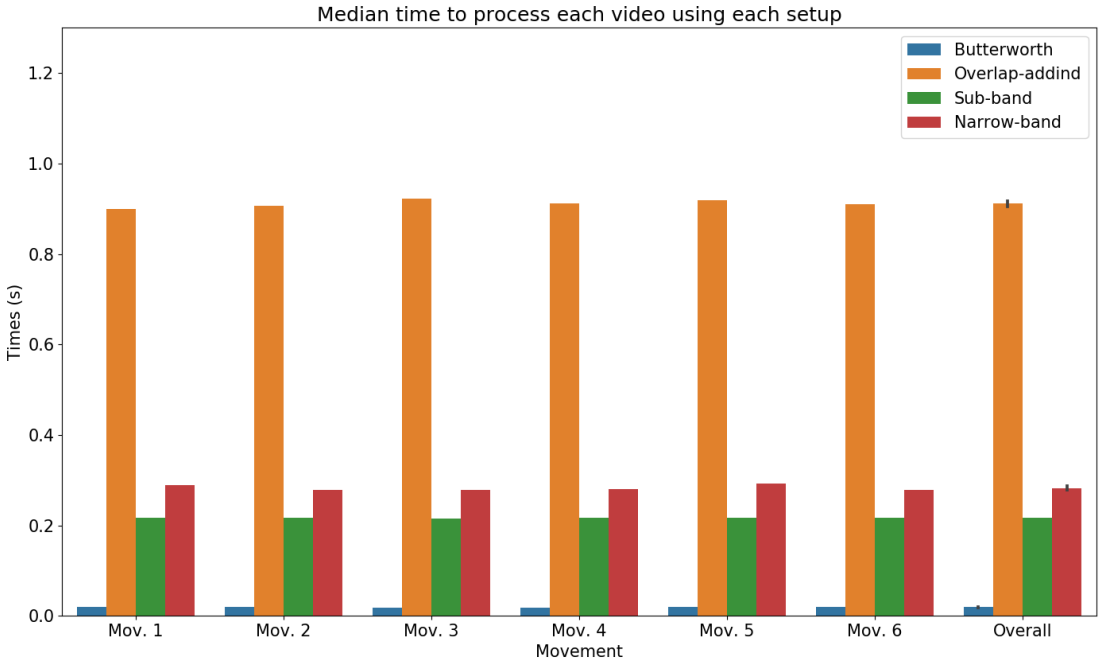
Figure 3.26 – Correlations between the rPPG of each filtering method and PPG.



Source – Author's database, 2020

about 0.2 s; and the Narrow-band spent 0.3 s. Besides the time difference, each of them seems to be suitable for an online technique, as all of them took less than 1 s to process a 60 s signal. To choose the best technique, more tests are needed since looking only at the correlation between rPPG and PPG with all methods the results were very similar. They will be again evaluated in the next chapter by precision of the peak location.

Figure 3.27 – Time spent by the raw and the enhancing techniques.



Source – Author's database, 2020

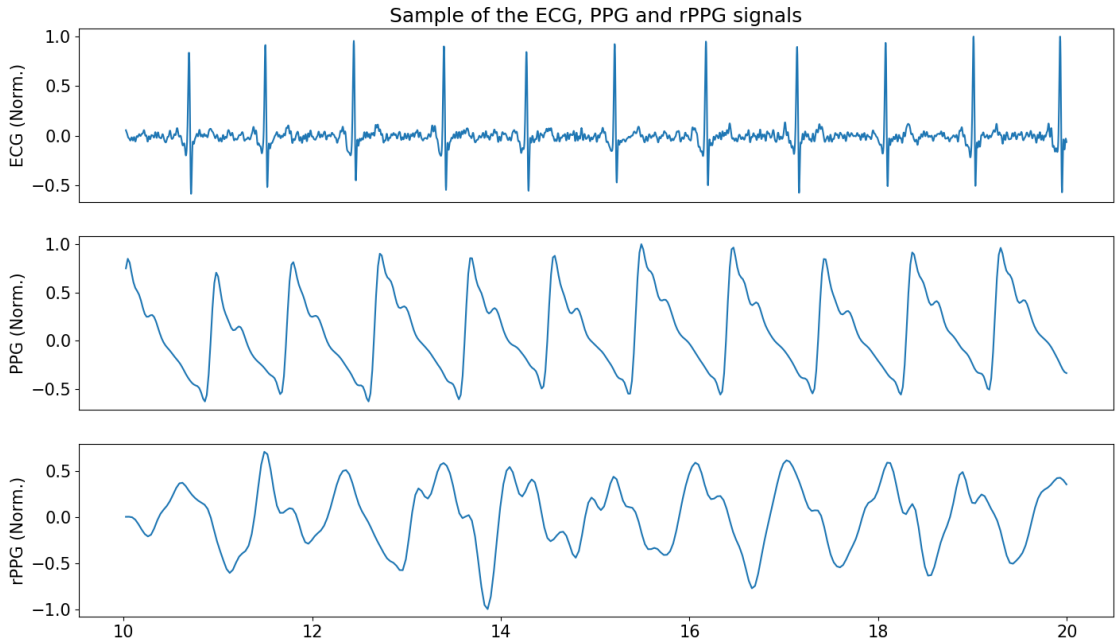
4 Results

4.1 Heart Rate and Interbeat Signal Estimation

The two cardiac parameters measured with the rPPG are the heart rate and the interbeat difference. In the studied literature, the main methodology used to estimate the heart rate from rPPG signal is a Power Spectrum Density (PSD) analysis. A frequency-time transformation is used in an rPPG segment of 10 to 20 s, and the frequency with the greatest peak in the human pulse rate frequencies band is selected to be the estimated HR for the window.

The inter-beat signal is based on peak detection. On the ECG signal, it is represented by the difference between consecutive R-R peaks. These peaks are normally easy to locate if there is not too much noise. For PPG, the peaks are not as prominent as ECG but may also be well defined, with a bit more processing. With rPPG this task is harder. Normally, there is too much noise at the signal, and the peaks are more diffuse. To exemplify, a sample of the ECG, PPG and rPPG signals are presented in Figure 4.1. It's possible to visualize that the peaks at ECG and PPG signals are easier to define than the rPPG's. For visualisation sake, it was chosen a sample of a static subject, when the subject is moving the rPPG signal is worse.

Figure 4.1 – Sample of the ECG, PPG and rPPG signals.



Source – Author's database, 2020

4.1.1 Frequency Analysis for Heart Rate Estimation

To estimate the heart rate using the rPPG signal, a common approach is to take the most powerful frequency on the rPPG spectra inside the band of 0.67 to 4 Hz (that represents the human common heart rate frequency). In this work, three algorithms were tested to get the frequency spectrum of the rPPG signal: the Fast Fourier Transform (FFT), the Welch Periodogram (WP) and the Lomb-Scargle Periodogram (LSP). One of the most used time-frequency transforms is the FFT because of its low computational cost, ideal for online applications. Welch's method divides the whole signal into overlapping sections, computes a modified periodogram for each part and average the overlapping regions. This method is more robust to noise but also reduces the resolution of the signal (WELCH, 1967). As both, FFT and WP, require that the signal has a constant sampling time, it should be interpolated as the sample rate from the camera may change a bit. Also it is possible that some samples are lost because VJ may not find the face or the tracker loses the target in some frames, or even the skin region may not be extracted, introducing gaps at the signal. With interpolation, a disturbance may be added to the rPPG signal. To avoid interpolate the signal, an approach capable to estimate the PSD of unevenly sampled signals may be used: the Lomb-Scargle Periodogram (LSP), created by Lomb (1976) and modified by Scargle (1982). This technique focuses on analyzing astronomic data, to detect a weak periodic signal hidden into noise. This approach has been used to analyze the rPPG in (CHEN et al., 2018).

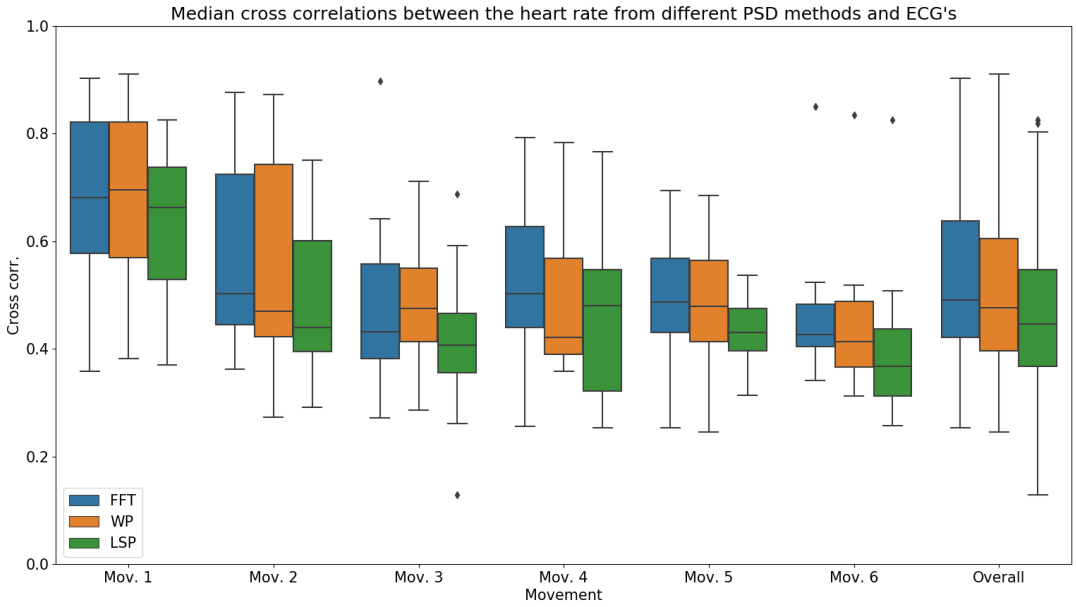
To test the three methods, the heart rate was calculated using each of them. The rPPG was extracted from the videos using the POS methodology and then, to estimate the heart rate, the FFT, the WP and the LSP were used.To apply the techniques, a 10 s sliding window was used to estimate the heart rate over time. The window is applied for each new sample, defining the overlap size as 299 samples ($10\text{ s} \times 30\text{ fps} - 1$ sample). Applying the setup the result to the 60 s rPPG signal, the result is a heart rate curve with length of 50 s (the first estimation is done after taking a sample of 10 s). To evaluate the result the heart rate curve is also calculated with the ECG. To perform it, the R-peaks on the ECG are identified using the library provided by Carreiras et al. (2015), them, taking the inverse of the time difference between consecutive peaks multiplied by 60 to take the bpm values of the curve.

To compare the results the ECG heart rate curve is interpolated to the same frequency sampling of the rPPG's heart rate curve (30 Hz), them, the cross-correlation between the heart rate found with the rPPGs and the ECG and their root mean squared error (RMSE)[1] were calculated, together with the processing time. The ECG is the gold standard method to extract the heart rate, so, because of it, the result was compared with the ECG instead of the PPG. Figure 4.2 presents the box-plots of cross-correlation found by getting the HR signal from the rPPG and the ECG using a 10 s window, and Figure 4.3 presents the RMSE by calculating the heart rate using the rPPG comparing the difference with the ECG.

In Figure 4.2, the blue boxes are the cross-correlation values of the rPPG's and ECG's

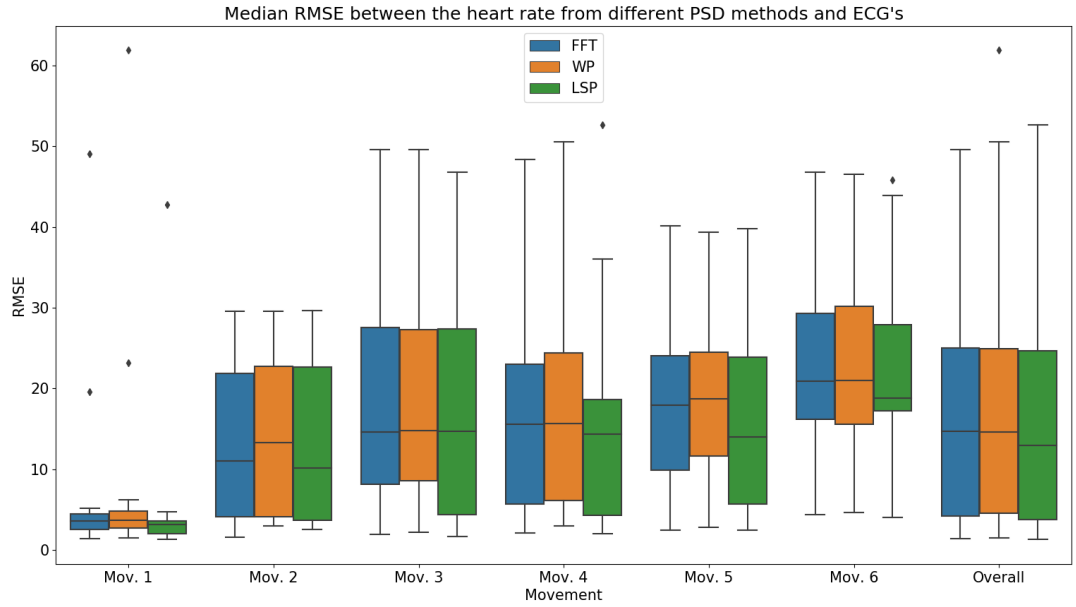
[1] See Appendix A.6.1 for details about the RMSE measurement

Figure 4.2 – Correlation of the heart rate curves from rPPG and ECG.



Source – Author's database, 2020

Figure 4.3 – RMSE of the heart rate curves from rPPG and ECG.



