Exploring Physiological Multimodality for Emotional Assessment

Joana Rosa Figueiredo Pinto

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisor(s): Prof. Ana Luísa Nobre Fred
Dr. Rui Cruz Ferreira

Examination Committee

Chairperson: Prof. Patrícia Margarida Piedade Figueiredo
Supervisor: Prof. Ana Luísa Nobre Fred
Member of the Committee: Prof. Susana de Almeida Mendes Vinga Martins

May 2019

Preface

The work presented in this thesis was performed at the Institute for Bioengineering and Biosciences of Instituto Superior Técnico (Lisbon, Portugal), during the period February 2018 - May 2019, under the supervision of Professor Ana Fred.

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

The accomplishment of this master thesis would have not been possible without the support, guidance, encouragement and love of my dearest ones, to whom I thank.

My parents and sister, who have always supported me throughout my academic journey and in the pursuit of my dreams, even when those imply long absences.

My best girl friends, who brought me continuous encouragement throughout this 5 years of the master degree, and with whom I lived some of my happiest moments.

My hometown friends, my erasmus friends, and my friends living abroad, who have always been present, even when not in person.

My friends at Clynx, with whom I have been sharing some of the most challenging and exciting experiences in the entrepreneurial world.

All my dearest friends who volunteered to participate in the experimental sessions of this study. Their availability allowed me to construct the dataset that was essential for the purpose of this work.

My friends at the IT Laboratory, with whom I spent a long time during the development and writing of this thesis, and a particular appreciation to my Professors Ana Fred and Hugo Silva, who have always been supportive and available to give me key insights and guidance throughout this work.

Resumo

A resposta emocional conjuga o sentimento subjectivo e processos cognitivos, com manifestações motoras e fisiológicas.

No estado da arte, têm sido propostas várias abordagens para reconhecimento de emoção. Estas geralmente diferem no que concerne aos métodos de elicitação de emoção utilizados, aos estados emocionais a reconhecer, às fontes de informação ou modalidades, e às técnicas de classificação automática.

Este trabalho explora uma abordagem multimodal, baseada em biossinais, para reconhecimento emocional durante a visualização de videos imersivos, um método de elicitação relativamente pouco explorado. Os dados adquiridos através de sensores eletrocardiográficos (ECG), de atividade eletrodérmica (EDA), do pulso de volume sanguíneo (BVP) e de respiração foram recolhidos durante uma sequência de quatro calibrações e sete vídeos imersivos, elicitadores de diferentes emoções. Os participantes reportaram os seus estados emocionais do dia (baseline), e auto-avaliaram, através das escalas *Self-Assessment Manikin (SAM)*, as emoções experienciadas durante cada vídeo, no espaço bidimensional *valence-arousal*. Várias características fisiológicas e estatísticas foram extraídas dos biossinais e utilizadas como input para um sistema de reconhecimento emocional, avaliado em duas vertentes: classificação dependente do utilizador e classificação independente do utilizador, reconhecendo três e duas classes por dimensão, respetivamente. Foram utilizadas Máquinas de Vetores de Suporte (SVM), considerando que são um dos algoritmos mais promissores nesta área.

A abordagem proposta levou a taxas de reconhecimento de 51.07% para *arousal* e 67.68% para *valence* no cenário dependente do utilizador, e 69.13% para *arousal* e 67.75% para *valence* no cenário independente do utilizador, resultados que incentivam a pesquisa e o trabalho futuro com uma população e conjunto de dados maiores.

Palavras-chave: Reconhecimento de Emoção, Biossinais, Realidade Virtual, Máquinas de Suporte Vectorial, Computação Afectiva.

Abstract

Emotional responses combine subjective feeling and cognitive processes expressed by both motor and physiological manifestations.

Many emotion recognition schemes have been proposed in the state-of-the-art. They generally differ in terms of the emotion elicitation methods, target emotional states to recognize, data sources or modalities, and classification techniques.

In this work a multimodal approach based on biosignals is explored for emotion assessment during immersive video visualization, an elicitation method relatively unexplored within the related work. Data from Electrocardiography (ECG), Electrodermal Activity (EDA), Blood Volume Pulse (BVP) and Respiration sensors was collected, during a sequence of tasks comprising four calibrations and seven immersive videos, capable of eliciting different expected emotions. Participants reported their emotional state of the day (baseline), and provided self-assessment of the emotion experienced in each video through the Self-Assessment Manikin (SAM), in the valence-arousal space. Multiple physiological and statistical features extracted from the signals were used as inputs to an emotion recognition workflow, targeting both user-dependent and user-independent classification scenarios, with three and two classes per dimension, respectively. Support Vector Machines (SVM) were used, as it is considered one of the most promising classifiers in the field.

The proposed approach led to accuracies of 51.07% for arousal and 67.68% for valence in the user-dependent approach, and 69.13% for arousal and 67.75% for valence in the user-independent approach, which are encouraging for further research with a larger training dataset and population.

Keywords: Emotion Recognition, Biosignals, Virtual Reality, Support Vector Machines, Affective Computing.

Contents

List of Tables

List of Figures

List of Abbreviations

ANS Autonomic Nervous System.

BVP Blood Volume Pulse.

ECG Electrocardiogram.

EDA Electrodermal Activity.

EMG Electromyogram.

HF High Frequency Band.

HR Heart Rate.

HRV Heart Rate Variability.

K-NN K-Nearest Neighbors.

LF Low Frequency Band.

PNS Parasympathetic Nervous System.

SAM Self-Assessment Manikin.

SCR Skin Conductance Response.

SNS Sympathetic Nervous System.

SVM Support Vector Machines.

VR Virtual Reality.

July 8, 2019

Chapter 1

Introduction

1.1 Scope and Context

Emotions represent a valuable source of information in the daily interaction amongst Humans. Communication widely relies on the interpretation of affective states [1], as the expression of emotions of an individual can considerably change the sense of their message [2], by molding the perception that another individual has on that message. This connection between emotion and thinking plays a significant role in intelligent functioning and decision making throughout all sorts of human interaction and communication [1, 3].

The modelling of emotion presents two major challenges, associated to the vague definitions and fuzzy boundaries of emotion, and to the methodology followed [4]. Researchers in the field of psychology and neuroscience have explained emotion through various theoretical models, from which the two most applied ones are the discrete emotional model proposed by Ekman [5] and the two dimensional valence-arousal model proposed by Lang [6]. Moreover, emotions can be referred as a mental state or feeling that occurs spontaneously rather than a conscious effort [7], having two main types of manifestation: a mental response of emotion, combining subjective feeling and cognitive processes; and a bodily expression, which includes motor and physiological responses [4, 8], which present patterns that are reflective of emotional expressions [9].

The fact that emotion manifestation through electrophysiological signals, or biosignals, is determined by the Autonomic Nervous System (ANS), which can hardly be controlled by the subjective conscious or intention of the individual, enables more objective and reliable results [7, 10] comparing with external manifestations, such as facial expression, speech or gestures. Moreover, biosignals can be assessed with wearable and non-intrusive sensing techniques [10, 11], where miniaturization [10, 12] is happening to an extent that will soon allow the integration of biosignals sensors into everyday objects, accessories or clothes [13].

Being a complex process, emotions are expressed through a variety of responses or behaviors that are difficult for a machine to fully assess [14], reason for which automatic emotion recognition remains an open research problem [8], and a challenge of utmost importance. In effect, nowadays, both working, learning and entertainment are tightly connected to the usage of computers, despite they still ignore the emotional state of the user [8]. Thus, for instance, by recognizing the emotional state of an user, the computer (or robot, or machine) could adapt and determine its appropriate reaction [15], empowering natural and intelligent communications [1, 16].

It is believed that the progress towards an automatic emotion recognition system will bring significant value for both scientific research and commercial activities in a number of application areas. Researchers and developers are trying to apply breakthroughs in automatic emotion recognition systems to real-life applications, within several fields such as psychology, security, health care, road safety, education, marketing, advertising, gaming and service robots, or telecommunications [2, 15, 17, 18].

Specific applications of emotion recognition can target, for example, their usage in call centers, handling the interaction with customers based on their mood, or in investigation departments, where it could be used to predict the activities of the criminals [2]. Another promising field concerns educational applications, e.g. to boost the potential of virtual tutors, since the detection of attention, motivation and learning would enable the virtual tutor to minimize boredom, frustration, and react to confusion, helping the students to learn faster than in another scenario that only relies on performance measures [8]. In the context of gaming, the usage of emotion recognition has been proposed for brain computer interface (BCI), i.e. a device that gives an in-depth inside view of the emotion of the player, and thus might use this information to enhance their interaction with games in Virtual Reality (VR) Virtual Reality (VR) [19].

Many emotion recognition schemes have been proposed in the literature. They generally differ in terms of the emotion elicitation methods, target emotional states to recognize, data sources or modalities, and classification technique. In fact, it is key to design methodologies able to collect and label emotional experiences effectively [4]. This work will explore a multimodal approach for emotion recognition, based on a VR emotion elicitation protocol and on the usage of SVM towards that purpose. The emotion recognition system is tested to classify emotional states in terms of the emotion dimensions of valence and arousal, for both User-Dependent and User-Independent scenarios, considering 3 and 2 classes per dimension, respectively.

1.2 Objectives

The objectives of the present work are related to the major purpose of proposing an automatic system for emotion recognition in VR, namely:

- Designing an effective emotion elicitation protocol based on VR stimuli;
- Collecting emotional self-assessments and corresponding physiological response from five biosignals;
- Implementing a signal processing workflow for preprocessing, segmentation, and outliers detection and removal, for each biosignal;
- Extracting the most relevant features from each biosignal;
- Fusing the biosignals information into a multimodal classifier;
- Evaluating the classifier in two scenarios: User-Dependent and User-Independent.