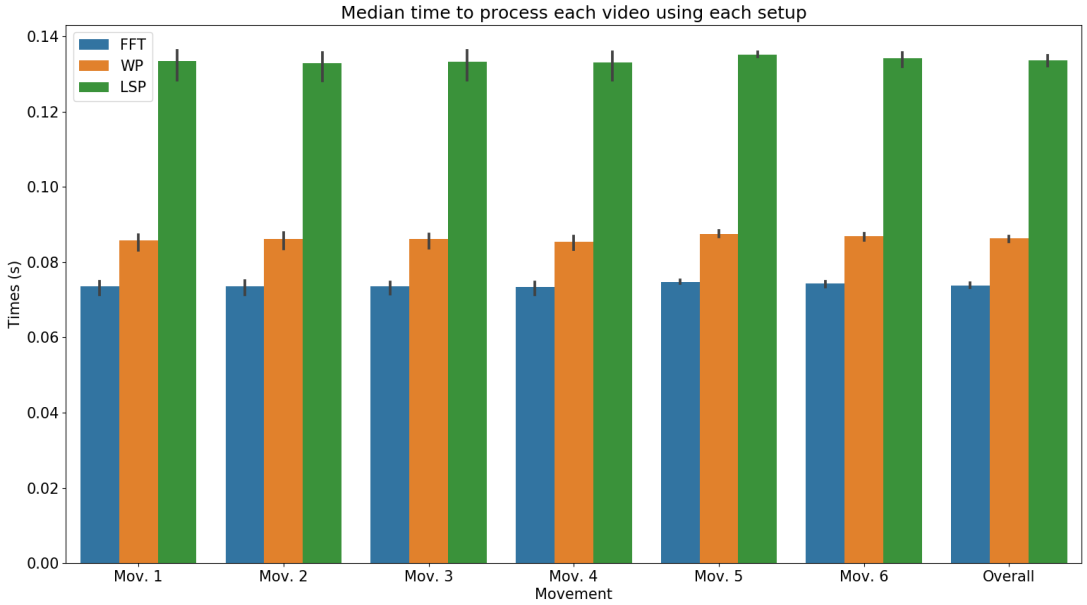
Source – Author's database, 2020

heart rate curves using the FFT; the orange boxes are the correlation achieved with the WP; and the green ones are the result obtained with LSP. The best correlations are achieved using the FFT algorithm, the correlations achieved with the Welch periodogram are close, however, a bit worse then the FFT. And the Worst results are achieved using the LSP method.

Looking at the RMSE values in Figure 4.3, the three algorithms got very close results,

being the LSP the one with the smallest error, besides getting a correlation a bit worse than the other two methods. But also, for most of the videos, the maximum error difference between the algorithms was of 2 bpm, except for Movement 4 that the LSP got a significantly smaller error (4 bpm less comparing to the FFT's result, and 5 bpm comparing to the Welch's). It is also notable in Figure 4.3 how much the movement affects the rPPG. At Movement 1, almost all error values were under 5 bpm, while in the other movements the median values of the error were over 10 bpm. The final feature that must be taken into account is the processing time. The median time that each method took for each video is presented in Figure 4.4.

Figure 4.4 – Total time spent by each setup.



Source – Author's database, 2020

As expected, the fastest was the FFT, then the WP and at last the LSP with almost double of the time spent by the FFT. Besides that, all three were fast enough for an online system. All three processed 60 s of signal in less than 0.15 s, however, the FFT took almost half of the LSP time.

Looking at the results, as the FFT got the best correlation (even all three being very close) and the smallest processing time, it will be used at the final system to calculate the heart rate.

4.1.2 Interbeats Difference Signal

The heart rate variation is defined by the time-differences between consecutive heart-beat cycles. This signal parameter is relatively easy to extract at the ECG, as the R peaks are normally the most prominent peak at the signal, as presented in Figure 4.1.

The ECG *Python* library *BioSPPy*, provided by Carreiras et al. (2015) was used to locate the R peaks at ECG. To locate the peaks in PPG, a 7^{th} order band-pass Butterworth filter,

with cutoff frequencies at 0.3 Hz and 8 Hz was used to eliminate noise and enhance the peaks. Then a peak finder function provided by the *Scipy* library (VIRTANEN et al., 2020) was used. This function accepts parameters to prune unwanted peaks, as the minimum distance between consecutive peaks, their prominence or their minimum height[2]. For the PPG peaks, only two parameters were used: the distance between the peaks, as a minimum of 0.25 s (time period at 240 bpm), and the minimum prominence, that was set as 30% of the whole signal's amplitude.

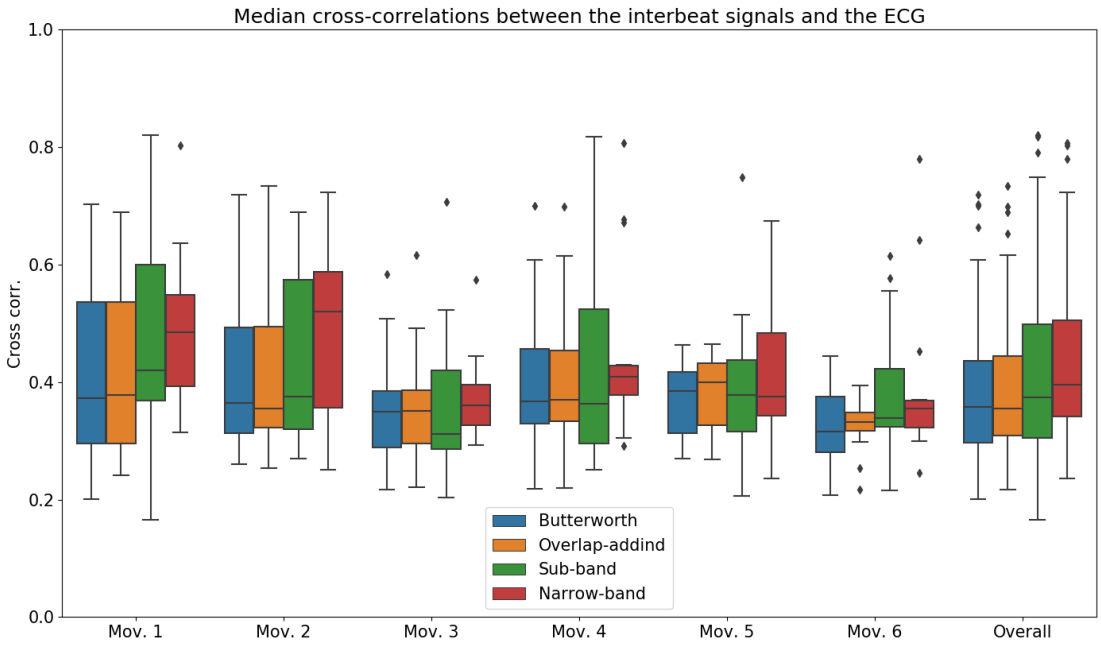
The peak location using rPPG signal is a harder task. The raw signal is normally very noisy, thus a hard filtering technique must be used to clean the signal. The four filtering techniques from Section 3.4.6 were tested together with the POS rPPG methodology to estimate the signal. Its peaks were located and the peak-to-peak difference was calculated and compared to ECG. To locate the peaks using rPPG, the same peak detection function of PPG was used, however with other parameters. First, the rPPG signal was split into segments of 128 samples (about 4.27 s). As the amplitude of the rPPG variates too much, a fixed value of amplitude at the peak finding function ignored a lot of peaks. For each segment, the minimum amplitude pruning parameter was set as 30% of the amplitude of the segment. The minimum distance was defined as the period of the maximum frequency in the window (the probable heart rate) plus 0.14 Hz (8.4 bpm). This procedure limits the minimum distance between adjacent peaks based on the average heart rate, it avoids possible noise that may appear between peaks. These values were defined empirically during the tests.

To evaluate the results, the peak-to-peak difference of each signal was calculated and correlated with the R-R difference from ECG. Before the cross-correlation calculus, the signals were interpolated to a sample rate of 7 Hz, as advised by Taskforce (1996). Some important features used to define emotions were calculated using the power spectrum density (PSD) for analysis of the interbeat difference signal. So, the PSDs from the rPPG and ECG R-R signals were correlated. Figure 4.5 presents the correlation of the rPPG interbeat signals and the ECG's, and Figure 4.6 shows the correlation of their PSD.

The signal with the highest correlation was the result of Narrow-band filtering, however, it was only slightly better than the others. However, looking at the result of the PSD correlation in Figure 4.6, the Narrow-band filtering scored significantly better than the other methods. So, this method will be used to generate the rPPG HRV features.

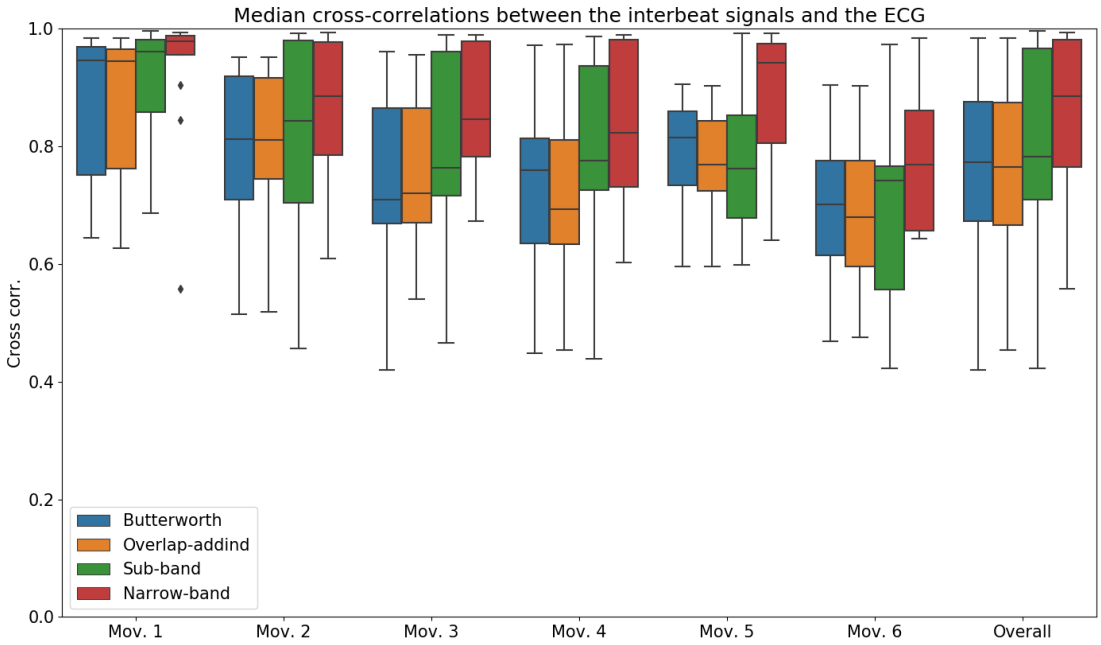
[2] for more details see: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.find_peaks.html>.

Figure 4.5 – Correlation of the interbeat signals.



Source – Author's database, 2020

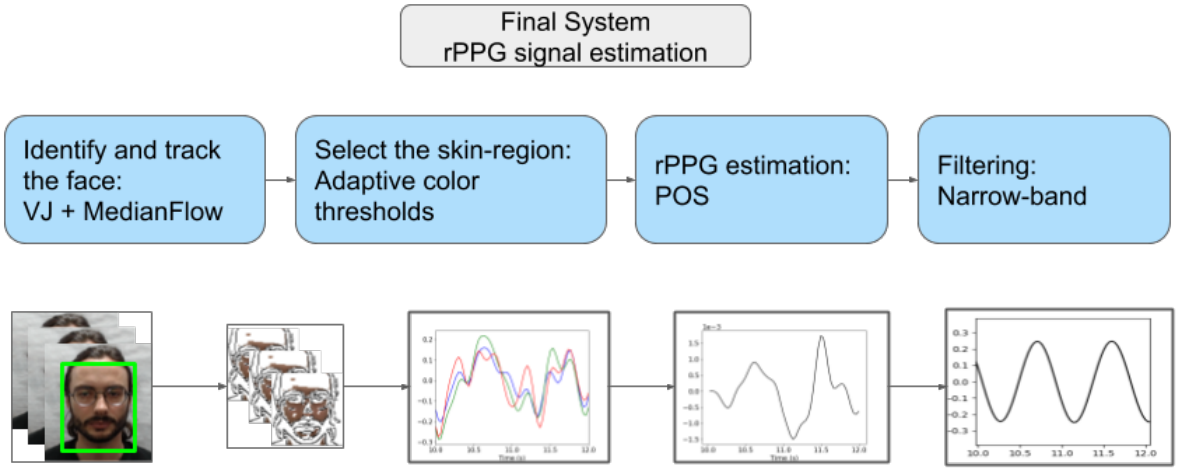
Figure 4.6 – Correlation of the interbeat PSD.



Source – Author's database, 2020

That evaluation finished the rPPG signal estimation procedure. Figure 4.7 summarizes all the algorithms chosen to perform the rPPG signal estimation.

Figure 4.7 – Algorithms chosen to perform the rPPG signal estimation.



Source – Author's database, 2020

4.2 Emotion Classification

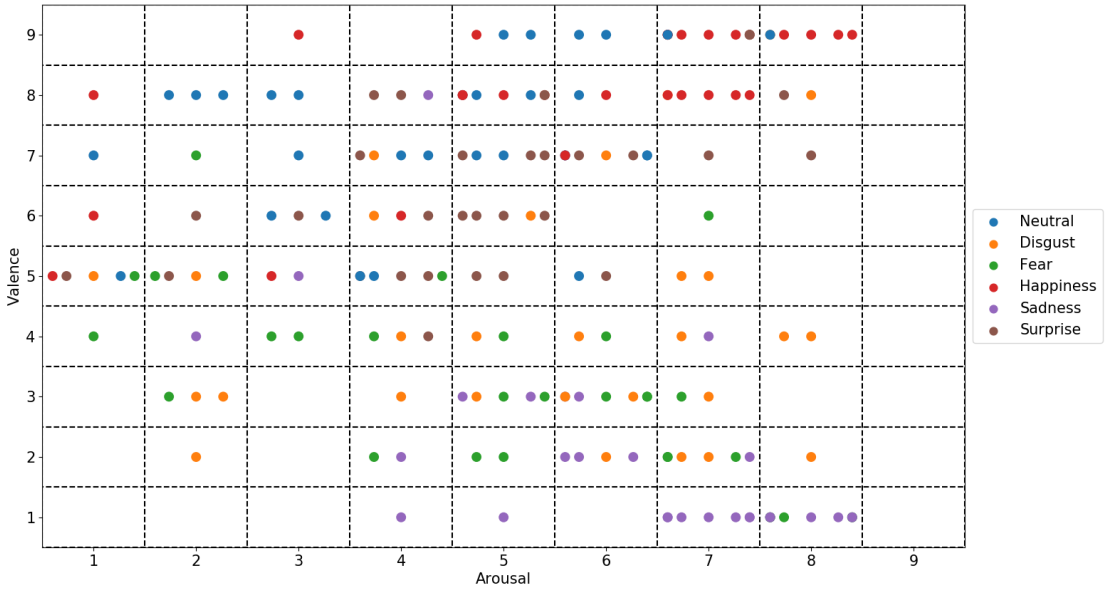
The database used here to evaluate the suitability of using rPPG as an indicator of emotional states was the Emotion Database detailed at Section 3.1.2. At this database, 30 volunteers were told to watch six short videos with emotional content (one neutral video, one disgusting, one with fear, one with happiness, one with sadness and one with surprise), and then asked to answer a questionnaire about the arousal and valence excited by each video using Self-Assessment Manikins (SAM). The videos are that the used in (GOULART et al., 2019a), they were select by a psychologist to trigger the emotions on children. This database contains face-videos, ECG and also PPG for each subject.