1.3 Contributions

This work provided the following contributions to the emotion recognition field of research:

- An automatic integrated system for emotional elicitation and assessment based on biosignals, in a Virtual Reality based elicitation protocol.
- Database with experimental data acquired with the proposed system, i.e. a multimodal dataset ECG, BVP, EDA and Respiration[1]) associated to the respective emotional labelings (valence and arousal).
- Implementation of a module (in python) for automatic data quality assessment of EDA signals.
- Implementation of preprocessing methods for Respiration signals, proposed to integrate the public biosignals processing library, BioSPPy [20].
- Design of multimodal machine learning algorithms based on SVM.
- Experimental evaluation of the overall automatic processing and classification framework over the acquired data, in both User-Dependent and User-Independent scenarios
- Abstract accepted and presented at the 6th IEEE Portuguese Meeting on Bioengineering (ENBENG 2019), with overall evaluation of 4,5/5, regarding the significance and innovative character of the topic, its relation with the state-of-the-art, technical and experimental contribution, and presentation quality.
- Paper accepted as Full Contributed paper at the 41st IEEE Engineering in Medicine and Biology Conference (EMBC 2019).

[1] The Skin Temperature was also included in the experiment, but the data acquired was not accurate to include in the dataset

1.4 Thesis Outline

The present work is divided in seven chapters, organized in the following manner.

Now that the scope of the study has been introduced, along with the definition of the problem, objectives and contributions of the work, Chapter 2 comprises the relevant theoretical background of this study, namely the theory of emotion, the data sources associated with both internal and external pathways that manifest emotion, in particular the biosignals used in this work, and the SVM, which will be used for the classification tasks. Chapter 3 will present a review of the state-of-the-art regarding the emotion elicitation methodologies and the emotion assessment methods, both the subjective self-assessment and the computerized assessment, exploring the performance of different machine learning algorithms and features extracted by different authors. Chapter 4 outlines the approach proposed for this emotion recognition system, presenting the experimental settings and protocol, the treatment of the data regarding signal processing, segmentation, outliers handling and feature extraction, as well as the steps followed in the classification task. Chapter 5 includes the most relevant results obtained regarding the aims of the present work, presenting a comparative analysis of the self-assessment ratings obtained, a visual inspection of the biosignals and their attributes under different emotional states, and an analysis of the classification performance for the proposed emotion recognition pipeline. Chapter 6 outlines the discussion regarding the results obtained throughout the work and provids a critical analysis over the limitations of the study and future work guidelines. Finally, Chapter 7 remarks the main conclusions obtained throughout the previous discussion.

Chapter 2

Theoretical Framework

This chapter comprises the relevant theoretical background in the field of emotion characterization and recognition. This work primarily relies upon the understanding of emotion, for which an overview of the most common theoretical models to describe emotions is provided in Section 2.1.

Then, considering both internal and external pathways for the manifestation of emotions, Section 2.2 introduces the different modalities tipically used to assess them. A detailed presentation of the physiological manifestations takes place in Section 2.3, particularly for the biosignals selected for the present work.

Finally, the chapter provides an overview of the SVM classifier, in Section 2.4, which will be used due to its promising performance in the field of emotion recognition, as will be further reviewed from the literature in Chapter 3.

2.1 Emotion Theory

Emotions can be referred as a mental state or feeling that occurs spontaneously rather than a conscious effort, and they are reflected by a set of physiological changes in the human body [7]. Researchers agree upon the point that emotion eliciting situations are related to changes of multiple variables within subjective, physiological and behavioral responses [21].

Affect, on the other hand, is referred to as the mental counterpart of internal bodily representations associated with emotions, actions that involve some degree of motivation, intensity, and force, or even personality dispositions [22]. It allows reference to the effect of some event or of the internal state of an individual, although often without providing the ability of specifying exactly what kind of an effect or state it is [22]. In order to better understand emotion, Psychologists and Neuroscientists have explained it through various theoretical models. The two most applied models are the discrete emotional model proposed by Ekman [5] and the two dimensional valence-arousal model proposed by Lang [6].

The discrete emotional model divides emotions into several basic emotions and claims their universality among all cultures [5]. This model is built on the assumption that an independent neural system subserves every discrete basic emotion. Nevertheless, neuro-imaging and physiological studies have failed to establish reliable evidence to support this theory [23], and the matter remains ambiguous. Despite the diversity in categories of emotion, there is a considerable agreement in six emotions, those being happiness, sadness, surprise, anger, disgust, and fear [7].

The dimensional model is the most popular approach, which categorizes emotions based on scales and characterizes them according to their valence and arousal [7]. The model uses these two dimensions to quantitatively describe human emotions. This theory emerged from the evidence that even simple organisms such as worms possess basic approach/avoidance responses, positing that in more complex animals, discrete emotions such as anger and fear result from these basic emotive processes coupled with cognitive appraisals of the self and the environment [24]. Valence represents the pleasantness and ranges from negative to positive, while arousal indicates the activation level and ranges from low to high [6]. In this two-dimensional circumplex model, all emotions can be understood as varying degrees of both valence and arousal [23], plotted at various positions on a 2D plane, as illustrated in Figure 2.1.

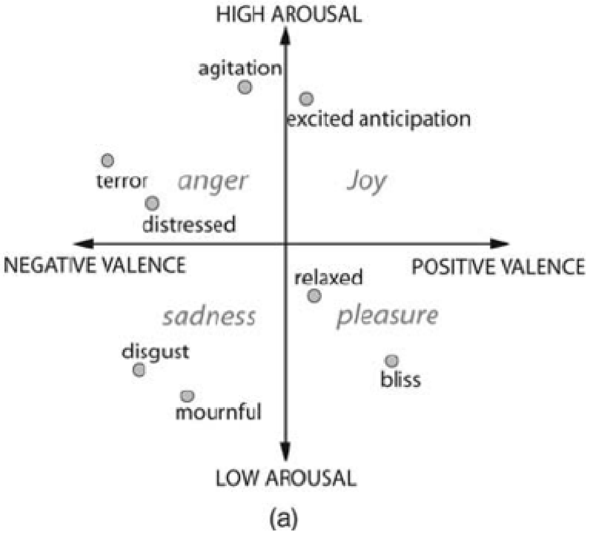


Figure 2.1: Visual summary of the two-dimensional valence/arousal model [25].

Harmon et al. [24] claim the value of both the dimensional and discrete models, considering that the value taken from describing emotions in terms of dimensions (e.g. valence and arousal) does not preclude the value of describing them through the discrete perspective. In fact, sadness differs qualitatively from anger, even though both have negative valence, as well as sadness differs qualitatively from relaxation, even though both have negative arousal.

Despite the two-dimensional model being the most commonly used in affective studies, some authors state that more than two dimensions are needed for a low-dimensional representation of the semantic space of emotion. Thus models with further dimensions have been proposed, e.g. a three-dimensional model is proposed by Schimmack and Grob [26] including valence, tension-arousal (tension-relaxation) and energy-arousal (awake-tiredness). A four-dimensional model has been proposed by Fontaine et al. [27], considering the dimensions evaluation-pleasantness (valence), potency-control (feelings of power vs weakness), activation-arousal (arousal), and unpredictability (differentiation appraisals of novelty vs familiar situations).

In terms of cross-cultural recognition of emotions, the meta-analysis conducted by Elfenbein et al. [28] shows that certain core components of emotions are universal and likely biological while, nevertheless, emotional expressions may lose some of their meaning across cultural boundaries. In effect, there is evidence of an in-group advantage, since emotions may be more accurately understood when they are judged by members of the same national, ethnic, or regional group [28]. As such, there is a suggestion that culture might have an important role in shaping the human emotional communication.

2.2 Data Sources

As stated, emotional manifestations can occur both on an internal or external basis. Namely, those can be reflected by the physiological activity, through the human biosignals, or by physical manifestations, such as facial expressions, voice or gestures. Those manifestations have been used as information modalities by researchers in the field of emotion recognition, as means of assessing emotion. Several studies have thus explored various modalities, including facial images and gestures [2, 14, 15], speech [2, 14], or physiological signals [29–34].

The group of methods that use facial images, gestures and speech can lack recognition accuracy, as they are not universal and depend on culture, gender and age [7]. Furthermore, when compared to physiological signals, those modalities are more susceptible to social masking, which can lead to wrong recognition of an emotional state. Regarding experimental settings, these modalities require special attention to lighting conditions and ambient noise for instance, which challenges their implementation in real world [7]. The major advantages of these methods concerns their interpretation, believed to be easier compared to other modalities [7].

Alternatively, physiological signals can be used for affect recognition, as they present patterns that are reflective of emotional expressions [9]. In fact, the focus has shifted towards the usage of biosignals, from both the peripheral and central nervous systems, since they can provide continuous measurements and appear to be more efficient and reliable in multiple aspects [14]. As those result from the activity of the ANS, they cannot be easily triggered by any conscious or intentional control [7], which allows the researchers to overcome the social masking problem described for the previous group of modalities. Moreover, when compared to visual data collection, e.g. facial expression, it is expected that the recording of biosignals is less disturbing for the users than being ”watched” by a camera, also avoiding the drawbacks of light levels issues and movements of the user [12]. Similarly, emotion recognition from speech has also been associated with critical difficulties in applications where users are listening to music or watching movies [12], as they are not expected to talk during these activities. Another major advantage of using biosignals concerns the miniaturization and integration capabilities of the underlying sensors [10–12], which is happening to an extent that will soon enable them to be incorporated into everyday objects, accessories or clothes [13].

However, biosignal processing poses several challenges related with their subjective and complex nature, the sensitivity to movement artifacts and corruption by power line interference, motion artifacts, or electrode contact noise [29], and the difficulty for the researcher to visually interpret or perceive emotions from the biosignals waves [30]. Those make the annotation and obtaining of the ground truth from the raw data more difficult. These artifacts can be encountered even in laboratory settings, and might become increasingly significant in real-life applications.