To evaluate the emotion content experienced by the volunteer when he/she was watching the videos, the volunteer rated each one for arousal and valence using the aforementioned SAMs questionnaire, as done in (GOULART et al., 2019a). Morris (1995) proposed this questionnaire to evaluate arousal and valence with discrete numbers from one to nine and also representing the rates with images to better explain the scale to the volunteer. The SAM scales are presented in Figure 3.6. For arousal, a value close to 1 represents a very weak emotion; close to 5 a medium intensity emotion; and close to 9, a strong emotion. For valence, a number close to 1 represents a very negative emotion; close to 5, a neutral; and close to 9, a very positive emotion.

The values that each volunteer rated the videos for arousal and valence are presented in Figure 4.8, where each one of the six videos is represented by a different color, the rates that the volunteers gave for arousal are represented in the horizontal axis and the rates for valence in the vertical axis. It's interesting to notice that the variation of the arousal rates was normally bigger than the valence ones. To better visualize it, see Figure 4.9. The wide range of arousal rates and the narrower range of the valence imply that the subjects normally agree if the video-valence triggers a bad, neutral or good feeling. But each video triggers different arousal strengths for each

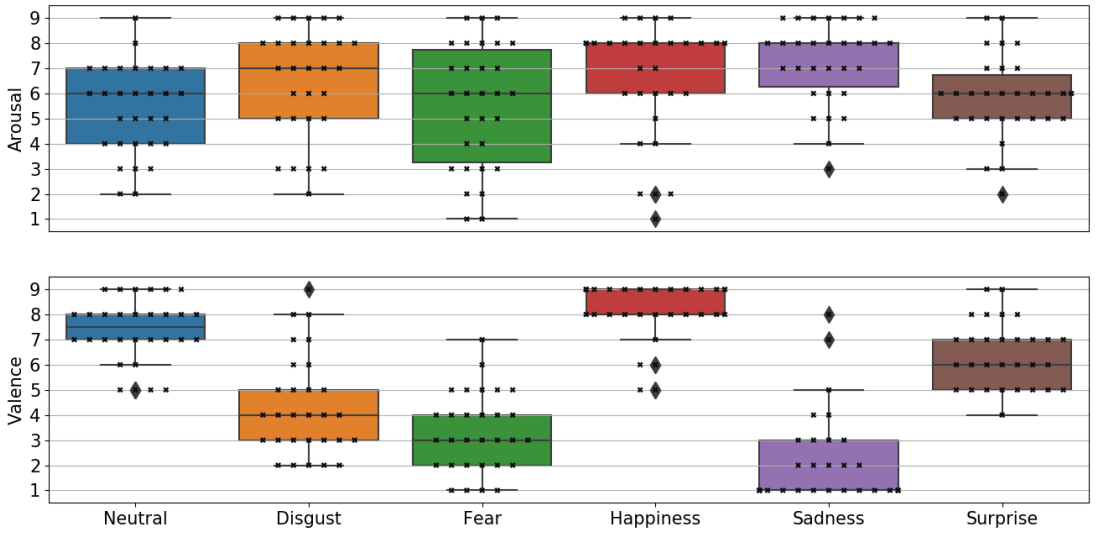
volunteer. This is more visible in the neutral, disgusting and fear videos.

Figure 4.8 – Arousal and valence overall rates distribution.



Source – Author's database, 2020

Figure 4.9 – Arousal and valence rates distribution.



Source – Author's database, 2020

The features extracted from HRV to quantify emotion were based on the work from (SCHAAFF; ADAM, 2013). The work evaluates the use of 12 different features, that are normally used for a 5 min signal (which is impracticable for an online system), for smaller lengths. They compared the use of ultra-short windows (15 s, 30 s, and 60 s) against the standard short window presented at the traditional literature (5 min) (TASKFORCE, 1996). However, as all the videos

have less than 3 min, it is impossible to use a 5 min array. The features used in that work are the 12 presented in Section 2.2.3.1. They are: $meanNN$, $SDNN$, $RMSSD$, $pNN12$, $pNN20$, $pNN50$, $SD1$, $SD2$, $SD1/SD2$, LF, HF and LF/HF [3].

For the rPPG classification, all of them were used as input for the Principal Component Analysis (PCA)[4] to select their combined values that have the highest variation. All 12 features were calculated using a 30 s sliding window and used in an emotion classification system, presented in the next section.

The rPPG feature values distribution for all volunteers is presented in Figure 4.10, and the features of ECG are presented in Figure 4.11. To represent a multidimensional feature matrix in a 2-D plot, the t-SNE algorithm implemented by was used (MAATEN; HINTON, 2008).

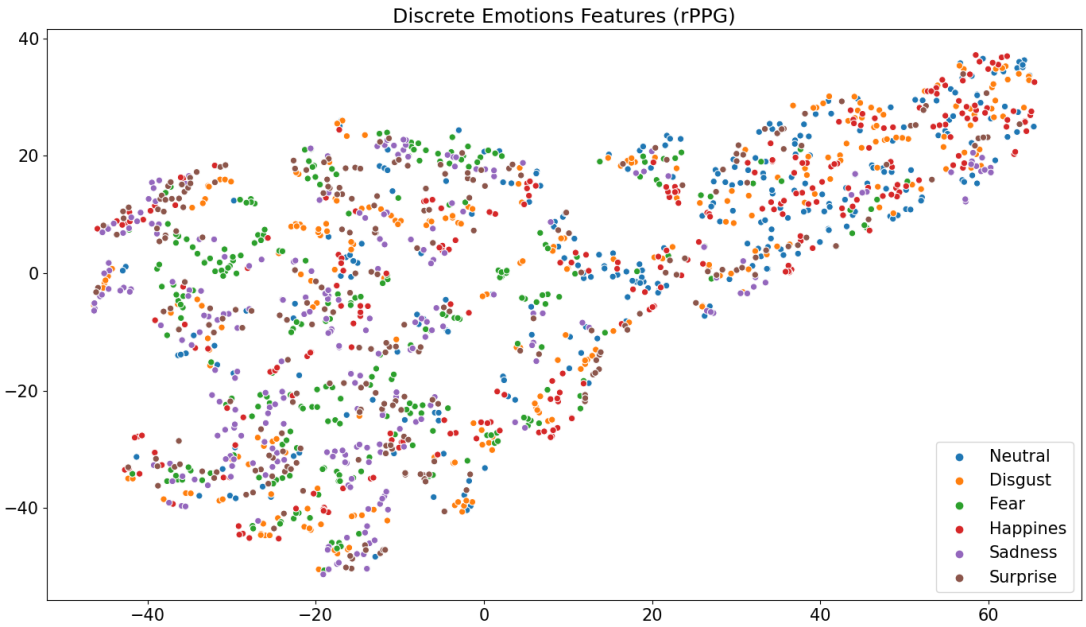
For both signals, the samples of each group are too mixed, it is not possible to identify clusters for each discrete emotion. To analyze the arousal and valence rates, the color of the points were changed to represent the arousal/valence value of each feature. The rPPG distributions for arousal and valence are presented in Figures 4.12 and 4.13, and the ones from ECG are presented in Figures 4.14 and 4.15.

It's possible to see in Figure 4.12 that for the rPPG arousal the samples are completely mixed, being very hard to define clusters for low or high rates. However, for the valence in Figure 4.13, a better visual separation may be observed. The lower grades (blue points) are more predominant at the left side; the neutral grades (black points) are spread through the graph; and the high grades (green points) valence are more concentrated at the right side. So, at a first look, the valence classification should have better results.

For the ECG data, it was expected a better separation between the classes in Figures 4.14 and 4.15. However, for arousal the samples are completely mixed up. For valence this also happens, but there is a small concentration of low grades at the right side, and the high grades are more predominant at the left side, indicating that the valence rates are more separable than the arousal ones. These statements were all made by visual inspection on the data, the classification results should provide a more reliable analysis.

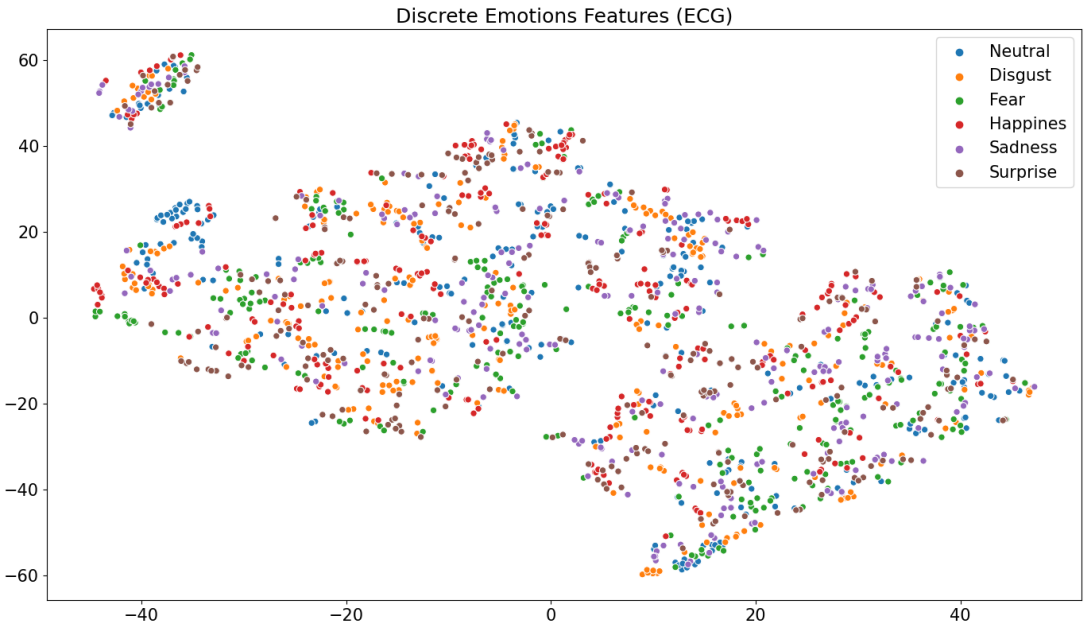
[3] Please see Section 2.2.3.1 for details about each of them.
[4] See Appendix A.4 for details about the PCA algorithm.

Figure 4.10 – HRV features distribution using rPPG.



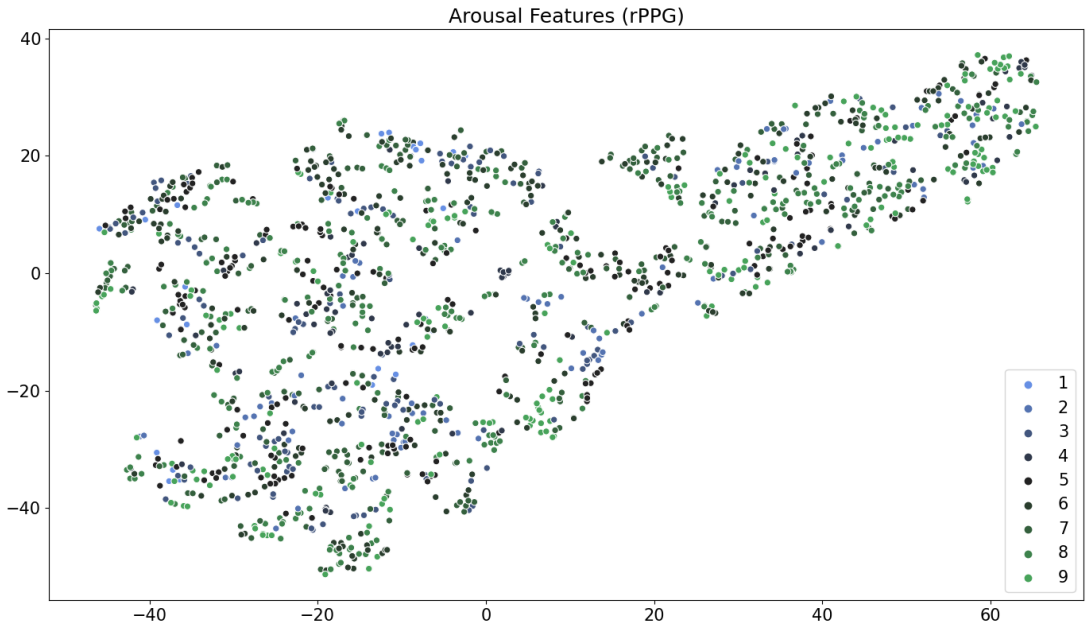
Source – Author's database, 2020

Figure 4.11 – HRV features distribution using ECG.



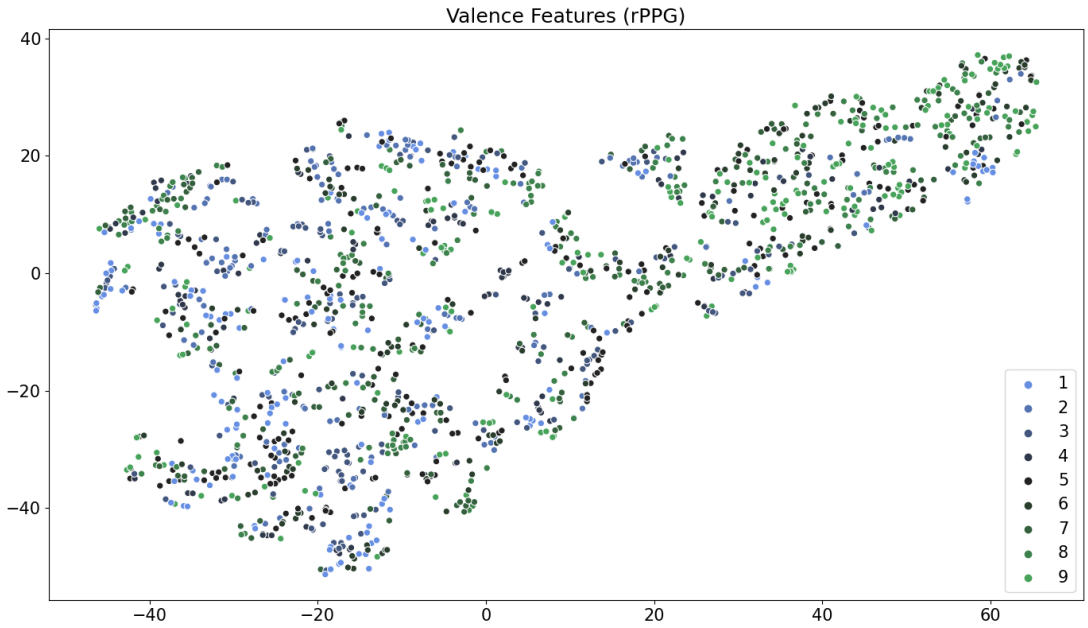
Source – Author's database, 2020

Figure 4.12 – Distribution of rPPG arousal features.



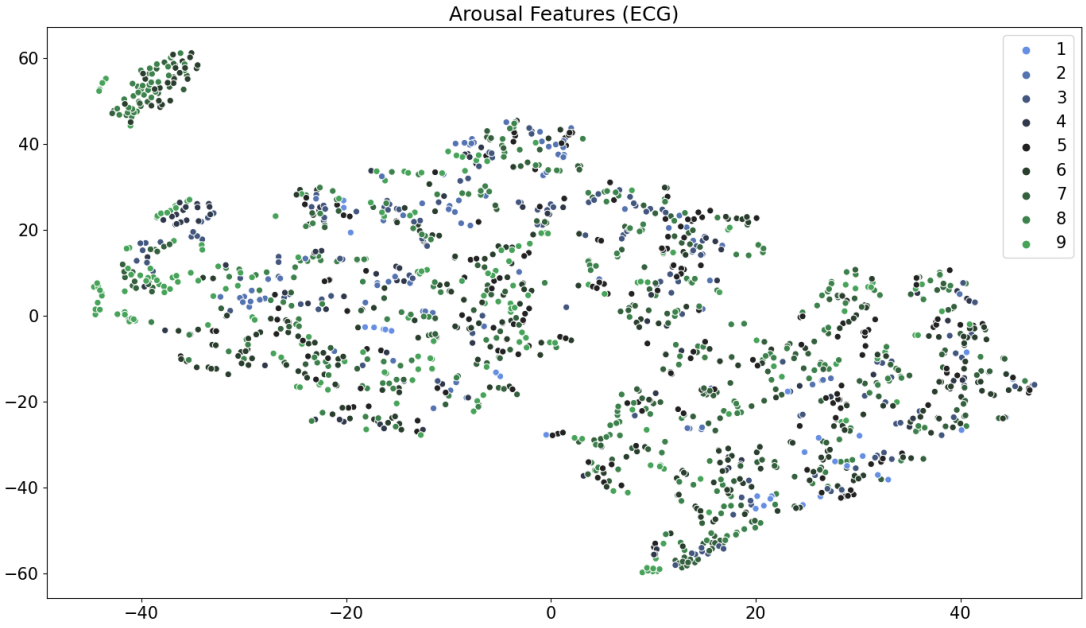
Source – Author's database

Figure 4.13 – Distribution of rPPG valence features.



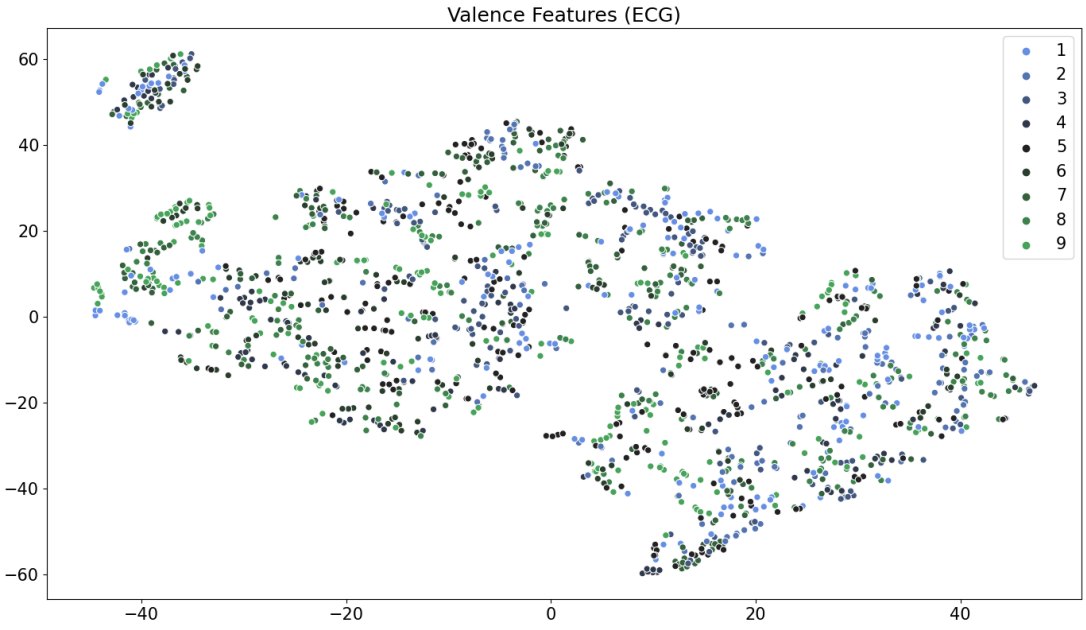
Source – Author's database

Figure 4.14 – Distribution of ECG arousal features.



Source – Author's database

Figure 4.15 – Distribution of ECG valence features.



Source – Author's database

4.2.1 Individual Emotion Recognition

The first classification setup was to create a personal classifier for each subject. There are six videos at the Emotion Database. Including the 4s countdown clip, the first is a neutral video 100.4 s long, the second a disgusting with 60.8 s, the third is a 61.8 s long scary video, the fourth

is a 43.8 s happy video, the fifth is a sad video with 134.2 s and the last is a surprising video with 85.5 s.

To calculate the HF and LF features, the minimal signal length is 30 s (SCHAAFF; ADAM, 2013). In an ideal situation, the signal would be split into non-overlapping segments of 30 s, and each segment would generate one sample. Unfortunately, the shortest video has 43.8 s, so, only one sample can be calculated that way.

To increase the number of samples from each video, the segments were overlapped. The overlapping region of consecutive windows was 28.5 s, that guarantees that at least one new beat in each segment (considering, the minimum HR as 40 bpm, with an interbeat length of 1.5 s). Using this setup, the first video generated about 46 samples, the second about 20, the third 21, the fourth 10, the fifth 69 and the sixth 37.

Classify each one of the six emotions at the Emotion Database is impossible, considering that the videos are short and the samples created using the sliding window share over 90% of the information with its neighbors since it is a 28.5 s of overlapping in a 30 s segment. So, to avoid the bias caused by the overlapped region, the videos used to train the classifier must be different from the ones to test it.

An individual classifier was tested for the system that tries to differentiate the arousal and valence rated by the subjects at each video. So, the classifier tries to predict the rates for arousal and valence given by each subject. As the values for arousal and valence using SAMs questionnaire are integers from 1 to 9, they were divided into three groups: the Low group, that goes from 1 to 3; the Neutral group, that goes from 4 to 6; and the High group, that goes from 7 to 9.

To create the training and test sets, the videos from each subject were tagged in Low, Neutral and High classes, according to the respective rate. Then, for each group, the longest videos in each class were chosen to train the classifier, and the others to test it. Since the videos have different lengths, to balance the training set, the signals were reduced to the same size by cutting off the beginning of each series until they all have the length of the shortest video duration.

It was chosen the end of the videos to balance the training dataset, because the end of an long experience (watch the whole video) will have more weight to evaluate the whole experience than the beginning of it (KAHNEMAN, 2011). So the rates that the volunteers gave for each video, probably, had a higher influence from the end of it.

The k-Nearest Neighbors (k-NN) classification algorithm was used. It is a classifier that tags a new sample in the same class as its nearest neighbor samples in the training set. The features were calculated using the 12 HRV parameters from (SCHAAFF; ADAM, 2013). As done in (GOULART et al., 2019a), PCA was calculated using the covariance matrix to reduce the dimension of the data, keeping the highest variance components until the accumulated variance of 90%.

The classification was done using ECG, PPG and rPPG signals. To evaluate the classification, the following metrics were calculated: accuracy, precision and Kappa index [5] for the classification with rPPG, ECG and PPG. The result of the arousal classification are presented in Table 4.1 and the valence results in Table 4.2.

Table 4.1 – Evaluation metrics for individual arousal classification using the three signals.

	Accuracy	**Precision**	**Kappa**
rPPG	0.26	0.34	-0.08
ECG	0.42	0.41	0.08
PPG	0.39	0.38	0.03

In Table 4.1 rPPG got the worst results, with 26% of accuracy, which does not even beat the random classifier[6], that would get 33% for three classes. The best accuracy results were achieved by ECG, which gets an accuracy of 42%, followed by PPG, which scored 39%.

Table 4.2 – Evaluation metrics for the individual valence classification using the three signals.

	Accuracy	**Precision**	**Kappa**
rPPG	0.45	0.45	0.18
ECG	0.37	0.37	0.05
PPG	0.34	0.34	0.01

The results for valence presented at Table 4.2 were much different than the arousal ones. The accuracy of rPPG got the highest result, reaching 45%, where as ECG and PPG were closer to the random classifier, getting only 37% and 34%, respectively. To better see the inter-class distribution of arousal and valence, the normalized confusion matrices are presented at Table 4.3.

Table 4.3 – Arousal confusion matrices.

	rPPG			**ECG**			**PPG**		
	Low	**Neutral**	**High**	**Low**	**Neutral**	**High**	**Low**	**Neutral**	**High**
Low	0.28	0.22	0.51	0.08	0.67	0.24	0.01	0.57	0.20
Neutral	0.43	0.29	0.28	0.11	0.47	0.42	0.11	0.56	0.33
High	0.52	0.28	0.20	0.29	0.16	0.55	0.38	0.25	0.37

In Table 4.3, the lines represent the true class of the test samples (normalized), and the columns the class that were assigned by k-NN. For rPPG, the Low and High samples were tagged oppositely: the Low samples were mostly tagged as High and the High as Low. For ECG and PPG classification, the best results were with the Normal and High classes.

[5] Details about the classification metrics are presented in Appendix A.7.

[6] A random classifier (or dumb classifier) is a method that tags each sample to a random class without any criteria, so the accuracy of this kind of classification normally is approximated to $1/N$, where N is the number of classes that the classifier knows. A classifier that gets results near to the random one, probably, cannot separate the classes at all.

Looking at the confusion matrix, it is possible to see that for all signals the classification results were low. Thus, it's difficult to affirm if the better results on the ECG are a causality or if there is a real tendency.

Table 4.4 – Valence confusion matrices.

	rPPG			**ECG**			**PPG**		
	Low	**Neutral**	**High**	**Low**	**Neutral**	**High**	**Low**	**Neutral**	**High**
Low	032	0.40	0.28	0.36	0.27	0.36	0.40	0.31	0.29
Neutral	0.35	0.45	0.20	0.23	0.50	0.27	0.33	0.34	0.33
High	0.14	0.25	0.60	0.46	0.31	0.23	0.36	0.37	0.27

At the valence confusion matrices, the ECG and PPG scores were almost completely similar to a random classifier. However, for rPPG, the results were better: the Low test samples were tagged mostly at the Neutral class (40%) than the Low (32%), followed by the smallest part of the High class (28%); for the Neutral class, 45% were classified as Neutral, 35% as Low and 20% as High; the High class got the best result, as 60% of the samples were rightly tagged in the High class, 25% as Neutral and only 14% as Low. These results were achieved using about 20 out of the 30 volunteers (10 were excluded with errors on the recording in one of the signals).

The highest results for the rPPG in Table 4.4 may have two possibilities: the first is that the best results achieved by rPPG are a coincidence. The second possibility is that the high low-pass filtering applied to the rPPG signal may enhance features that facilitate the identification of valence relating features on the cardiac signal. However, as the accuracy scores were low, more tests are necessary to confirm this hypothesis.

4.2.2 Multi-person Emotion Recognition

This section presents the results of the multi-person emotion classification. The classifier used was the same as the individual classification, the k-NN. However, to train and test the classifier, the *k-fold* technique was used. This method makes the classification k times using $(k-1)N/k$ subjects to train, and N/k to test the algorithm, where N is the number of subjects. It was defines a 5 as the k values. So, each fold had 80% of the volunteers for training (approximately 22), and 20% for testing (5 volunteers). The face detection failed to find the face of 3 out of 30 volunteers, so these ones are not used in the testes. The classification and testing process was repeated 5 times (one for each fold), and the mean result is calculated.

After extracting the features from the signals, putting the samples from all subjects together, there were at least 170 samples for video. So, as the number of samples is higher, besides the arousal and valence classification, it was also tested the result of the classification using each one of the six videos as an output. So, for rPPG, ECG and PPG, k-NN was trained in 5 folds using different volunteers for training and testing. To balance the train dataset, samples of the most populated classes are excluded until all classes had the same number of samples. The

classification metrics for each output (discrete emotions, arousal and valence) from rPPG signal are presented in Table 4.5. The classification results of ECG are presented in Table 4.6, and the results from PPG in Table 4.7.

Table 4.5 – Evaluation metrics for the multi-person classification using rPPG.

	Accuracy	**Precision**	**Kappa**
Video	0.28	0.37	0.13
Valence	0.41	0.44	0.12
Arousal	0.31	0.43	-0.01

Table 4.6 – Evaluation metrics for the multi-person classification using ECG.

	Accuracy	**Precision**	**Kappa**
Video	0.17	0.21	0.00
Valence	0.34	0.37	0.01
Arousal	0.35	0.43	0.02

Table 4.7 – Evaluation metrics for the multi-person classification using PPG.