The difficulty of finding objective and measurable signals that contain affective information brings additional complexity to the process of emotion recognition [4]. Therefore, several methods such as the fusion of multimodal physiological signals are sought to enhance the efficacy of the process. The term ”multimodal information fusion” is described as the aggregation of information from several modalities to elicit decision-making strategies [16]. Comparing to emotion recognition based on a single biosignal, the fusion of multimodal emotion-related biosignals provides robustness [35], by eliminating anomalous changes not caused by emotional elicitation, which often appear in a specific biosignal. Furthermore, this fusion can boost the emotion recognition accuracy, since each individual modality can provide complementary information [18], and the recognition reliability is thus enhanced when taking into account the complementarity between classifiers [1]. Hence, multimodality has been increasingly and widely implemented for emotion recognition [18].

A representative example is the emotion study conducted by Setz et al. [29], where six physiological signals (ECG, Electromyogram (EMG), Electrooculogram (EOG), EDA, respiration and finger temperature) were recorded during the elicitation of five emotions (amusement, anger, contentment, neutral and sadness) through film clips. Their results suggested that modality fusion approaches have significantly increased the recognition accuracies.

2.3 Biosignals

In light of the prominent currents of thought concerning emotion theory, and taking into account the advantages of using physiological modalities over the alternatives described in the previous section, this work follows an approach based on multimodal physiological data to pursue the goal of emotion recognition.

Considering that emotion is mostly expressed by means of internal bodily manifestations, such as cardiorespiratory, intergumentary and thermal regulation activities, and that ECG, EDA, BVP, Respiration and Temperature are some of the most commonly modalities found in emotion assessment, those were selected for this study. Nevertheless, only the former four biosignals were actually included in the analysis, as the reliability of the temperature signals acquired during the experimental protocol was not satisfactory to proceed into the analysis, as will be detailed in Section 4.2.2.

This section will comprise a description of each of the adopted biosignals, along with the characterization of their physiological responses.

2.3.1 Electrocardiogram (ECG)

The cardiovascular system ensures the blood flow throughout the body. In addition to blood volume and contractile strength, the heart must sustain a regular cycle of relaxation and contraction [36]. This regularity is predicated on a series of complex electrophysiological events within the cardiac tissues, illustrated in Figure 2.2 and Table 2.1. The ECG is a test that records the cardiac electrophysiological [9, 37] and cardiovascular activity [38].

In an ECG signal, there are eight main electrophisiological events, from which three are visible in each cardiac cycle, these being the P wave, as evidence of the depolarization of the atrial muscle, the QRS complex, as evidence of the depolarization of the ventrical muscle, and the T wave, as evidence of the repolarization of the ventricular muscle [36, 39].

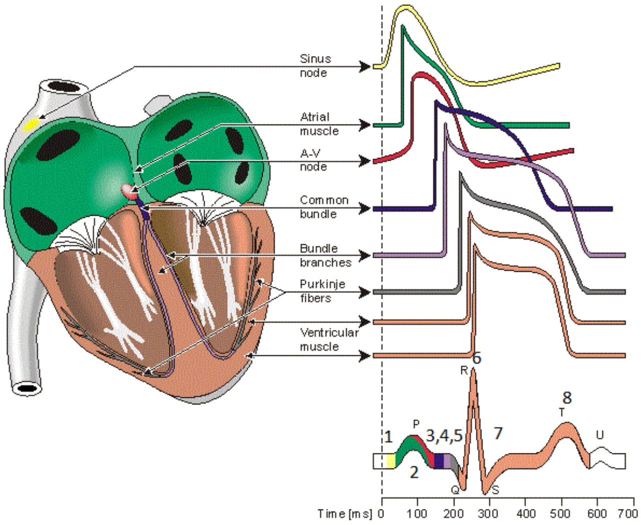


Figure 2.2: Summary of the 8 main electrophysiological events in a cardiac cycle, three of which are clearly evidenced in an ECG signal [39].

Typically, ECG measurements may be conducted with electrodes placed over the chest or limbs, even though the recordings from the limbs are more vulnerable to artifacts, yet less intrusive [12, 40]. From the ECG, it is possible to derive the rate and regularity of heartbeats [40]; furthermore, one can determine

Table 2.1: Summary of the 8 main electrophysiological events in a cardiac cycle, three of which are clearly evidenced in an ECG signal [36].

Physiologic Event		ECG Evidence
1.	SA node initiates impulse	Not visible
2.	Depolarization of atrial muscle	P wave
3.	Atrial contraction	Not visible
4.	Depolarization of AV node & Common Bundle	Not visible
5.	Repolarization of atrial muscle	Not visible
6.	Depolarization of ventricular muscle	QRS complex
7.	Contraction of ventricular muscle	Not visible
8.	Repolarization of ventricular muscle	T wave

the Heart Rate Variability (HRV), using the Heart Rate (HR) and/or Inter-beat Intervals (IBI) [37].

ECG responses to external stimuli (e.g. emotion elicitation), have been proved to produce large variability in a given subject's physiological signal [38]. Nonetheless, in the field of emotion recognition, HR is informative in the sense that low or increased HR can indicate a state of relaxation or frustration/mental stress, respectively [40]. The HR has thus been used in numerous studies that assess mental stress [14]. An increased heart rate has also been associated to emotions with high arousal ratings, such as joy [31].

2.3.2 Electrodermal Activity (EDA)

The EDA, also referred to as Skin Conductance (SC) or Galvanic Skin Response (GSR) [12], is one of the most promissor noninvasive peripheral measures of the ANS neural pathway [41].

It expresses the changes in electrical properties of the skin [41], namely the skin resistance, due to the activity of sweat glands, which is physically interpreted as conductance. The sweat glands, distributed on the skin surface, are controlled by the sympathetic nervous system only [42]. Thus, electrodermal signals enable the estimation of ANS, as they are a manifestation of the activity in eccrine sweat glands that are innervated by the sympathetic branch of the ANS, mainly by the sudomotor nerves [41].

The recordings of EDA signals can easily be performed on the body surface, by applying an electrical potential between two points of skin contact. This way, it is possible to observe a flow of current through them, due to the movement of free ions present in the skin structures [41]. Regarding the strength of the EDA responses in various body locations, the palmar and finger sites are the most suggested ones for the electrodes placement [41]. A practical issue concerns the fact that this biosignal, being influenced by external factors such as ambient temperature [12], requires cautions regarding the laboratory setup and its conditions.

Figure 2.3 illustrates a typical electrodermal response waveform, where onset and peak of the responses are marked, as well as other relevant features retrieved in terms of time and amplitude, such as the rising

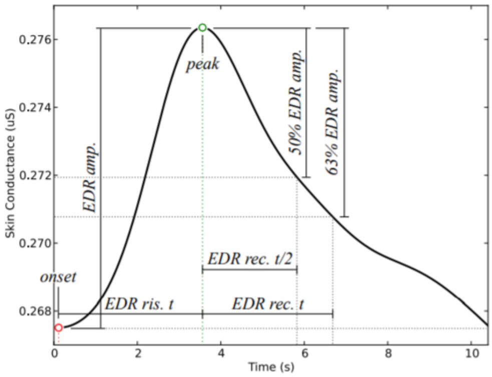


Figure 2.3: Skin Conductance Response (SCR) in the EDA signal [43].

time (EDR ris. t), the skin response amplitude (EDR amp.), and the recovery time (EDR rec. t), which occurs once the amplitude has decreased 63% with respect to the peak of the response.

The association between EDA and the ANS makes this biosignal widely popular in psychophysiology research, including information processing, and quantification of arousal levels during emotional and cognitive processes [41]. This way, the EDA is extensively used to infer perceptual affective states from physiological peripheral signals [44]. Particularly in the field of emotion recognition, this biosignal takes special relevance, since it is a good indicator of arousal level, due to external sensory and cognitive stimuli [14, 42], and stress, helping differentiate between conflict and no conflict situations [12], i.e.intense and relaxing emotions. For instance, emotions of joy, happiness and anger have been characterized by high skin conductance [31], as those are highly rated in terms of arousal. On the other hand, sadness has been associated with low skin conductance levels [31]. In terms of valence, positive emotions have been characterized by both low and high skin conductance [31], reinforcing the greater ability of the EDA to differentiate emotions along the axis of arousal rather than in the valence dimension [12, 41].

2.3.3 Blood Volume Pulse (BVP)

The BVP or Photoplethysmography (PPG) monitors the blood volume changes in the capillaries and arteries [38]. The BVP is measured with optical and non-invasive sensors [38, 45], generally consisting of a light source and photo sensor [42], that assess cardiovascular dynamics by detecting changes in the arterial translucency (or reflectance), and can be placed anywhere on the body [38], with the finger being the most common location [12, 42].

The BVP signal results from the fact that when the heart pumps blood, the arteries become more opaque (due to the momentary increase in volume), allowing less light to pass from the emitter on the sensor through to the receiver (or reflecting more light from the emitter on the sensor through to the receiver) [45], since the amount of light passed or reflected depends on the blood volume [42]. Being a periodic biosignal, the inter-beat time intervals of BVP can be used to, similarly to ECG, compute the heart rate (through their inverse) [4, 11]. Figure 2.4 illustrates one typical BVP signal, with three waves or periods,

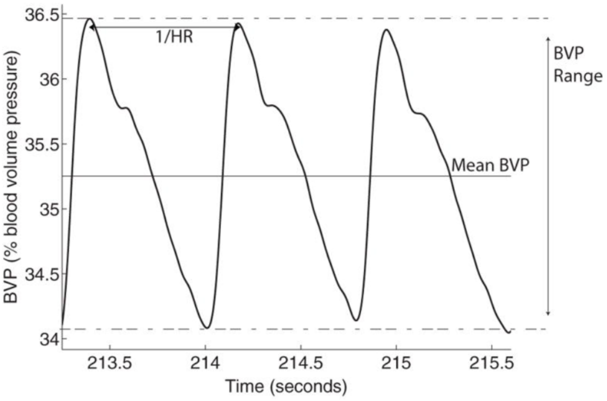


Figure 2.4: Typical waveform of a Blood Volume Pulse (BVP) signal [46].

marking some of its commonly extracted features.