	Accuracy	**Precision**	**Kappa**
Video	0.18	0.25	0.02
Valence	0.34	0.37	0.00
Arousal	0.35	0.41	0.01

The classification results were quite unexpected, as the prediction with ECG and PPG were completely compatible with a random classifier. For six classes (one per video), a random classifier would get an accuracy of 17%, which is the same accuracy score reached by ECG. The PPG was very close to it, with an accuracy rate of 18%. The arousal and valence parameters have three classes each, so the random accuracy value was 33%. The PPG and ECG got 34% of accuracy for the valence classification, and 35% for arousal, pretty close to a random classifier.

Surprisingly, the results of rPPG were better than the other two signals for the videos and valence classification. For the videos, the accuracy score reached 28%, almost 65% higher than the 17% of the random classifier. Even the Kappa got a score higher than 0.1. For valence, the accuracy reached 41%, 24% higher than the 33% of the random classifier. For arousal, the accuracy got a score of 31% a bit lower than the random classifier. For details, Tables 4.8, 4.9 and 4.10 present the confusion matrices of each classification. The values highlighted in green are the classes where the majority of the samples were tagged, and in red are the ones where the second biggest group is tagged.

The confusion matrix of video classification using rPPG enhances the hypothesis that different emotional states may be reflected at the signal. The majority of the greatest scores were in the right class, the only two that did not get the best score in the right class, getting the second higher. These results were obtained separating the subjects into two groups, the training and the

Table 4.8 – Normalized confusion matrix of the general rPPG video classification.

	Neutral	**Disgust**	**Fear**	**Happiness**	**Sadness**	**Surprise**
Neutral	0.24	0.20	0.08	0.26	0.07	0.15
Disgust	0.18	0.25	0.09	0.24	0.13	0.11
Fear	0.07	0.11	0.30	0.15	0.26	0.11
Happiness	0.22	0.22	0.10	0.27	0.10	0.10
Sadness	0.05	0.12	0.24	0.11	0.34	0.14
Surprise	0.10	0.11	0.17	0.17	0.24	0.21

Table 4.9 – Normalized confusion matrix of the general ECG video classification.

	Neutral	**Disgust**	**Fear**	**Happiness**	**Sadness**	**Surprise**
Neutral	0.15	0.13	0.14	0.22	0.23	0.12
Disgust	0.14	0.07	0.10	0.33	0.17	0.14
Fear	0.06	0.12	0.23	0.22	0.22	0.15
Happiness	0.13	0.19	0.19	0.23	0.10	0.15
Sadness	0.21	0.13	0.17	0.24	0.16	0.10
Surprise	0.16	0.12	0.14	0.23	0.17	0.17

Table 4.10 – Normalized confusion matrix of the general PPG video classification.

	Neutral	**Disgust**	**Fear**	**Happiness**	**Sadness**	**Surprise**
Neutral	0.13	0.15	0.16	0.28	0.17	0.11
Disgust	0.13	0.14	0.19	0.26	0.17	0.12
Fear	0.10	0.19	0.19	0.22	0.19	0.12
Happiness	0.11	0.13	0.20	0.29	0.17	0.11
Sadness	0.11	0.18	0.17	0.23	0.22	0.10
Surprise	0.12	0.13	0.19	0.23	0.17	0.15

test group. This separation avoid biasing the classifier since a person is used for train or for test the classifier, never both. Using this method, a reasonable amount of samples is generated per video, at least 170. In the classification results achieved using ECG and the PPG, the samples were mostly tagged at the happiness video. The respective confusion matrix for arousal and valence are presented at Tables 4.11 and 4.12.

Table 4.11 – Arousal confusion matrices of the three signals.

	rPPG			**ECG**			**PPG**		
	Low	**Neutral**	**High**	**Low**	**Neutral**	**High**	**Low**	**Neutral**	**High**
Low	0.38	0.35	027	0.35	0.38	0.25	0.41	0.35	0.24
Neutral	0.35	0.29	0.37	0.33	0.40	0.28	0.29	0.36	0.34
High	0.35	0.34	0.31	0.35	0.36	0.29	0.30	0.39	0.31

The arousal confusion matrices show that the separation of the classes for all signals was not successful. Almost all classes got between 30% to 40% of accuracy spread by all classes. Thus, it is not possible to affirm that a pattern was caught by the classifier.

Table 4.12 – Valence confusion matrices of the three signals.

	rPPG			**ECG**			**PPG**		
	Low	**Neutral**	**High**	**Low**	**Neutral**	**High**	**Low**	**Neutral**	**High**
Low	0.54	0.17	0.29	0.34	0.34	0.32	0.41	0.28	0.31
Neutral	0.34	0.32	0.34	0.31	0.35	0.34	0.35	0.27	0.37
High	0.33	0.34	0.33	0.28	0.38	0.34	0.34	0.34	0.31

For valence, results were almost the same, as the arousal ones, except by the Low rPPG class. About 54% of the rPPG Low valence samples were correctly classified, enforcing the result of the individual classifications that the valence rate of the videos may be caught through rPPG.

After the defense of this dissertation, a new evaluation of the algorithms used in the classification was done. The results are presented in Appendix B.

4.3 Heart Rate Variation for Emotion Classification in the Literature

According to Appelhans and Luecken (2006) HRV can reflect the individual emotional states responses, however, the literature also presents a limitation for it. Choi et al. (2017) tested the suitability of using the HRV extracted from ECG to separate the "happy", "unhappy" and "neutral" emotional states excited through visual stimulation by pictures. They concluded that is only possible to use an HRV-based emotion classifier when a high emotional level is expressed.

Guo et al. (2016) Used ECG to separate 2 (negative and positive) and 5 (sad, angry, fear, happy, relax) emotional states. They also used a set of videos to trigger these emotions on 25 subjects. And using segments of 90 s to extract the features they could recognise 71.4% of the emotions correctly for the 2 emotional states, and for the 5 different emotional states they could reach an accuracy of 56.9%.

Benezeth et al. (2018) was the only work found (until 31 May 2020) that used rPPG to separate the emotional states. According to them, they could differentiate a high and neutral arousal states using rPPG for 12 out of 16 volunteers (building a different system for each volunteer). They used 2 min length signals to extract the features.

The present work got worse results comparing to the literature. It shows that the emotion recognition step, may be improved, and in future works will be more productive to perform the analysis firstly with ECG, to achieve the results on the literature, then, repeat it for rPPG. Unfortunately due to the small time, it was not possible to neither refine the emotion classification techniques and either build an proper database for HRV emotion recognition (with recordings longer than 2 mins).

4.4 Computational Time of the Whole System

The last part is to integrate all the best methods tested at this work in a system capable to process a video, find the face of the user on it, extract the skin region through the frames, compute the mean RGB values of the skin region, estimate the cardiac signal, find the peaks on it, extract the cardiac parameters and then classify the signal to identify the user's emotion (all of this is required to work online).

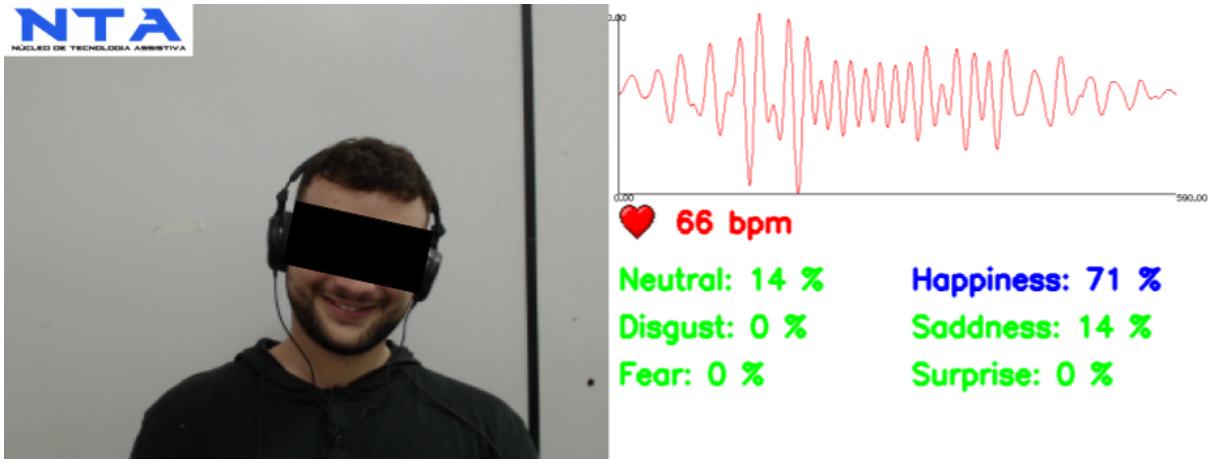
Putting all these parts together, the procedure spent 0.125 s[7] from the face-detection step to the classification step, so the system worked at 8 frames per second (fps), which was pretty slow. The Nyquist theorem sets: the minimum frequency sample of a signal should be the double of the highest frequency inside it. So, to capture a 4 Hz signal (240 bpm), the minimum sample rate is 8 Hz (or 8 fps). Which theoretically, it would be enough, however, to work at this sample rate, too much information would be lost (and there is higher frequencies on the signal besides the heart hate, increasing the aliasing effect). Also, not only the frequency of the signal is analyzed: to locate the rPPG peaks, the higher the sample rate, the higher the precision of the location of the peaks.

To improve the sampling rate, multiprocessing was used to split the whole processing. The main thread was used just to capture images and avoid losing frames. Another thread processed the face location and tracking, skin segmentation and extraction of mean RGB signal. The last thread was used for the rPPG estimation and filtering, peak finding, features extraction and classification. After using multiprocessing, the RGB signal generation needed 0.017 s to execute, equivalent to 59 fps, more than enough to process the frames of the webcam at 30 fps. However, the results of the classification were slower, as they took took 0.11 s to be processed, but is still fast enough to work online. Also, a friendly Graphic User Interface (GUI) showing the classification result and the heart rate was designed. The GUI is presented in Figure 4.16.

At the interface, the current image, a 30 s sample of the estimated cardiac signal, the estimated heart rate of the 30 s signal, and also the result of the classification are shown, with the most probable class highlighted in blue.

[7] The tests were computed with a i7 8th generation *Windows* computer, with 16 GB of RAM and no dedicated graphics board.

Figure 4.16 – Graphic User Interface presenting the results of the classification.



Source – Author's database, 2020

5 Conclusion

This work proposed an online emotion recognition system based on remote photoplethysmography (rPPG), by extracting the cardiac signal through RGB face-videos. Thus, the system is conceived to be used on a robot that helps a therapist in the interaction with autistic children. A research was done about the autism spectrum disorder, emotion recognition based on heart rate and the state of art on color image based rPPG.

To select the methods for the system, each step was evaluated for precision and processing time. At the face selection step, the Viola-Jones face detection method presented a good success rate and the smallest processing time. The chosen tracker, MedianFlow, spent less than 2 s to process 1800 frames and was able to track the great majority of face-area after a 60 s test. To extract the skin region, two methods were tested: the first based on specific regions of the face, and the second segments the skin based on its color. It was found that the color-extraction based method presented the best results, and also that the bigger the skin-area extracted, the better the results. After finding the skin-region, its mean RGB value was calculated and concatenated through frames to generate a temporal color signal. Four methods to estimate the cardiac signal from the raw RGB series were tested, being the plane-orthogonal-to-skin (POS) the one that got the rPPG signal most similar to the PPG. To get the heart rate using the frequency analysis, three methods to get the power density spectrum were tested, all of them with similar results, but being the Fast Fourier Transform (FFT) the fastest. Before identifying the peaks of the rPPG to calculate the HRV parameter, the rPPG was filtered. Thus, four filtering techniques were used and the one that best correlated the rPPG heart-beat differences to the ECG's was the one called Narrow-band, which works as a very narrow band-pass filter to overlapping parts of the signal, and then, average them.

With the interbeat signal extracted from the rPPG, twelve features were extracted, and used to classify videos with emotional content in two ways: a custom classifier for each subject and a multi-person classifier, both using k-Nearest Neighbor (k-NN) for classification. For comparison sake, the classification was also done with ECG and PPG signals. The database used was composed of 30 subjects, where each of them watched six short videos (about 40 s to 240 s each video) to trigger six specific emotions (neutral, disgust, fear, happiness, sadness, surprise), while their face-images, EGC and PPG signals were recorded. They also rated each video for arousal and valence. As the videos were short, there was not much data to make an individual classifier for each emotion. So at the individual classifier, only their rates for arousal and valence were classified. The arousal and valence rates were separated into three groups (Low, Neutral and High), and a set of a Low, Neutral and High arousal and valence videos were separated to train the k-NN, and the remaining videos used to test. To the multi-person classifier, besides arousal and valence classification, the six emotional states were also separated using k-NN.

In general, the classification scores were pretty low but quite interesting. The expected was that, using ECG, the classification results would get the best scores, then using PPG the scores would get a bit worse, and then, using rPPG, the scores wold be the worst. However, for the valence classification, the best score was obtained using rPPG, and the using the other two signals the accuracy results were close to the random classifier. For the multi-person classification, all the results got using ECG and PPG were close to the random classifier, but using rPPG, the classifier was able to identify the emotions better for the six videos classes and the valence rates outputs. The rPPG results were still pretty low, as the video classification got 28% of accuracy against 17% from the random classifier, and the valence got 41% against 33% from the random (the best class was the Low valence class, with an accuracy of 54%; the other two scored near to the random value of 33%).

This result brings some hypothesis, the first is that the high processing and filtering of the rPPG signal enhanced some important emotional features that were hidden at the ECG and PPG signals. So in future works, this procedure should be repeated, but applying a low-pass filter in ECG and PPG signals and evaluating the results. The second hypothesis is that the skin-color changes insert additional emotional information at the cardiac signal, which were extracted in the process and used in the classification. And the third one is that, as ECG and PPG, rPPG signal could not separate the emotions, and the better results for rPPG are a coincidence. It is worth to consider that the 30 s window may be too short to calculate the features, and two of the six videos got similar arousal and valence rates, even they representing different emotions, which difficulties the classification using the cardiac signal.

5.1 Future Works

Due to the rPPG estimation, it was noticed that the results are better for lighter skin-tones, probability because it is easier to catch the color variations on them. However, both linear rPPG estimators (POS and CHROM) used a standardized skin model to select the color-channel weight at the calculus, which may turn harder to get the color-variations on darker skin-tones. It is possible that a model with adaptive weights based on the skin-tone may improve the quality of the estimation.