The blood vessel diameter is controlled by the vasomotor activity, which is regulated by the sympathetic nervous system of the ANS. Hence changes in the BVP amplitude reflect instantaneous sympathetic activation [38], thus being also linked with emotion. In particular, the heart rate information that can be extracted from BVP might indicate states of emotional relaxation (low heart rate) or high arousal (high heart rate) such as mental stress, joy, and anger [14, 31, 40].

2.3.4 Respiration

Respiration sensors can measure how deep and fast a subject is breathing, and are generally applied on the chest or at the diaphragm level [12, 42].

The respiration biosignal can be useful in emotion assessment, as the respiration rate tends to reflect arousal [14]. It has also been observed that emotions of negative valence have a correlation with higher breathing rates [31]. For instance fast and deep breathing can indicate excitement such as anger, fear, or joy, while rapid shallow breathing can indicate tense anticipation, including panic, fear or concentration, and slow and deep breathing indicates a relaxed resting state while slow and shallow breathing can indicate states of withdrawal, passive-like depression or calm happiness [42]. Nevertheless, these characteristics are highly user- and dataset- dependent, having that joy has been associated to, contrarily to the reported by Jang et al. [42], deep and slow breathing [31].

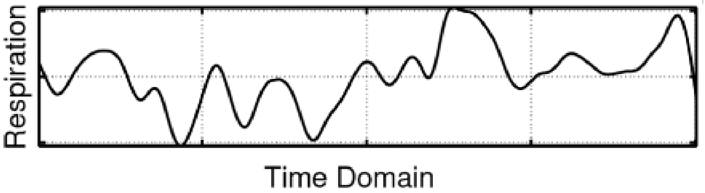


Figure 2.5: Typical waveform of a Respiration signal [21].

2.3.5 Skin Temperature

Skin temperature is determined by measuring the temperature on the surface of the skin. Similarly to the EDA, the skin temperature also depends on external factors and is reported to be related with emotion responses [12, 42]. For instance, under strain, the muscles become tensed and the blood vessels will be contracted, having a decreasing effect on the temperature [42].

In emotion assessment, one must take into account that it is, however, a relatively slow indicator of changes [12, 42], being indicated as able to carry valence cues [14].

2.4 Support Vector Machines

As will be further detailed in Chapter 3, SVM is considered one of the most promising algorithms in the field of emotion recognition, reason for which it was the method chosen for this classification task. This section will theoretically introduce the SVM, proceeding to present the algorithm formulation, and describing some concrete aspects of its implementation in the context of the work.

SVM belong to a set of supervised learning methods and it has seen extensive applications in learning tasks such as classification, regression and outliers detection [47, 48]. SVM were firstly proposed by V. N. Vapnik and A. Ya. Chervonenkis [49]. Their fundamental ideas, many of which are now being developed in the framework of SVM, were firstly published in 1964 [50].

SVM are based on statistical learning theory and intend to determine the location of decision boundaries that produce the optimal separation of classes [49, 51], through a so called ”separating hyperplane” [52]. In 1992, Boser et al. further introduced it as a training algorithm that maximizes the margin between the training patterns and the decision boundary [53], and proved its good efficiency and performance. Being based on the concept of separating different classes with a surface that maximizes the margin between them, this technique is said to be independent of the dimensionality of the feature space [51, 54]. The term ”support vectors” corresponds to the data points that are closest to the hyperplane [49, 51].

2.4.1 Formulation

2.4.1.1 Linear SVM

2.4.1.1.1 The Separable Case Given a set of training vectors $\mathbf{x_i} \in \rm I\!R^n$, i=1,...,*l* of dimension *n*, belonging to either of two classes, A and B. The input to the training algorithm is thus a set of *l* examples

x_i and an indicator vector y $\in \mathbb{R}^l$ such that $y_i \in 1, -1$, referred as the corresponding labels [53].

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_l, y_l) \tag{2.1}$$

Supposing that we have a hyperplane able to separate the positive from the negative samples [52], the points **x** which lie on the hyperplane satisfy:

$$w.x + b = 0 \tag{2.2}$$

where **w** is normal to the hyperplane, $||w||$ is the Euclidean norm of **w**, and $|b|/||w||$ is the perpendicular distance from the hyperplane to the origin. Defining $d_+(d_-)$ as the shortest distance from the separating hyperplane to the closest positive (negative) class [52], one has the *margin* of the separating hyperplane defined by $d_+ + d_-$

In this case, the SVM looks for the separating hyperplane with largest margin, satisfying the following constraints:

$$\mathbf{x_i}.\mathbf{w} + b \geq +1 \text{ for } y_i = +1 \tag{2.3}$$

$$\mathbf{x_i}.\mathbf{w} + b \leq -1 \text{ for } y_i = -1 \tag{2.4}$$

Equations 2.3 and 2.4 can be combined into the inequality:

$$y_i(\mathbf{x_i}.\mathbf{w} + b) - 1 \geq 0 \; \forall i \tag{2.5}$$

The support vectors (indicated by the extra circles in the Figure 2.6) are the points that define the hyperplane that maximizes the margin between the two classes, defined by Equations 2.3 and 2.4, which have a perpendicular distance of $|1 - b|/||w||$ and $|-1 - b|/||w||$ from the origin [52]. Thus $d_+ = d_- = 1/||w||$ and the margin is $2/||w||$. Hence the objective is to find that pair of hyperplanes that provides the maximum margin, minimizing $||w||$, or alternatively $\frac{1}{2}||w||^2$, subject to the constraints, for which the optimization problem to be solved can be formulated as

$$\min \frac{1}{2}||w||^2 \tag{2.6}$$

Aiming at easing the handling of the constraints in Equation 2.5, one can switch to a Lagrangian formulation [52], by replacing them with Lagrange multipliers themselves. Moreover, this reformulation will allow the training data to appear in the form of dot products between vectors, enabling the generalization of the procedure to the nonlinear case [52]. The positive Lagrange multipliers take the form of $\alpha_i, i = 1, ..., l,$, one for each of the constraints in Equation 2.5. The Lagrangian is formed by multiplying the constraint equations by positive Lagrange multipliers and subtracted from the objective function [52]. The Lagrange

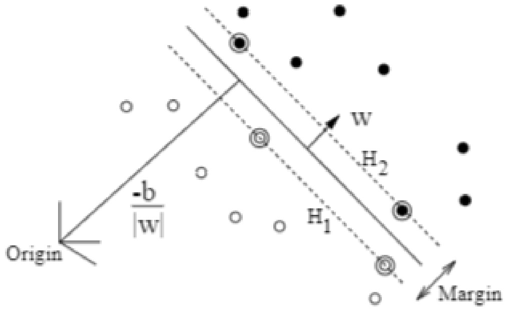


Figure 2.6: Linear separating hyperplanes for the separable case [52], where the decision boundary is the hyperplane lying half way between H1 and H2, having that H1 and H2 are the hyperplanes of the negative and positive class, respectively, and the respective support vector points are represented with an extra circle.

formulation thus becomes:

$$L_P \equiv \frac{1}{2}||w||^2 - \sum_{i=1}^{l} \alpha_i y_i(\mathbf{x_i}.\mathbf{w} + b) + \sum_{i=1}^{l} \alpha_i \tag{2.7}$$

Hence the objective is now to minimize L_P with respect to **w**, b, while requiring that the derivatives of L_P with respect to all $\alpha_i \geq 0, \forall i$ [52]. This corresponds to a convex quadratic programming problem, since the objective function is itself convex, and those points which satisfy the constraints also form a convex set, which means that one can solve the following “dual” problem instead. Hence, the goal becomes maximizing L_P, now imposing the gradient of L_P with respect to **w** and b vanish, also for $\alpha_i \geq 0, \forall i$. Considering these constraints, the following conditions are yielded:

$$\frac{\partial L_P}{\partial w} = 0 \longrightarrow \mathbf{w} = \sum_{i=1} \alpha_i y_i \mathbf{x_i} \tag{2.8}$$

and

$$\frac{\partial L_P}{\partial b} = 0 \longrightarrow \sum_{i=1} \alpha_i y_i = 0 \tag{2.9}$$

Substituting these equality constraints, which are in the dual form, into Equation 2.7, it yields the Dual Lagrangian formulation:

$$L_D = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x_i}.\mathbf{x_j}) \tag{2.10}$$

The two formulations were thus associated with two different Lagrangian labels (P for primal, D for dual), arising from the same objective function but with different constraints. The solution is found by minimizing L_P or by maximizing L_D. It matters to notice that, for this linear machines, the support vectors are the critical elements of the training set, which lie the closest to the decision boundary, having that if all other training points were removed or moved around without crossing the respective hyperplanes, the same separating hyperplane would be found in the training [52].

Considering the particular case of equality in Equation 2.5, which occurs for the support vectors (s ∈ S, subset of training points that are support vectors), yields:

$$y_s(\mathbf{x_s}.\mathbf{w}+b)-1\geq 0 \iff b=\frac{1}{y_s}-(w.x_s)\longrightarrow b=y_s-(w.x_s) \tag{2.11}$$

and thus:

$$b=y_s-\sum_{m\in S}\alpha_m y_m(x_m.x_s) \tag{2.12}$$

In the testing phase, the algorithm simply determines on which side of the decision boundary each testing point is, assigning the corresponding class label, as follows:

$$f(x)=sgn(x.w+b) \tag{2.13}$$

where f(x) is the class label of the point x, predicted by the SVM.

2.4.1.1.2 The Non-Separable Case The algorithm previously described cannot handle non-separable data. Hence, the previous ideas can be extended to the non-separable case by introducing a further cost (i.e. an increase in the primal objective function), the so called positive slack variables $\xi_i, i=1,...,l$ in the constraints [49], which represent the deviation from the ideal condition of linear separability [30], while still verifying Equations 2.3 and 2.4, by relaxing their hard margins [52]. These equations thus become:

$$\mathbf{x_i}.\mathbf{w}+b\geq +1-\xi_i \text{ for } y_i=+1 \tag{2.14}$$

$$\mathbf{x_i}.\mathbf{w}+b\leq -1+\xi_i \text{ for } y_i=-1 \tag{2.15}$$

$$\xi_i\geq 0\forall i \tag{2.16}$$

The objective function that was minimized in the previous case, $||w||^2$, now becomes

$$\frac{1}{2||w||^2+C(\sum_{i=1}\xi_i)^k}, \tag{2.17}$$

to assign an extra cost for errors, where C is a parameter chosen by the user, having that the larger it is, the higher penalty is assigned to errors. Similarly, the optimization problem must then also take the soft margin penalty term into account

$$\min\frac{1}{2}||w||^2+C(\sum_{i=1}\xi_i)^k \tag{2.18}$$

And the solution is given again by

$$w = \sum_{i=1}^{Ns} \alpha_i y_i \xi_i \tag{2.19}$$

where Ns is the number of support vectors, and the Lagrangian formulation is used (detailed in [52]).

2.4.1.2 Non-Linear SVM

The methods previously described can be generalized to the case where the decision function is not a linear function of the data, by applying the Kernel trick introduced by Cortes and Vapnik [53, 55]. One can map the data to some other (possibly infinite dimensional) Euclidean space $\mathcal{H}$, computing a mapping called ϕ:

$$\phi : \mathrm{I\!R}^{\mathrm{d}} \mapsto \mathcal{H} \tag{2.20}$$

Then, the training algorithm would only depend on the data through dot products in $\mathcal{H}$, thus on functions of the form $\phi(x_i).\phi(x_j)$. Defining a Kernel function, K, such that

$$K = \phi(x_i).\phi(x_j) \tag{2.21}$$

would allow to use it in the training algorithm without need to explicitly even know what ϕ is [52]. This trick allows the holding of all the considerations previously described, as it still computes a linear separation, only in a different space. At the testing phase, the SVM classification is obtained by computing, for each testing point, the sign of:

$$f(x) = \sum_{i=1}^{N_S} \alpha_i y_i \phi(s_i).\phi(x) + b = \sum_{i=1}^{N_S} \alpha_i y_i K(s_i, x) + b \tag{2.22}$$

where s_i are the support vectors and N_S is the number of support vectors, having that one was again able to use the Kernel, $K = \phi(s_i).\phi(x)$, instead of computing $\phi(x)$ explicitly [48, 49, 52, 55].

Various Kernel functions satisfy the condition on Equation 2.21 and are thus suitable for SVM tasks, as will be discussed in Section 4.5.3

2.4.2 Application

Although SVMs were initially developed to perform binary classification, most of the practical applications involve multiclass classification, as will happen with the user-dependent classification in this work. To perform multiclass classification with SVM, various methods have been proposed, namely the One against the Rest approach, the One against One approach, the Directed Acyclic Graph based approach, the Multiclass Objective Function, and the Error-Correcting Output Code based approach [51].

This work will follow the One against One Approach, since it appears to be the one that responds better to the tradeoff between accuracy and computation cost [51]. In this approach a set of SVM classifiers is created for all possible pairs of classes [51, 56]. Considering k as the number of classes, then $k(k-1)/2$ classifiers are constructed and each one trains data from two classes [47]. The output is obtained from each classifier in the form of a class label and the class label that occurs the most is assigned to that point in the data vector [51], *i.e.* using a sort of voting strategy, where each binary classification is considered to be a voting where votes can be cast for all data points, and each point is designated to be in the class with the maximum number of votes [57].

The dataset of this work corresponds to a Non-Linear SVM case, as the decision function is not a linear function of the data. Naturally, the Kernel trick is going to be applied to transform the input data into the Hilbert space, having two kernels considered during the model selection (Linear and RBF), as will be described in Section 4.5.3. Furthermore, taking into account the noise and ambiguous self-assessment of emotion, it was found convenient to consider this problem as a Non-Separable case, thus computing a non-rigid hyperplane and allowing some outliers in the training. Hence, the amount of misclassified data that is tolerated in each classification model will be determined through the adjustment of the parameter C, also during the step of model selection, aiming at accomplishing a better generalization ability.

Chapter 3

State of the Art

This chapter comprises an overview of the state of the art in the scope of emotion recognition. Section 3.1 presents the main techniques used for emotion elicitation, and 3.2 further details the usage and advantages VR towards that goal. Section 3.3 comprises a synthesis of both self-reporting and automatic techniques for emotion assessment, along with a review of the machine learning algorithms previously explored for emotion classification and the features most typically extracted from biosignals in related work. The chapter is concluded with Table 3.1 which depicts the main characteristics of the works conducted in the literature reviewed, allowing the comparison of these studies.

3.1 Emotion Elicitation

The study of emotion has been consistently associated with the ANS activity [1, 11], being the result of two interconnected components, the psychological and the physiological, as described in Section 2.2. Regarding the emotion elicitation process, most researchers agree upon the point that emotions usually occur as a response to internal or external stimuli or events that are significant to the organism [8]. As the emotion processes are correlated with the activity of ANS, whose manifestation is translated into the physiological signals of the subject, any experimental setup for the elicitation of emotions must be as natural and as close as possible to a real-life scenario [1].

As stated in the previous chapter, the elicitation of emotions can be performed as an external stimulus [34], for which several kinds of techniques have been tested, from visual [3, 12], audiovisual [29, 33], audio [23], personalized imagery [58], recall paradigm [7], to a multimodal approach [30] elicitation. Concrete methods include images, sounds, musics, video-clips and immersive videos, as will be hereafter summarized. A common goal of such procedures is to obtain the most validity, which relies on creating a strong sense of realism in the subject [12].

A set of normative emotional stimuli for experimental investigations of emotion and attention has been created and widely used during the last decades. Two important examples concern the visual and audio stimulus modalities, the International Affective Picture System (IAPS) and the International Affective Digitized Sound system (IADS), respectively. The purpose of IAPS is to provide a large set of internationally-accessible standardized color photographs that includes contents across a wide range of semantic categories, capable of evoking emotion [59, 60]. Similarly, the IADS provides a set of internationally-accessible standardized acoustic emotional stimuli, including contents across a wide range of semantic categories [61].

The usage of music has also been explored; despite the complex phenomena of music-mediated emotion elicitation is still lacking clear understanding, music is believed to have the ability to convey powerful emotions, for which it has seen a significant usage in emotion elicitation protocols [23].

Audio-visual film clips have widely been used [34, 62], since they are able to join the two techniques previously mentioned, combining both dynamic visual and auditory stimuli. Moreover, videos suit in elicitation protocols due to their properties such as being readily standardized, involving no deception, and being dynamic rather than static [34]. For instance, in the work by Li and Chen, videoclips were used to elicitate three target emotions (fear, neutral and joy) [63].

Other approaches such as the recall paradigm have been used [58], where the participant is asked to recall of past emotional life episodes, as emotional instances and dyadic interaction where a facilitator helps in inducing the various emotions [58].

Many studies have already confirmed the emotion elicitation ability of films, TV programs and imagery techniques [64]. Moreover, a new approach for emotion elicitation has emerged in the last two decades, through the usage of Virtual Reality [33, 64, 65], which can be understood as an extension of the audio-visual film clips, adding extra benefits in terms of immersion, which contribute to the purpose of higher reliability in the emotion elicitation process.

Considering its promising features and topical nature, this work explores emotion elicitation through VR content, hence further explored in the following section.

3.2 Virtual Reality

The term ”Virtual Reality” was coined by Jaron Lanier in the 80’s, during a period of intense research activity into this form of technology. At that time, VR was described as a computer simulated environment with, and within which, people interact [64]. Using visual, aural and haptic devices, the user is allowed to experience the environment as if they were part of the world, while the synthetic environment is able to modify its behaviour in real time, providing interaction and immersion [64].

More recently, a virtual environment has been defined as one that, through synthetic sensory information, leads to perceptions of environments and their contents as if they were not synthetic [66], which is closely related to the concept of immersion. VR is, by nature, “unreal” and relies on perceptual stimulation, from visual cues to sounds, and sometimes touch and smell, to trigger emotional reactions [65].

For the last two decades, VR has stepped into the field of emotion recognition in two ways. On one hand, VR can be used as a more reliable emotion elicitation agent and an unprecedented opportunity to investigate human behavior in well controlled designs [33, 64, 65], while on the other hand it can correspond to the end-application of emotion recognition in the context of Human Computer Interfaces (HCI), and specifically for Brain Computer Interfaces (BCI), where the assessment of emotions can enhance the interaction of a user with games in VR [19].

Moreover, VR can benefit emotion elicitation procedures in terms of immersion ability [64], as well as in several methodological aspects. As described by Blascovich et.al [66], those aspects are at least three, namely: (i) the experimental control–mundane realism trade-off; (ii) lack of replication, and (iii) unrepresentative sampling [66]. These are further detailed as:

(i) Elicitation settings must be chosen in order to facilitate the experimental control, implying precise manipulation of independent variables, while facilitating mundane realism, through situations similar to everyday’s life. This has key impact in the increasing of participants’ engagement within well-controlled experimental situations, although it is difficult to optimize [66].
(ii) Researchers often experience difficulties in implementing and using the exact methods and procedures of other investigators, and do not share physically identical laboratories, thereby eliminating perfect replications of scenarios [66], which can be mitigated through VR, to some extent.
(iii) There is a lack of random assignment and selection of participants, posing a major threat to external validity and generalization. For instance, most experimental social psychologists still use samples of convenience, typically college students, whom are not selected randomly, even from their own cohort [66].