To test the above hypotheses, another dataset should be recorded, with fewer emotions (trying to separate low/high arousal and valence only), and longer videos to excite the same emotion. The dataset should be planned and rated in partnership with psychologists, as the volunteer rates may be subjective.

At all, the proposed system can work online, but, using only cardiac information, it was not possible to get high accuracy on the emotion classification. A possible way to improve accuracy is to add the skin color values as additional information. For example, Goulart et al. (2019a) showed that different emotions change the temperature of the face in different ways. Thus,

if these temperature changes are caused by the increasing/decreasing of the blood volume in a particular area of the face, they may also change the color of this region, so it may be recorded by an RGB camera. This would be an interesting test for future works.

Another research suggestion is to use the HSV (Hue, Saturation, Value) color space to identify and track the skin region on the frames. A skin detection algorithm based on HSV could cope with the face-detector, face trackers and skin extraction parts of the work, which could save a considerable processing time.

5.2 Publications

Along the development of this research, the results achieved were published in two national and one international conferences:

1. LAMPIER, L.; BARROS, R.; RIVERA, H.; VETTORACI, P.; DELISLE-RODRIGUEZ, D.; CALDEIRA, E.; BASTOS-FILHO, T.. (2019). Medição de Batimentos Cardíacos via Câmera RGB: Comparando Diferentes ROIs e Janelas de Tempo. Vitória - Epírito Santo, Brazil. *2º International Workshop on Assistive Technology (IWAT)*, 2019.

2. LAMPIER, L; FLORIANO, A.; DELISLE-RODRIGUEZ, D.; BASTOS-FILHO, T.; CALDEIRA, E. Effect of Image Resolution on Remote Photoplethysmography: Towards Emotion Detection in Children with Autism Spectrum Disorder. Ouro Preto - MG, BRAZIL. *14º SIMPÓSIO BRASILEIRO DE AUTOMAÇÃO INTELIGENTE(SBAI)*, 2019 DOI: <http://dx.doi.org/10.17648/sbai-2019-111242>.

3. LAMPIER, L.; CALDEIRA, E.; DELISLE-RODRIGUEZ, D.; FLORIANO, A.; BASTOS-FILHO, T. Remote Estimation of Surrogate Heart Rate Indices from RGB Videos for Emotion Recognition in ASD Children-Robot Interaction. Buenos Aires, Argentina. *10º Congreso Iberoamericano de Tecnologías de Apoyo a la Discapacidad (IBERDISCAP)*, 2019.

Bibliography

AL-NAJI, A. et al. Monitoring of Cardiorespiratory Signal: Principles of Remote Measurements and Review of Methods. *IEEE Access*, v. 5, p. 15776–15790, 2017. ISSN 2169-3536.

APPELHANS, B. M.; LUECKEN, L. J. Heart rate variability as an index of regulated emotional responding. *Review of General Psychology*, v. 10, n. 3, p. 229–240, 2006. ISSN 10892680.

BALAKRISHNAMA, S.; GANAPATHIRAJU, A. *LINEAR DISCRIMINANT ANALYSIS - A BRIEF TUTORIAL*. Mississippi - EUA: Mississippi State University, 1998. 8 p. Available from Internet: <http://www.music.mcgill.ca/$\sim$ich/classes/mumt611_07/classifiers/lda_theory.pdf>.

BALAKRISHNAN, G. et al. Detecting Pulse from Head Motions in Video. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2013. p. 3430–3437. ISBN 1063-6919 VO -.

BENEZETH, Y. et al. Remote heart rate variability for emotional state monitoring. In: *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. [S.l.: s.n.], 2018. p. 153–156. ISBN VO -.

BOCCANFUSO, L. et al. A low-cost socially assistive robot and robot-assisted intervention for children with autism spectrum disorder: field trials and lessons learned. *Autonomous Robots*, v. 41, n. 3, p. 637–655, 2017. ISSN 1573-7527. Available from Internet: <https://doi.org/10.1007/s10514-016-9554-4>.

BOLME, D. S. et al. Visual object tracking using adaptive correlation filters. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2010. p. 2544–2550. ISSN 1063-6919.

BOUSEFSAF, F. et al. Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate. *Biomedical Signal Processing and Control*, v. 8, n. 6, p. 568 – 574, 2013. ISSN 1746-8094. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1746809413000840>.

BRENNAN, M. et al. Do existing measures of poincare plot geometry reflect nonlinear features of heart rate variability? *IEEE Transactions on Biomedical Engineering*, v. 48, n. 11, p. 1342–1347, Nov 2001. ISSN 1558-2531.

CARREIRAS, C. et al. *BioSPPy: Biosignal Processing in Python*. 2015. [Online; accessed 10th Feb. 2020]. Available from Internet: <https://github.com/PIA-Group/BioSPPy/>.

CHEN, W. et al. Estimating carotid pulse and breathing rate from near-infrared video of the neck. *Physiological Measurement*, {IOP} Publishing, v. 39, n. 10, p. 10NT01, 2018. Available from Internet: <https://doi.org/10.1088{\%}2F1361-6579{\%}2F>.

CHOI, K.-H. et al. Is heart rate variability (HRV) an adequate tool for evaluating human emotions? - A focus on the use of the International Affective Picture System (IAPS). *Psychiatry research*, v. 251, p. 192–196, may 2017. ISSN 1872-7123 (Electronic).

COHMER, S. *Autistic Disturbances of Affective Contact" (1943), by Leo Kanner*. Arizona State University. School of Life Sciences. Center for Biology and Society. Embryo Project Encyclopedia., 2014. Available from Internet: <http://embryo.asu.edu/handle/10776/7895>.

EKMAN, P. An argument for basic emotions. *Cognition and Emotion*, Routledge, v. 6, n. 3-4, p. 169–200, 1992. Available from Internet: <https://doi.org/10.1080/02699939208411068>.

GONZALEZ, R. C.; WOODS, R. E. *Digital image processing*. 3. ed. [S.l.]: Pearson Education do Brasil, 2010. ISBN 978-85-8143-586-2.

GOULART, C. et al. Visual and thermal image processing for facial specific landmark detection to infer emotions in a child-robot interaction. *Sensors*, v. 19, p. 2844, 06 2019.

GOULART, C. et al. Social robot for interaction with children. In: *XXVI Brazilian Congress on Biomedical Engineering*. [S.l.]: Springer Singapore, 2019. p. 711–715. ISBN 978-981-13-2118-4.

GOULART, C. et al. MARIA: Um Robô para Interação com Crianças com Autismo. *XII Simpósio Brasileiro de Automação Inteligente (SBAI)*, p. 557–562, 2015. Available from Internet: <http://swge.inf.br/SBAI2015/anais/164.pdf>.

GUDI, A. et al. Efficient Real-Time Camera Based Estimation of Heart Rate and Its Variability. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. [S.l.: s.n.], 2019. p. 1570–1579. ISBN 2473-9944 VO -.

GUO, H. et al. Heart Rate Variability Signal Features for Emotion Recognition by Using Principal Component Analysis and Support Vectors Machine. In: *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*. [S.l.: s.n.], 2016. p. 274–277. ISBN 2471-7819 VO -.

HAAN, G. de; JEANNE, V. Robust pulse rate from chrominance-based rPPG. *IEEE transactions on bio-medical engineering*, v. 60, n. 10, p. 2878–2886, oct 2013. ISSN 1558-2531 (Electronic).

HELD, D. et al. Learning to track at 100 fps with deep regression networks. In: LEIBE, B. et al. (Ed.). *Computer Vision – ECCV 2016*. Cham: Springer International Publishing, 2016. p. 749–765. ISBN 978-3-319-46448-0.

HYVARINEN, A.; OJA, E. Independent component analysis: algorithms and applications. *Neural networks : the official journal of the International Neural Network Society*, v. 13, n. 4-5, p. 411–430, 2000. ISSN 0893-6080 (Print).

IZARD, C. E. Basic Emotions, Natural Kinds, Emotion Schemas, and a New Paradigm. *Perspectives on Psychological Science*, v. 2, n. 3, p. 260–280, 2007. Available from Internet: <https://doi.org/10.1111/j.1745-6916.2007.00044.x>.

KAHNEMAN, D. *Thinking, fast and slow*. New York: Farrar, Straus and Giroux, 2011. ISBN 9780374275631 0374275637. Available from Internet: <https://www.amazon.de/Thinking-Fast-Slow-Daniel-Kahneman/dp/0374275637/ref=wl_it_dp_o_pdT1_nS_nC?ie=UTF8&colid=151193SNGKJT9&coliid=I3OCESLZCVDFL7>.

KALAL, Z. et al. Forward-Backward Error: Automatic Detection of Tracking Failures. In: *2010 20th International Conference on Pattern Recognition*. [S.l.: s.n.], 2010. p. 2756–2759. ISBN 1051-4651.

KALAL, Z. et al. Tracking-learning-detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE Computer Society, USA, v. 34, n. 7, p. 1409–1422, jul. 2012. ISSN 0162-8828. Available from Internet: <https://doi.org/10.1109/TPAMI.2011.239>.

KANNER, L. Autistic Disturbances of affective contact. *Nervous Child*, v. 2, p. 217–250, 1943.

KAZEMI, V.; SULLIVAN, J. One millisecond face alignment with an ensemble of regression trees. In: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*. USA: IEEE Computer Society, 2014. (CVPR '14), p. 1867–1874. ISBN 9781479951185. Available from Internet: <https://doi.org/10.1109/CVPR.2014.241>.

KING, D. E. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, v. 10, p. 1755–1758, 2009.

KLIN, A. Autismo e síndrome de Asperger: uma visão geral Autism and Asperger syndrome: an overview. *Rev Bras Psiquiatr*, v. 28, n. Supl I, p. 3–11, 2006. Available from Internet: <http://www.scielo.br/pdf/rbp/v28s1/a02v28s1.pdf>.

KO, B. C. A Brief Review of Facial Emotion Recognition Based on Visual Information. *Sensors*, v. 18, n. 2, 2018. ISSN 1424-8220. Available from Internet: <https://www.mdpi.com/1424-8220/18/2/401>.

KOELSTRA, S. et al. Deap: A database for emotion analysis ;using physiological signals. *IEEE Transactions on Affective Computing*, v. 3, n. 1, p. 18–31, 2012. ISSN 2371-9850.

KRANJEC, J. et al. Non-contact heart rate and heart rate variability measurements: A review. *Biomedical Signal Processing and Control*, v. 13, n. 1, p. 102–112, 2014. ISSN 17468108.

KWON, O. et al. Electrocardiogram Sampling Frequency Range Acceptable for Heart Rate Variability Analysis. *Healthcare informatics research*, Korean Society of Medical Informatics, v. 24, n. 3, p. 198–206, jul 2018. ISSN 2093-3681. Available from Internet: <https://www.ncbi.nlm.nih.gov/pubmed/30109153>.

LAI, M.-C. et al. Autism. *The Lancet*, Elsevier, v. 383, n. 9920, p. 896–910, mar 2014. ISSN 0140-6736. Available from Internet: <https://doi.org/10.1016/S0140-6736(13)61539-1>.

LANG, P. J. The emotion probe. Studies of motivation and attention. *The American psychologist*, v. 50, n. 5, p. 372–385, may 1995. ISSN 0003-066X (Print).

LEWANDOWSKA, M. et al. Measuring pulse rate with a webcam — A non-contact method for evaluating cardiac activity. In: *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*. [S.l.: s.n.], 2011. p. 405–410. ISBN null VO -.

LOMB, N. R. Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, v. 39, n. 2, p. 447–462, 1976. ISSN 1572-946X. Available from Internet: <https://doi.org/10.1007/BF00648343>.

LU, S. et al. Can Photoplethysmography Variability Serve as an Alternative Approach to Obtain Heart Rate Variability Information? *Journal of Clinical Monitoring and Computing*, v. 22, n. 1, p. 23–29, 2008. ISSN 1573-2614. Available from Internet: <https://doi.org/10.1007/s10877-007-9103-y>.

LUKEŽIČ, A. et al. Discriminative Correlation Filter Tracker with Channel and Spatial Reliability. *International Journal of Computer Vision*, v. 126, n. 7, p. 671–688, 2018. ISSN 1573-1405. Available from Internet: <https://doi.org/10.1007/s11263-017-1061-3>.

MAATEN, L. van der; HINTON, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, v. 9, n. 15324435, p. 2579–2605, 2008. Available from Internet: <http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf>.

MAHMOUD, T. M. A New Fast Skin Color Detection Technique. *World Academy of Science, Engineering and Technology*, v. 43, p. 501–505, 2008.

MEHRABIAN, A. Comparison of the PAD and PANAS as models for describing emotions and for differentiating anxiety from depression. *Journal of Psychopathology and Behavioral Assessment*, v. 19, n. 4, p. 331–357, 1997. ISSN 1573-3505. Available from Internet: <https://doi.org/10.1007/BF02229025>.

MORETTIN, P. A.; BUSSAB, W. d. O. *Estatística Básica*. 6. ed. [S.l.]: Editora Saraiva, 2010. 493 p. ISBN 9788502081772.

MORRIS, J. D. Observations: SAM: The self-assessment manikin: An efficient cross-cultural measurement of emotional response. *Journal of Advertising Research*, Advertising Research Foundation, US, v. 35, n. 6, p. 63–68, 1995. ISSN 1740-1909(Electronic),0021-8499(Print).

NUSSBAUMER, H. J. *The Fast Fourier Transform*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1981. 80–111 p. ISBN 978-3-662-00551-4.

OLIPHANT, T. E. *A guide to NumPy*. [S.l.]: Trelgol Publishing USA, 2006. v. 1.

OPENCV. *Face Detection using Haar Cascades*. 2018. Available from Internet: <https://docs.opencv.org/3.4.1/d7/d8b/tutorial_py_face_detection.html>.

PASTORE, C. et al. Diretrizes da sociedade brasileira de cardiologia sobre Ánalise e emissão de laudos eletrocardiográficos. *Arquivos Brasileiros de Cardiologia*, scielo, v. 93, p. 1 – 19, 00 2009. ISSN 0066-782X.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PENNISI, P. et al. Autism and social robotics: A systematic review. *Autism Research*, v. 9, n. 2, p. 165–183, 2016. ISSN 1939-3806.