Through the usage of virtual environments, one expects them to help ameliorate those methodological problems [66], particularly (i) and (ii). In fact, the advantage of leading to more robust replication and representative sampling as been pointed out as one of the key benefits by multiple researchers [33, 65].

Furthermore, in order to optimize the VR experience, according to Riva et al. [67] one must design it in a way that allows integration of three consciousness layers: *proto* presence, *core* presence, and *extended* presence. One should provide as much immersion as possible, by integrating *proto* (spatial) and *core* (sensory) presence; the events and entities experienced must be significant for the participant so as to integrate extended presence [67].

In fact, the reason why some of the elicitation methods described in the previous section are inefficient arises from their different ability to influence those three consciousness layers. For instance, in a compelling book reading, only extended consciousness is involved, and with a movie experience we can modify both core presence and extended presence but not proto presence [67]. The immersive VR is the only method that is able to influence all the three layers of consciousness, which gives suggestion that immersive VR is a privileged status as a medium for meaningful experiences [67].

It is thus believed that VR induces a sense of *presence* in a computer-generated world, referred as the "sense of being there" [68], and this ability to induce that feeling in the user experience is seen as the major feature of VR [64]. This sense of presence is closely connected with the immersive experience of VR, that is characterized as a psychological state in which the individuals perceives themselves to be enveloped by, included in, and interacting with an environment that provides a continuous stream of stimuli [66]. Moreover, there is consensus upon the point that the experience of presence is a complex, multidimensional perception, formed through an interplay of raw (multi-) sensory data and various cognitive processes [64]. Hence presence is seen as a necessary mediator that allows real emotions to be activated by a virtual environment, even though research has not yet been able to clarify the relationship between presence and emotional experience in VR or that causality role for presence [65].

Blascovich et al. [66] firstly proposed the usage of immersive VR technology as a new paradigm for experimental social psychology, as this is one of the fields that will most likely benefit from this promising research tool.

In the field of emotion recognition, Li et al. [33] conducted an emotion study compiling the several characteristics described for VR, and made available one database of standardized content to help establish and allow public access to a database of immersive VR video clips that can be a resource for studies on emotion induction using VR, with reliable valence and arousal ratings for each video.

3.3 Emotion Assessment

3.3.1 Self-Assessment

An important part of the emotion modelling process is the collection of the self-reported emotional states of the user. This self-assessment might be seen as a ground truth, providing the researchers with relevant information, since the emotion actually felt by an individual can strongly differ from the expected one (i.e. from the target emotion elicited) [69]. Considering this labeled data is thus of utmost importance for classification tasks and the critical analysis of the emotion recognition results.

The Self-Assessment Manikin (SAM), initially proposed by Lang [60], is an acknowledged technique and the most widely used scale for the measurement of the emotional states [3]. Several variations of these scales have been proposed, using different number of options to rate emotional states. Figure 3.1 illustrates two 9-point SAM scales [70], for valence and arousal, where the former is rated from negative

(1, sad) to positive (9, happy), and the latter is rated from low activation level (1, calm) to high activation level (9, excited).

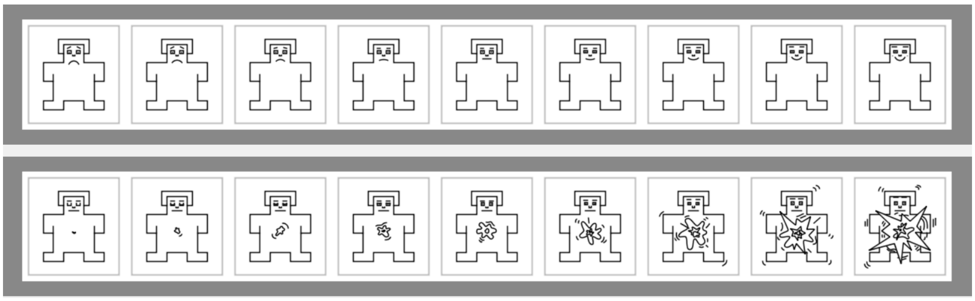


Figure 3.1: The 9-point SAM scales for valence and arousal [70].

3.3.2 Computerized Assessment

3.3.2.1 Machine Learning Algorithms

A wide range of machine learning methods has been used to infer emotional states [58], and both supervised and unsupervised learning have been explored towards that goal.

In supervised learning, one is provided with labeled data from two or more groups of objects [37], designated as training set. For instance, in emotion recognition the appropriate features might be available for a set of data from the emotion labeled as ”sad” and another set from the emotion ”happy”, or with different valence or arousal labels. The classifier is thus trained with those labeled examples to create a set of mathematical rules or models, which can then evaluate new data and predict its classification based on its features. Often, a testing set is used to test the classifier trained with the training set [37].

Alternatively, unsupervised learning is a similar process, despite more complex than classification [37]. The main difference refers to the fact the data provided to the classification method (the training set) is not labeled [37]. Clustering techniques (e.g.) retrieve not only the clusters (groups) of the data, but also a set of rules or mathematical equations to distinguish the groups from each other. Ciaccio et.al [37] claim that clustering is even more natural and more useful in medical research than supervised techniques, and K-means is one of the most popular techniques used.

The reason behind such a wide range for machine learning algorithms is that the best predictor will not be the same for all the datasets, as stated in the No-Free-Lunch theorem [71]. In fact, various studies [3, 72–74] have been conducted to test the performances achieved using classifiers from several families, in different contexts and datasets.

Moreover, there might be a big potential in combining different methods into the so called multiple classifier systems [54, 75]. For instance, Kittler et al. [75] noticed that often, when experimentally assessing different classifier designs for a problem, despite one of the designs would yield the best performance, the sets of patterns misclassified by the different classifiers would not necessarily overlap. This evidence suggested that different classifier designs could potentially provide complementary information about the patterns to be classified, which could improve the performance of the selected classifier [75].

Rigas et al. [3] computed an emotion classification system for three emotions (happiness, disgust, and fear) using four biosignals (facial EMG, ECG, respiration and EDA), comparing the accuracy of two classifiers, the Random Forests and K-Nearest Neighbors (K-NN). Their results showed statistically similar performance, concluding that K-NN performed slightly better, with an accuracy of 62,70%, largely inferior to what is reported in the other studies hereafter explored.

In their work, using physiological signals (EDA, ECG, BVP, and temperature), Jang et al. [73] conducted an analysis on the performance of four machine learning algorithms: LDA (linear discriminant analysis) as one linear model, CART (Classification And Regression Tree) as one decision tree model, SOM (self-organizing maps) as one artificial neural network, and SVM as one non-linear model, all well-known approaches used in emotion recognition. Aiming at identifying three single emotions (boredom, pain, and surprise), their results showed that SVM was the best algorithm being able to classify those emotions [73]. In another study, the same group of researchers [34] reinforced the previous result, by testing five additional machine learning algorithms, which were the FLD (Fisher linear discriminant) as one linear function, the CART, the SOM, the Naïve Bayes classifier based on density, and the SVM with the Gaussian radial basis function kernel. This time, seven emotional states (happiness, sadness, anger, fear, disgust, surprise, and stress) were under classification, yielding to an accuracy of 99,04% in the emotion classification by SVM, the highest accuracy amongst the algorithms tested, while FLD was the least accurate, with only 30,14%. The authors thus claim that these results should help new studies and lead to better chance of recognizing various human emotions by physiological signals [73], pointing SVM as the optimal emotion recognition algorithm for the data used in that study [34].

Changchun et al. [74] have reached to the same conclusion. Their empirical study comparing four machine learning techniques (K-NN, Regression Tree, Bayesian Network and SVM) using physiological features (of ECG and facial EMG) for emotion recognition found an advantage for SVM over the remaining algorithms [74]. SVM has reached a classification accuracy of 85,81%.

In their work, Haag et.al [12] used a Neural Network Classifier to classify user-dependent emotion states, split into arousal and valence, from several biosignals (ECG, EDA, respiration, BVP, EMG and Temperature). Their emotion recognition results showed that the estimation of the valence value was a harder task than the estimation of arousal, despite the overall results were considered good for both, accomplishing 89,7% of correct classification for arousal and 63,8% for valence, when allowing a bandwith of 10% for the output to be counted as correct, i.e. if the estimated and the target value lie within a range of a 10% distance. A similar result was found by Wagner et al. [31], whom also accomplished better results for arousal classification than for valence classification, which suggests that emotion is easier to differentiate along the arousal axis than the valence axis.

Most of the works in this field only performed user-dependent emotion recognition [1, 25, 31, 34, 73, 74, 76], as it is easier to be computed, even though a user-independent system is believed to be more significant and more applicable in the area of emotional intelligence [33]. The user-independent approach proposed by Li and Chen [33] yielded encouraging results, using four physiological signals (ECG, EDA, temperature and respiration) from multiple subjects and adopting a canonical correlation analysis to find the relationship between three emotions and extracted features. Their recognition accuracy is reported to be 85,3% [33]. Similarly, Kim et al. [30] developed a user-independent emotion recognition system using short term monitoring of physiological signals (ECG, EDA and Temperature), by classifying the extracted patterns with SVM, yielding to the correct classification ratios of 78,4% and 61,8% (for 50 subjects), for the recognition of three (sadness, anger, stress) and four (sadness, anger, stress, surprise) emotions, respectively. These results suggest the feasibility and importance of a user-independent emotion recognition system.

At the end of this chapter, Table 3.1 compiles a review of the related works that perform emotion recognition based on physiological signals assessment, considering their methodology in terms of elicitation method, biosignals, emotions to recognize, and overall accuracy obtained with the algorithm used. Hence, taking into account the evidence of classification accuracy of the several algorithms found in the state-of-the-art, the algorithm chosen for this work was the SVM. Moreover, due to the importance of further research towards user-independent systems, this work will explore both user-dependent and user-independent scenarios.

3.3.2.2 Feature Extraction

The choice of the right features is critical for the outcome of the classifier [37]. In fact, finding the features of the physiological signals that will correlate with certainty with the affective status of an individual is a challenging task, as one must take into account that they vary frequently from one individual to another, and that they are also very sensitive to day-to-day variations [58].

In emotion recognition, the process of feature extraction is understood as the transformation of the multimodal signals into a set of inputs or attributes, that are representative of the signal and suitable for a computational predictor [4], and is the first step in both supervised and unsupervised learning [37], previously described. This step takes place when the preprocessing is concluded, once the biosignals are reliable and thus appropriate to be subject to the feature extraction procedures.

Broadly, one can point out two types of features often used in biomedical sciences, described by Ciaccio [37] as (i) Biomedical and Biological features and (ii) Signal and Image Processing features, or simply (i) Physiological features and (ii) Statistical features, the terminology that will be used throughout this work. By definition, the former derive from the knowledge available about the biomedical system under study, while the latter are often purely mathematical concepts that do not necessarily present a direct physiological or biomedical meaning for the users [37]. These are potentially able to enhance the predictive capabilities of machine learning algorithms.

According to Martinez et al. [4], for unidimensional continuous signals such as the biosignals under analysis in this work, the features most typically extracted are common statistical features, such as average and standard deviation values, calculated on the time or frequency domains. The utilization of more complex feature extractors, such as deep learning methods, is increasing due to their capacity of automatically deriving characteristics from the signals. In fact, in the research by Martinez et al. [4], learned attributes led to more accurate affective models and gained simplicity in the sense that multiple signals were fused and fed directly (with limited preprocessing) to the model for training. Moreover, deep learning models are believed to retain more abstract characteristics and thus expressive representations of the input data at each layer [77]. This is an important property, since traditional machine learning systems require extensive domain knowledge and careful feature engineering, in order to find a suitable representation that can be fed to the next learning module [78]. An autoencoder has been proposed as a special type of feedforward neural network, whose purpose is to learn how to reconstruct the inputs belonging to a given dataset. This model is a self-supervised technique, trained to attempt to map its input to its output [77], and has been used to learn lower dimensional representations of the original data and to pre-train other deep learning networks [78].

Wavelet derived features are also a possible approach commonly used for biosignals classification [37], since the wavelet transform provides a number of coefficients that decompose a signal at different scales. In fact, the role of harmonic analysis has increasingly become substantial in the field of biomedical signal processing, benefiting from the theory of wavelets and their generalizations [79]. In a study for automatic classification of ECG atrial fibrillation, the feature extraction used the power spectral density of the wavelet decomposition of the signals [80], suggesting its suitability in ECG signals and thus a possibility for feature extraction approach in the context of emotion recognition.

Another hypothesis for the assessment of biosignals explored in the state-of-the-art [62] is that the momentary variations that occur in the physiological signals during different emotions can be evaluated by understanding the degree of similarity and short term correlations using the Hurst parameter. It assesses the smoothness of a time series, using self similarity and correlation properties, and can be obtained based on several methodologies supported by rescaled range statistics, finite variance scaling, wavelet transform and empirical mode decomposition. However, not much work has been done to find its emotional content, and the Hurst analysis has been mostly used for early pathologies detection [62].

As stated, statistical features are broadly common in the context of emotion recognition using biosignals data. Those features can be extracted from most of the recorded signals, whereas some special features are only extracted from a specific signal [32], as it will be discussed hereafter. Signal mean amplitude has been used for ECG [3, 12], BVP [4, 12, 42], EDA [3, 4, 12, 32, 42], respiration [3, 12, 32] and Temperature [12, 42], while standard deviation has been used for ECG [12, 32], BVP [4, 12], EDA

[4, 12], respiration and Temperature [12] and median for ECG, EDA and respiration [3]. Another time-domain feature extracted specifically for the ECG signal has been the root mean square of differences between RR intervals [32]. Means of absolute values of first differences has been also proposed by Rigas et al. [3] for ECG, EDA and respiration signals.

Unlikely those conventional statistics, Higher Order Statistics (HOS) refer to functions of orders three or more; HOS features have been widely used in the analysis of physiological signals, since their non-linear and non-Gaussian characteristics can be assessed [62]. HOS analysis has shown to be an accurate tool in the assessment of human emotional states [69], where the recognition of emotions was performed using EEG signals. HOS is believed to be useful to seek emotional information from other physiological signals [69]. Skewness and kurtosis are HOS features that measure the presence of transients in the signal and are robust to noise, which are normalized versions of third and fourth order functions, respectively. Skewness measures the symmetry of a distribution around its mean, while kurtosis measures the relative heaviness of the tail of a distribution with respect to its normal distribution [62]. In the work by Selvaraj et al. [62], ECG information was used to explore six emotional states through new non-linear features combining HOS and the Hurst parameter. Two new features, the skewness based Hurst and the kurtosis based Hurst, were derived and found to contain better emotional information compared to the Hurst parameter derived in the traditional way.

Regarding the type (i) of features, several physiological-based features have been suggested for each biosignal. The most used ECG features are the HR and HRV [3, 32, 34, 42]. In the frequency domain, the power in Low Frequency Band (LF) and in High Frequency Band (HF) are commonly extracted. The activities of the ANS (which consists of the Sympathetic Nervous System (SNS) and Parasympathetic Nervous System (PNS)), are believed to be reflected in the LF and HF. In fact, empirical evidence has suggested that the activity of the the SNS influences the LF of the HRV, from 0,04 to 0,15 Hz, while the PNS is predominantly reflected in the HF, from 0,15 to 0,4 Hz [81], also considered in the work by Jang et al. [42], while Murugappan et al. [82] considered slightly different band values, of 0,03 to 0,12 Hz and 0,12 to 0,488 Hz, respectively. From these basic features, further can be extracted, such as the ratio LF/HF and sum LF+HF [34].

Concerning the EDA signals, a wide range of physiological features has been explored in previous works, not being completely clear which of those can retrieve more valuable information. Some instances are the zero crossing rate [32, 83], the average of absolute derivative [32], the Skin Conductance Response (SCR) [11, 32, 34, 42] and non-specic skin conductance response (NSCR) [32, 34, 42], time and amplitude differences between the onset and peak [11], the power spectral, the rise time, the fall time [83], the initial skin conductance (SCi), the final skin condition (SCf) and their difference (SCi-SCf) [4, 11].

Physiological information of BVP signals includes the average inter-beat amplitude [4], whereas the HR and HRV can also be computed by their inverse [4, 11].

For respiration signals, the respiration rate is the most used physiological feature [11, 32].

Table 3.1: Review of related works on emotion recognition based on physiological signals assessment, considering their methodology in terms characteristics of the study and elicitation method, biosignals, emotions to recognize, type of recognition system, and overall accuracy obtained with the algorithm used.

Classification Algorithm	Elicitation method	Physiological Signals	Emotional assessment	Accuracy	System	Characteristics of the Study	Study
Random Forests	Set of Pictures from IAPS	Facial EMG ECG Respiration EDA	Happiness, disgust and fear	50,81%	User-Independent	9 users; 30-55 recordings per emotion	[3]
K-NN				62,70%			
LDA	Audio-visual film clips	EDA ECG BVP Temperature	Boredom, pain, and surprise	78.6%	User-Dependent	200 users (21,7± 2,3 years old)	[73]
CART				93,3%			
SOM				70,4%			
SVM				100,0%			
FLD	Audio-visual film clips (from movies, documentary and TV shows)	EDA ECG BVP Temperature	Happiness, sadness, anger, fear, disgust, surprise, and stress	30,14%	User-Dependent	6 males (20,8± 1,26 years old) and 6 females (21,2± 2,70 years old)	[34]
CART				74,93%			
SOM				33,67%			
Naïve Bayes				66,44%			
SVM				99,04%			
K-NN	Computer-based cognitive tasks	ECG Facial EMG	Anxiety, engagement, boredom, frustration and anger	75,16%	User-Dependent	15 users (21-57 years old)	[74]
Regression Tree				83.5%			
Bayesian Network				74,03%			
SVM				85,81%			
Neural Network	Pictures from IAPS	facial EMG EDA Temperature BVP ECG Respiration	Arousal (A) and Valence (V) (3 levels in each)	A: 89,73% V: 63,76%	User-Independent	1 user; 5 recordings per emotion	[12]
Canonical Correlation	Film clips	ECG Temperature EDA Respiration	Fear, Joy, Neutral	85,3%	User-Independent	60 female users (18-23 years old)	[63]
LDF	Music	facial EMG ECG EDA Respiration	Arousal (A) and Valence (V) (2 levels in each)	A: 96,59% V: 86,36%	User-Dependent	25 recordings per emotion	[31]
K-NN				A: 94,32% V: 86,36%			
Multilayer Perceptron				A: 94,32% V: 88,64%			
SVM	Multimodal approach (audio, visual and cognitive)	ECG BVP Temperature EDA	Sadness, Anger, Stress, (Surprise)	3 classes: 78,4% 4 classes: 61,8%	User-Independent	50 users (7-8 years old)	[30]
LDA	Music	facial EMG ECG EDA Respiration	Joy, Sadness, Anger, Pleasure	95%	User-Dependent	3 male users (25-38 years old); 90 recordings per emotion	[25]
Hidden Markov Model	Robot Actions	facial EMG ECG EDA	Arousal (A) and Valence (V) (3 levels in each)	A: 83% V: 80%	User-Dependent	36 subjects (19-56 years-old); 2-3 recordings (25 minutes each) per emotion	[84]
				A: 66% V: 66%	User-Independent		
SVM	Video-clips	ECG EDA Temperature Respiration	Arousal (A) and Valence (V) (2 levels in each)	A: 64,23% V: 65,03%	User-Dependent	24 users; 20 recordings	[76]

Chapter 4

Proposed System and Methodology

The approach followed for emotion recognition is similar to the method proposed in [3] and based on the following general steps: emotion elicitation; biosignals acquisition; biosignal preprocessing; feature extraction; model selection; and classification. Throughout this chapter, each section will describe the practical implementation of each of the steps that compose the whole system.