PHUNG, S. L. et al. A novel skin color model in ycbcr color space and its application to human face detection. In: *Proceedings. International Conference on Image Processing*. [S.l.: s.n.], 2002. v. 1, p. I–I. ISSN 1522-4880.

REDFIELD, R. R. et al. Morbidity and Mortality Weekly Report Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years-Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014 Centers for Disease Control and Prevention MMWR Editorial and . v. 67, n. 6, 2014.

ROUAST, P. V. et al. Remote Heart Rate Measurement Using Low-Cost RGB Face Video: A Technical Literature Review. *Frontiers of Computer Science*, v. 12, n. 5, p. 858–872, 2018. ISSN 2095-2236.

RUTTER, M. Diagnosis and definition of childhood autism. *Journal of autism and childhood schizophrenia*, v. 8, n. 2, p. 139–161, 1978. ISSN 1573-3432. Available from Internet: <https://doi.org/10.1007/BF01537863>.

SADIKU, M. et al. Correlation: A brief introduction. *International Journal of Electrical Engineering Education*, v. 51, p. 33, 04 2014.

SANGHVI, J. et al. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In: . [S.l.: s.n.], 2011. p. 305–312.

SCARGLE, J. D. Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, v. 263, p. 835–853, dez. 1982.

SCHAAFF, K.; ADAM, M. T. P. Measuring emotional arousal for online applications: Evaluation of ultra-short term heart rate variability measures. In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. [S.l.: s.n.], 2013. p. 362–368. ISSN 2156-8103.

SHU, L. et al. A Review of Emotion Recognition Using Physiological Signals. *Sensors*, v. 18, n. 7, 2018. ISSN 1424-8220. Available from Internet: <https://www.mdpi.com/1424-8220/18/7/2074>.

STOCK, A. *Quais são as teorias e as pesquisas sobre as possíveis causas do autismo*. 2018. Available from Internet: <http://www.bbc.com/portuguese/geral-43577510>.

SURASAK, T. et al. Histogram of oriented gradients for human detection in video. *Proceedings of 2018 5th International Conference on Business and Industrial Research: Smart Technology for Next Generation of Information, Engineering, Business and Social Science, ICBIR 2018*, p. 172–176, 2018.

TAPUS, A. et al. Children with autism social engagement in interaction with nao, an imitative robot: A series of single case experiments. *Interaction Studies*, John Benjamins, v. 13, n. 3, p. 315–347, 2012. ISSN 1572-0373.

TASKFORCE. Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. *Circulation*, v. 93, n. 5, p. 1043–1065, mar 1996. ISSN 0009-7322 (Print).

THOMAZ, A. et al. Computational Human-Robot Interaction. *Foundations and Trends® in Robotics*, v. 4, n. 2-3, p. 105–223, 2016. ISSN 1935-8253. Available from Internet: <http://dx.doi.org/10.1561/2300000049>.

TIPPING, M. E.; BISHOP, C. Mixtures of probabilistic principal component analyzers. *Neural Computation*, v. 11, p. 443–482, January 1999. Available from Internet: <https://www.microsoft.com/en-us/research/publication/mixtures-of-probabilistic-principal-component-analyzers/>.

UNAKAFOV, A. M. Pulse rate estimation using imaging photoplethysmography: Generic framework and comparison of methods on a publicly available dataset. *Biomedical Physics and Engineering Express*, v. 4, n. 4, p. 1–17, 2018. ISSN 20571976.

VENTURE, G. et al. Recognizing Emotions Conveyed by Human Gait. *International Journal of Social Robotics*, v. 6, n. 4, p. 621–632, 2014. ISSN 1875-4805. Available from Internet: <https://doi.org/10.1007/s12369-014-0243-1>.

VERKRUYSSE, W. et al. Remote plethysmographic imaging using ambient light. *Optics express*, v. 16, n. 26, p. 21434–21445, dec 2008. ISSN 1094-4087. Available from Internet: <https://www.ncbi.nlm.nih.gov/pubmed/19104573https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2717852/>.

VIOLA, P.; JONES, M. Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, v. 1, p. I–511–I–518, 2001. ISSN 1063-6919. Available from Internet: <http://ieeexplore.ieee.org/document/990517/>.

VIRTANEN, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 2020.

WANG, W. et al. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, v. 64, n. 7, p. 1479–1491, July 2017. ISSN 1558-2531.

WANG, W. et al. Robust heart rate from fitness videos. *Physiological Measurement*, IOP Publishing, v. 38, n. 6, p. 1023–1044, 2017. ISSN 13616579.

WELCH, P. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, v. 15, n. 2, p. 70–73, June 1967. ISSN 1558-2582.

ZORCEC, T. et al. Getting engaged: Assisted play with a humanoid robot kaspar for children with severe autism. In: KALAJDZISKI, S.; ACKOVSKA, N. (Ed.). *ICT Innovations 2018. Engineering and Life Sciences*. Cham: Springer International Publishing, 2018. p. 198–207. ISBN 978-3-030-00825-3.

Appendix

APPENDIX A – General Equations and methods

A.1 Statistics

The statistical equations in this appendix are based on (MORETTIN; BUSSAB, 2010).

A.1.1 Mean

The mean value of the samples inside the **X** array is defined by Equation A.1:

$$\mu = \frac{1}{n}\sum_{i=1}^{n}(x_i) \tag{A.1}$$

where n is the number of samples of the **X** array.

A.1.2 Median

The median value of an array of samples is defined as the middle value of the sorted array if it has an odd number of samples, or the mean value of the two values at the very middle if it has an even number of samples.

A.1.3 Standard Deviation

The standard deviation of the **X** array is defined by Equation A.2:

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{A.2}$$

where n is the number of samples of the **X** array and μ is the mean value of the array.

A.2 Cross-Correlation

According to Sadiku et al. (2014) the cross-correlation measures the similarity between two signals. It involves sliding one function over another and calculate the overlapping area.

The cross-correlation equation, at a specific time instant t may be defined for continuous functions as:

$$R_{xy}(t) = \int_{-\infty}^{\infty} x(\tau)y(t+\tau)d\tau \tag{A.3}$$

For discrete series, the cross-correlation, for a lag n may be calculated as follows:

$$R_{xy}[n] = \sum_{-\infty}^{\infty} x[k]y(k+n) \tag{A.4}$$

A.3 Fourier Transform

A.3.1 Continuous Fourier Transform

According to Gonzalez and Woods (2010) the Fourier Transform represents a continuous time function $x(t)$ in the frequency domain as a sum of sines and cosines functions using the following transform:

$$\Im\{x(t))\} = X(f) = \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft}dt \tag{A.5}$$

Where $j = \sqrt{-1}$, f is the frequency value, t the instant of time and $X(f)$ the representation of the function at the frequency domain. To get the signal back to the time domain the inverse transform is defined as:

$$\Im^{-1}\{X(f)\} = x(t) = \int_{-\infty}^{\infty} X(f)e^{j2\pi ft}df \tag{A.6}$$

A.3.2 Discrete Fourier Transform (DFT)

When dealing with discrete samples a set of properties may be used to reduce the computational cost of the Fourier transform. The Discrete Fourier transform may be defined as (NUSSBAUMER, 1981):

$$\Im\{x[mT]\} = X[kf] = \sum_{m=0}^{N-1} x[mT]W^{mk} \tag{A.7}$$

Where $k \in 0, 1....., N-1$, $W = e^{-j2\pi/N}$, and N is the number of consecutive samples of $x[mT]$. The inverse transform is presented bellow:

$$\Im\{X[kf])\} = x[\bar{m}T] = \frac{1}{N}\sum_{k=0}^{N-1} X[kf]W^{-mk} \tag{A.8}$$

where $x[\bar{m}T] \equiv x[mT]$. The computational cost of the DFT is equal to N^2. Some algorithms exploits the properties of the DFT to reduce the number of operations to $\frac{N}{2}log_2N$ complex multiplications plus $Nlog_2N$ additions, this algorithms are called Fast Fourier Transform. For the details of the algorithm see (NUSSBAUMER, 1981).

A.4 Principal Component Analysis (PCA)

According to Tipping and Bishop (1999) The objective of the PCA algorithm is to find a set of orthogonal axes in which the variation of the data projection on those axes is maximal. So, for a set of a $N - dimensional$ the $q - dimensions$ with the higher variation may be selected.

For a set of N-dimensional data $\mathbf{t}_n$ with $n \in \{1, 2, ...N\}$, the covariance matrix of $\mathbf{t}_n$ is defined as $\mathbf{S} = \sum_n (\mathbf{t}_n - \bar{t})(\mathbf{t}_n - \bar{t})^T/N$. Such that $\mathbf{S}\mathbf{w}_j = \lambda_j \mathbf{w}_j$ where λ_j is the eigenvalue associated to the $\mathbf{w}_j$ eigenvector and $\bar{t}$ is the mean value of the samples. Defining the vector $\mathbf{x}_n$ as $\mathbf{x}_n = \mathbf{W}^T(\mathbf{t}_n - \bar{t})$, where $\mathbf{W}$ is composed of the q eigenvectos, $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_1, \mathbf{w}_1, \ldots, \mathbf{w}_q)$ and $q < N$. This makes $\mathbf{x}_n$ a q-dimensional representation of $\mathbf{t}_n$. Choosing the q principal eigenvectors associated to the highest eigenvalues of S, $\mathbf{x}_n$ will be represented in the axes in which the samples $\mathbf{t}_n$ have the greatest variance.

A.5 Independent Component Analysis (ICA)

The ICA algorithms is programed to solve the following situation: let's define $\mathbf{x}$ a set of signals of n linear mixtures of n independent components: $x_j = a_{j1}s_1 + a_{j2}s_2 + a_{j3}s_3 \cdots + a_{jn}s_n$, or $\mathbf{x} = \mathbf{A}\mathbf{s}$. This represents the j input linear mixtures from $\mathbf{x}$ as a weighted mixture of n independent sources, where each source, s, can not be directly observed.

ICA assumes that the components of $\mathbf{s}$ are non-Gaussian and statistically independents, meaning that s_i does not offer information of any other components. To find the independent components, first both $\mathbf{A}$ and $\mathbf{s}$ are estimated based on $\mathbf{x}$. Then the inverse of $\mathbf{A}$ is computed: $\mathbf{s} = \mathbf{W}\mathbf{x}$ (HYVARINEN; OJA, 2000).

There are different methodologies to infer $\mathbf{W}$, and $\mathbf{s}$, the code used by Lewandowska et al. (2011) is the *fastICA*, it was also used at the tests in this work. The code is already implemented by the library *scikit-learn* (PEDREGOSA et al., 2011).[1]

A.6 Error Meassurements

A.6.1 Root Mean Squared Error (RMSE)

The RMSE is defined by Equation A.9 (VIRTANEN et al., 2020):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{pred_i} - x_{truei})^2} \tag{A.9}$$

Where N is the number of samples, x_{pred_i} the predicted value for x_i and x_{truei} the true value of x_i.

[1] For more details see <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.FastICA.html>

A.7 Classification Metrics

The classification metrics are implemented at the library *scikit-learn* provided by (PEDREGOSA et al., 2011).

A.7.1 Accuracy

The accuracy score is the rate between the samples that are assigned in the right class divided by the total number of samples. The worse score is 0 and the best score is 1.

A.7.2 Precision

Precision measures the ability of the classifier to do not label samples outside a class as a member of it. The calculus is done as presented in Equation A.10.

$$Prec. = \frac{TP}{TP + FP} \tag{A.10}$$

where TP is the number of samples that are rightly assigned in the class and FP is the number of samples that are wrongly assigned in the class. The best value is 1 and the worst value is 0.

A.7.3 Kappa

The Kappa score represents the agreement level of two taggers about a series. A value close to one, means that they agree at the classification, closer to minus one, means that the disagree completely, and closer to zero means a random classification. the Kappa's formula is presented in Equation A.11

$$Kappa = \frac{p_o - p_e}{1 - Pe} \tag{A.11}$$

where p_o is the empirical probability of agreement on the assigned label and p_e is the probability when both taggers assign the tags randomly.

APPENDIX B – Algorithms Testing for Emotion Classification

This Appendix presents the tests to measure the influence of the algorithms used in two parts of the emotion classification process: the features dimension reduction and the classification algorithm itself. Dimension reduction is an important step to filter redundancy of the characteristics and, depending on the algorithm used, to better separate the clusters, which improves the accuracy of the classifiers. However, this process need some precautions. By reducing the dimensions of the samples, information is eliminated, if it is a redundant information or has no correlation with the classes that one want to separate, cutting these characteristics may improve the classification, however, if they are important information, the tendency is that the classification results are worse.

The algorithm used to reduce the dimensions in the original test was the Principal Component Analysis (PCA), which calculates a transformation matrix that maximizes the variance in the main components that have a higher probability to contain important information, and the components of less variance may be ignored. An alternative to PCA to reduce dimensions is the Linear Discriminant Analysis (LDA). LDA is a widely used technique for data classification and dimensionality reduction. It calculates a transformation matrix where the data is projected to a plane in which it the ratio of its between-class variance to the within-class variance is maximized (BALAKRISHNAMA; GANAPATHIRAJU, 1998). The algorithm used to perform the LDA was provided by (PEDREGOSA et al., 2011) [1].

As it is supervised, the LDA training is more expensive than PCA, however, after being trained, for the same number of dimensions, it is as fast as to transform the samples. A point to consider when using the LDA is that it generates a maximum of $N - 1$ dimensions, where N is the number of classes of the input data. That is, if the number of classes is low, it is possible that the algorithm eliminates important information for the classification, while in the PCA the number of dimensions selected can be higher (in this study the dimensions of greater variation which together accumulate 90 % the total variation of the data are maintained).

In this test it is also tested the use of all 12 characteristics extracted, without dimension transformation, applying to them only normalization, as presented in Equation B.1 (which is also applied before PCA and LDA).

$$\mathbf{X_{norm}} = \frac{\mathbf{X} - \mu(\mathbf{X})}{\sigma(\mathbf{X})} \tag{B.1}$$

[1] For more information see <https://scikit-learn.org/0.16/modules/generated/sklearn.lda.LDA.html>

where $\mathbf{X}$ is array with the feature's values, $\mu(.)$ is the mean operation and $\sigma(.)$ is the standard deviation operation.

Another tested point was the classifier algorithm. As the k-Nearest Neighbors (k-NN) algorithm has a high computational cost in the classification, mainly when there is a large number of samples and/or dimensions in the training database, since it calculates the distance between the test sample and all the others training bench samples. The linear Suport Vector Machine (SVM) is compared to the k-NN. The SVM creates a set o hyper-planes that better separates the classes to their respective classes (PEDREGOSA et al., 2011) [2]. Once trained, this algorithm is faster to classify samples than k-NN.

So, to compare the accuracy and spent time, of the dimension reduction algorithms and the classifiers, both individual and the multi-person classifications were repeated to compare PCA to LDA and the use none of them, and also the k-NN to the SVM in the classification.

B.1 Individual Emotion Classification

In the individual classification, the procedure was the same as previously done, the outputs of the classifications were the arousal and valence scores divided into three groups, neutral high and low. Where high represents the signal from the videos that were rated between 7 and 9 for valence/arousal, the neutral were between 4 and 6 and the low from 1 to 3. To perform the classification it was necessary that each volunteer had at least one video in each group, so if the volunteer had only given high/low grades to arousal/valence, he/she was discarded. This resulted in data from 14 volunteers for the individual classification of arousal, and 22 for valence, out 30 volunteers in the emotion dataset.

As in the tests from Section 4.2.1, the training bases were balanced, the signals recorded from each video were cut to have the same number of samples as the smaller video of the training base (the end of the signal in each one was maintained). Then, the training set of each volunteer was normalized using the norm of Equation B.1. The classification was performed using three preprocessing methods, just normalization, and normalization combined with PCA and LDA. And, for each one, a k-NN algorithm (with 5 neighbors) and a linear SVM algorithm were trained. The average results for the arousal classification of the volunteers are shown in Table B.1, and for valence in Table B.2, the metrics used were average accuracy, precision and the Kappa index.

B.1.1 Arousal

Looking at the results of Table B.1, the it is possible to notice that that the classification failed to separate the signs of rPPG and PPG. The results using these two signals in different configurations were close to 33 % (result of a random classifier for 3 classes). The rPPG had

[2] for more information see <https://scikit-learn.org/stable/modules/svm.html#svm-mathematical-formulation>.

Table B.1 – Individual Classification: Arousal Results

	k-NN			SVM		
	Normilized Features					
	Accuracy	Precision	Kappa	Accuracy	Precision	Kappa
rPPG	0.30	0.32	-0.02	0.41	0.42	0.13
ECG	0.41	0.42	0.14	0.36	0.35	0.04
PPG	0.31	0.30	0.01	0.28	0.30	-0.04
	PCA					
	Accuracy	Precision	Kappa	Accuracy	Precision	Kappa
rPPG	0.26	0.29	-0.07	0.29	0.30	-0.05
ECG	0.39	0.38	0.10	0.41	0.41	0.12
PPG	0.33	0.31	0.04	0.30	0.32	0.00
	LDA					
	Accuracy	Precision	Kappa	Accuracy	Precision	Kappa
rPPG	0.40	0.40	0.10	0.38	0.37	0.06
ECG	0.36	0.37	0.00	0.37	0.38	0.02
PPG	0.42	0.44	0.18	0.42	0.45	0.17

accuracy varies between 26 % to 41 %, and the PPG varying between 28 % to 42 %, these values indicate that the separation was highly influenced by randomness. The ECG was the only one that had all the values above (but close to) 33 %, the worst value was 36 % achieved by the configurations LDA/k-NN and Normalized Features (NF)/SVM. The best combinations were NF/k-NN and PCA/SVM, both reaching 42 % in accuracy. However, these results are not very trusty, since the number of data is low, only one training video on each class for each volunteer.

The number of volunteers used in the classification was low, only 14, and still only one video from each class used to train the classifier. Thus, these results represent more an indication than a definitive result. It would be interesting in future works, to retake the test with a greater amount of videos per volunteer. With these results, it is not possible to affirm a better successful configuration.

B.1.2 Valence

A Valency results reaffirm the suitability of rPPG to separate this emotional characteristic. While ECG and PPG showed accuracy close to 33 %, for all configurations, rPPG had accuracy greater than 47 %, with the best results reaching 51 % of using the Normalized features (that is, without the use of PCA / LDA in the preprocessing step) with both classifiers k-NN and SVM. The accuracy of SVM was 1 % higher than that of k-NN.

B.2 Multi-person Emotion Classification

In the Multi-person Emotion Classification, another modification was done in relation to that presented in Section 4.2.2. Instead of using the *k-fold* method with 5 folds, the *leave-one-out*

Table B.2 – Individual Classification: Valence Results

	k-NN			SVM		
	Normalized Features					
	Accuracy	Precision	Kappa	Accuracy	Precision	Kappa
rPPG	0.51	0.51	0.26	0.51	0.52	0.26
ECG	0.36	0.36	0.04	0.25	0.26	-0.11
PPG	0.38	0.38	0.08	0.36	0.36	0.04
	PCA					
	Accuracy	Precision	Kappa	Accuracy	Precision	Kappa
rPPG	0.47	0.46	0.19	0.49	0.48	0.23
ECG	0.37	0.36	0.05	0.30	0.30	-0.05
PPG	0.35	0.35	0.02	0.33	0.33	-0.01
	LDA					
	Accuracy	Precision	Kappa	Accuracy	Precision	Kappa
rPPG	0.49	0.49	0.23	0.49	0.50	0.23
ECG	0.29	0.29	-0.06	0.32	0.33	-0.02
PPG	0.33	0.33	0.00	0.38	0.38	0.08

methodology was used. In this method, one volunteer is used for testing and the others to train the classifier, the process is repeated until all the volunteers are tested. The average of the results of each volunteer tested were calculated, the results for the classification of arousal, valence (with the three classes: low neutral and high) and the six individual emotions are presented in Tables B.3, B.4 and B.5, respectively.

B.2.1 Arousal

Table B.3 – Multi-person Classification: Arousal Results

	k-NN			SVM		
	Normalized Features					
	Accuracy	Precision	Kappa	Accuracy	Precision	Kappa
rPPG	0.32	0.32	-0.01	0.30	0.32	-0.02
ECG	0.35	0.34	0.01	0.33	0.35	0.03
PPG	0.34	0.33	-0.01	0.25	0.28	-0.08
	PCA					
	Accuracy	Precision	Kappa	Accuracy	Precision	Kappa
rPPG	0.31	0.32	-0.03	0.29	0.30	-0.04
ECG	0.34	0.34	0.01	0.37	0.38	0.07
PPG	0.36	0.34	0.01	0.28	0.28	-0.07
	LDA					
	Accuracy	Precision	Kappa	Accuracy	Precision	Kappa
rPPG	0.28	0.29	-0.06	0.29	0.31	-0.03
ECG	0.35	0.35	0.03	0.34	0.36	0.03
PPG	0.33	0.32	-0.02	0.22	0.24	-0.13

For the multi-person classification of arousal, the results did not show any significant improvement or worsening, all of them remained close to 33 % (result of a random classification).

B.2.2 Valence

Table B.4 – Multi-person Classification: Valence Results

	k-NN			SVM		
	Normalized Features					
	Accuracy	Precision	Kappa	Accuracy	Precision	Kappa
rPPG	0.37	0.37	0.06	0.42	0.40	0.14
ECG	0.36	0.36	0.04	0.32	0.31	-0.01
PPG	0.30	0.31	-0.04	0.34	0.33	0.00
	PCA					
	Accuracy	Precision	Kappa	Accuracy	Precision	Kappa
rPPG	0.38	0.39	0.09	0.41	0.38	0.14
ECG	0.35	0.35	0.03	0.32	0.31	-0.01
PPG	0.33	0.33	-0.01	0.34	0.33	0.01
	LDA					
	Accuracy	Precision	Kappa	Accuracy	Precision	Kappa
rPPG	0.38	0.38	0.09	0.42	0.41	0.15
ECG	0.34	0.33	0.01	0.33	0.31	-0.01
PPG	0.35	0.35	0.02	0.35	0.34	0.03

The results of Valencia were similar to those of Section 4.2.2, where only the classification with the rPPG generates results significantly above 33 %. Regarding the preprocessing algorithms, there was no significant difference in the classification of the rPPG using Normalized features, PCA or the LDA, however the use of SVM has improved considerably in relation to k-NN.

B.2.3 Video

Table B.5 – Multi-person Classification: Video Results

	k-NN			SVM		
	Normalized Features					
	Accuracy	Precision	Kappa	Accuracy	Precision	Kappa
rPPG	0.25	0.25	0.11	0.28	0.27	0.18
ECG	0.18	0.18	0.01	0.20	0.19	0.04
PPG	0.18	0.18	0.02	0.20	0.20	0.04
	PCA					
	Accuracy	Precision	Kappa	Accuracy	Precision	Kappa
rPPG	0.26	0.27	0.13	0.26	0.23	0.17
ECG	0.18	0.18	0.02	0.21	0.20	0.04
PPG	0.18	0.17	0.01	0.19	0.18	0.02
	LDA					
	Accuracy	Precision	Kappa	Accuracy	Precision	Kappa
rPPG	0.26	0.26	0.13	0.26	0.26	0.17
ECG	0.16	0.18	0.01	0.20	0.20	0.04
PPG	0.20	0.20	0.04	0.20	0.20	0.04

For the classification of the six emotions generated by the videos, again the classification with the rPPG had the best accuracy, with results concentrated in 26 % (against 17% of a random classifier) and there was no significant difference between the use of PCA, LDA and none of them. The Kappa index, was better using SVM over k-NN.

B.3 Processing Time of the Tested Techniques

The processing time of the algorithms was also calculated. The classification of arousal was repeated using all volunteers for training and testing (only for measuring time spent), so the classification was made with 5156 samples.

Both the PCA and the LDA took 0.001 s to transform the data. Regarding the classifiers, k-NN took 0.170 s to classify the data while SVM was much faster, with 0.002 s (both using all 12 Features).

Annex

ANNEX A – Consent Form - Emotion Database

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Nome do voluntário:
Data de nascimento:
Endereço:
Telefone:

Eu,_____
portador do registro de identidade número _____, fui convidado a participar da pesquisa cujos termos são explicitados abaixo:

1. Título da Pesquisa: Identificação de Estados Emocionais por Meio de Imagens e Sinais Biométricos.

2. Pesquisadores responsáveis: Dr. Teodiano Freire Bastos Filho, Dra. Eliete Maria de Oliveira Caldeira, Dr. Denis Delisle Rodriguez e Eng. Lucas Côgo Lampier.

3. Justificativa: A medição remota de parâmetros fisiológicos é mais confortável aos usuários, e os dados fisiológicos medidos como variabilidade do batimento cardíaco possibilitam a inferência do estado emocional de uma pessoa. Em uma próxima fase um sistema desenvolvido para inferir emocoes será usado durante a interação de uma criança com autismo e um robô.

4. Objetivos da pesquisa: Estudar a factibilidade do uso de câmeras de vídeos colorida e térmica de baixo custo, para inferir emoções a partir da medição remota parâmetros cardíacos.

5. Procedimentos: A pesquisa contará com voluntários adultos saudáveis cujos critérios de exclusão são: ocorrência de fobias ou vivência de episódios traumáticos. O voluntário será convidado a assistir seis vídeos para evocar emoções positivas e negativas enquanto é filmado por uma câmera térmica, uma colorida, e tem seus parâmetros cardíacos monitorados por meio de eletrocardiografia e fotopletismografia. As imagens serão analisadas por meio de métodos que permitam identificar e quantizar as emoções do voluntário durante o experimento, e o mesmo responderá um questionário informando seu estado emocional após cada vídeo.

6. Duração e Local de pesquisa: Os procedimentos serão realizados em laboratório dentro da Universidade Federal do Espírito Santo (UFES), e terá duração aproximada de 30 a 40 minutos.

7. Riscos e desconforto: A pesquisa não envolve procedimentos invasivos. Podendo causar leve desconforto físico apenas pelo uso de sensores obtrusivos (eletrodos e o oxímetro), e pela evocação de emoções negativas, como susto, tristeza e nojo. O projeto foi aprovado pelo Comitê de Ética: DESENVOLVIMENTO DE DISPOSITIVOS DE TECNOLOGIAS ASSISTIVAS

E REABILITAÇÃO BASEADO EM SINAIS BIOLÓGICOS E REALIDADE VIRTUAL, número CAAE: 64797816.7.0000.5542

8. Benefícios: Desenvolvimento de um método remoto de captura de parâmetros cardíacos e detecção de emoções utilizando sinais biomédicos.

9. Garantia e recusa de participar da pesquisa: Entendo que não sou obrigado(a) a participar da pesquisa, além disso terei direito a desistir de participar da pesquisa a qualquer momento, sem que isto traga prejuízos a mim. Entendo que tenho direito a todas as informações pertinentes à pesquisa, mesmo que isto possa interferir na minha decisão de participar.

10. Garantia de manutenção de sigilo e privacidade: Autorizo a divulgação e publicação dos resultados a partir de fotografias ou vídeos dos procedimentos experimentais exclusivamente para fins acadêmicos e científicos.

11. Esclarecimento de dúvidas: em caso de dúvidas sobre a pesquisa devo contatar o pesquisador Lucas Côgo Lampier, nos telefones 4009-2661 ou no endereço Av. Fernando Ferrari, 514, UFES, Campus Goiabeiras, 29075-910 Vitória-ES. Também posso contatar o Comitê de Ética e Pesquisa do CSS/UFES para resolver dúvidas ou relatar algum problema através do telefone: (27) 3335-7211 ou correio: Universidade Federal do Espírito Santo, Comissão de Ética e Pesquisa com Seres Humanos, Av. Marechal Campos, 1468, Campus Maruípe, Prédio da Administração do CSS, 29040-090, Vitória-ES.

Declaro que li e entendi os termos acima expostos, como também os meus direitos. Concordo com as afirmações acima relacionadas e dou meu consentimento livre e esclarecido para participar da pesquisa.

Na condição de pesquisador responsável por esta pesquisa, Prof. Dr. Teodiano Freire Bastos Filho declara ter cumprido as exigências do item IV.3 da resolução 466/12 a qual estabelece diretrizes e normas regulamentadoras envolvendo seres humanos.

Vitória, ___ de _____ de _____ .

_____ Voluntário

_____ Pesquisador responsável