Figure 4.1 illustrates the pipeline proposed, from the point where the biosignals input the system, until the prediction of the emotional state, performed by the multimodal SVM-based classifier .

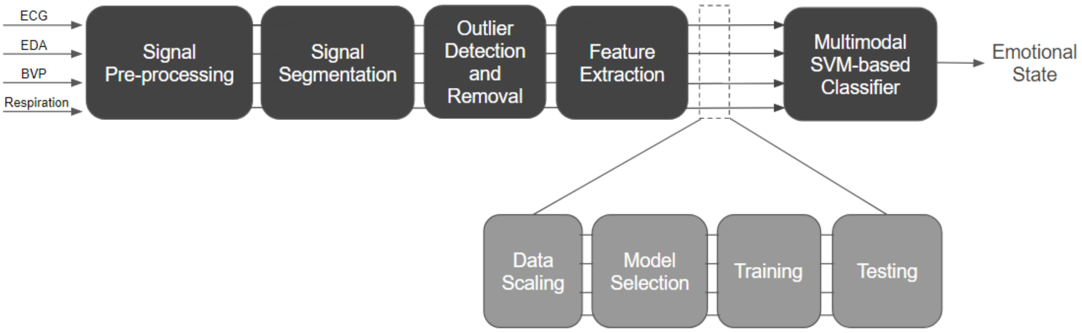


Figure 4.1: Overall structure of the emotion recognition system proposed.

4.1 Experimental Settings

4.1.1 Acquisition System and Experimental Setup

Three main components compose the acquisition system: the sensor modules for multimodal biosignal acquisition; the computer application that communicates with these modules; and the VR headset. The setup used during this experimental procedure is depicted in Figure 4.2.

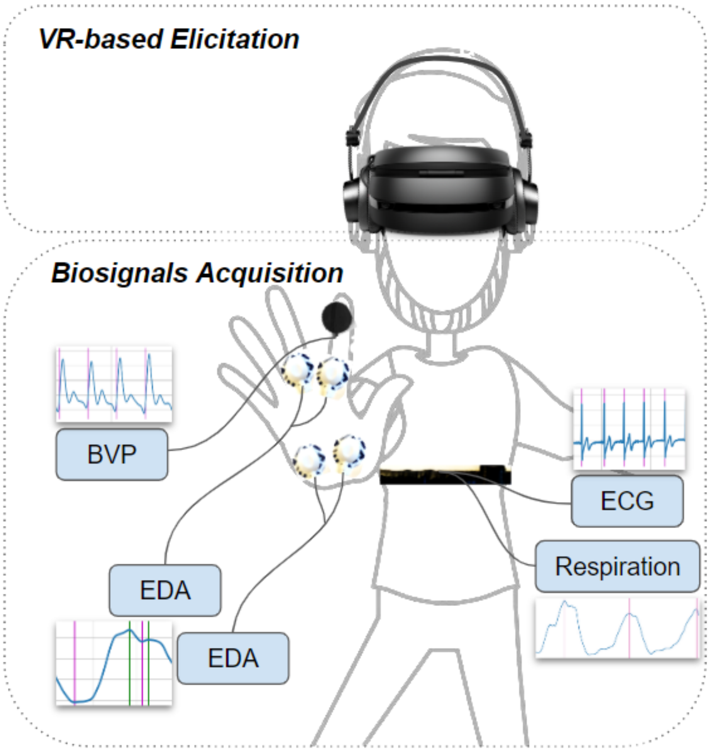


Figure 4.2: Illustration of the setup used during this experimental procedure.

The physiological data was collected using two modules built upon BITalino devices [85, 86], one placed on the participant's arm, and the other on their chest. During the experiment of each participant, 11 time series of data were collected per biosignal, each of them with different length and emotion goals, as detailed in the next section. Figure 4.3 summarizes the data that is acquired for each participant.

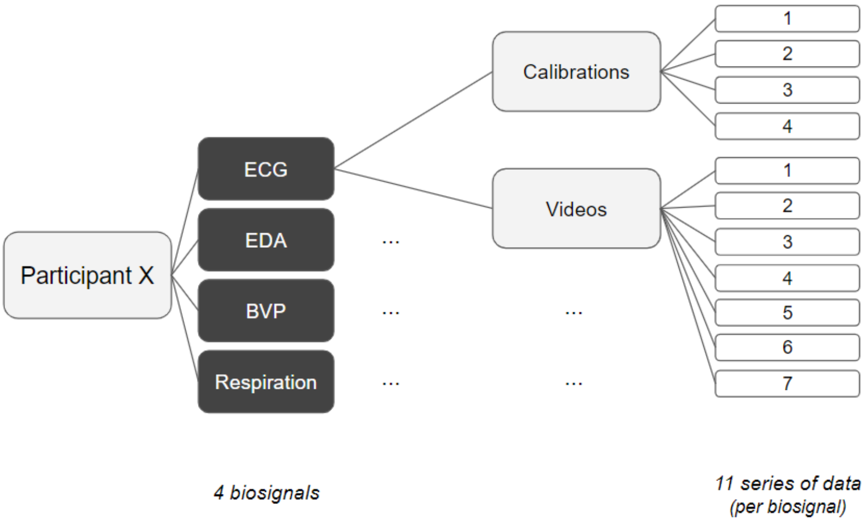


Figure 4.3: Summary of the protocol followed and data acquired for each participant.

The module placed on the arm encloses two EDAs, one BVP, and one skin temperature sensors. One EDA and the BVP sensors were placed at the base of a finger, while the other EDA's sensor and the temperature sensor were placed directly on the palm of the hand. Data was recorded from the right hand for all the participants, as illustrated in Figure 4.2.

Regarding the chest module, it incorporates ECG and respiration sensors. This data was acquired at the chest level, through their integration in an adjustable band.

To complete the experimental setup, the participants used the HP Windows Mixed Reality Headset [87] and headphones, so as to fully immerse on the VR videos of the elicitation protocol. Only the visual stimulus was included in the experimental setup, i.e. the participants only used the headset while the motion controllers were left aside, in order to avoid extra sources of artifacts from the upper-limbs movements that would result from their usage.

4.1.2 Emotion Elicitation Protocol

As referred in the Section 3.2, this protocol used immersive VR video clips for the emotion elicitation, combined with the SAM scale for its self-assessment by the participants.

Prior to the experiment, the participants had an adaptation time to feel comfortable in the laboratory setting while they were being introduced to the protocol procedures. At the beginning of the experience, the participants were asked to self annotate their emotional state of that day. This information was considered as the day's emotion baseline, to be taken into account in the analysis.

The wearable acquisition system was then attached on their chest, wrist, and finger, as described in Section 4.1.1. Participants were told to enjoy the VR experience and explore the video environments by turning their heads. Nevertheless, in order to maximize the reliability of the acquisition, it was asked to the participants to avoid hand or arm movements that could impact the chest's module.

Before the visualization of the emotion elicitation videos, there was a calibration period, in which participants were asked to think of, remember, or imagine situations in which they felt each of four emotions. Those four calibrations were relative to sadness, anger, happiness, and relaxation, chosen due to their relevant difference in the two dimensional valence/arousal plane, representing the extremes of the emotion states, as illustrated in the Figure 4.4.

Once the calibration was completed, a total of seven videos was shown, all with distinct target emotions, as shown in the Table 4.1. Each video had a target emotion and its choice was based on the emotion ratings for arousal and valence listed in [88], and on the association to the ratings available in several videos available at a public database of immersive VR videos [33].

Each video visualization was preceded by 5 seconds of a black screen, the goal of which was to immerse in a neutral emotional state before a new emotion was elicitated. In fact, the elicitation method should induce the target emotion while eliminating the chances of inducing multiple emotions [1].

Table 4.1: Protocol summary, including the system preparation and sequence of calibrations and videos, with respective target emotions and reference to the their ID in the VR video database [33]. In the column that indicates the duration of each step, the periods in which there was acquisition o biosignals are in parenthesis.

	Target Emotion	Duration (s)	ID in the Database [33]	Description of the Video
Reading and consenting with the purpose of the study.	-	60	-	-
Annotation of the emotional state of the day.	-	30	-	-
Wearing the acquisition system and VR setup.	-	120	-	-
Adaptation time and final recommendations.	-	60	-	-
Calibration 1	Sad	(30)		
Calibration 2	Anger	(30)		
Calibration 3	Happiness	(30)		
Calibration 4	Relaxation	(30)		
Neutral (black) screen.	Neutral	5	-	-
Video 1	Boredom	(43)	1	Upclose view of the inside of a bee hive with a large number of bees [33]
SAM annotation.	-	30	-	-
Neutral (black) screen.	Neutral	5	-	-
Video 2	Joyfulness	(250)	70	The viewer experiences snorkeling and surfing on a Tahitian beach [33]
SAM annotation.	-	30	-	-
Neutral (black) screen.	Neutral	5	-	-
Video 3	Panic/Fear	(160)	-	The viewer finds themself in an intense scenario and soundtrack, inside a submarine that is grabbed by a craken.
SAM annotation.	-	30	-	-
Neutral (black) screen.	Neutral	5	-	-
Video 4	Interest	(65)	42	Video clip showing four kittens playing with one another [33]
SAM annotation.	-	30	-	-
Neutral (black) screen.	Neutral	5	-	-
Video 5	Anger	(75)	-	Street fight that is experienced in first person by the user.
SAM annotation.	-	30	-	-
Neutral (black) screen.	Neutral	5	-	-
Video 6	Sadness	(120)	6	Virtual environment of a desolate valley [33]
SAM annotation.	-	30	-	-
Neutral (black) screen.	Neutral	5	-	-
Video 7	Relaxation	(210)	32	Sun rising over the horizon at a beach [33]
SAM annotation.	-	30	-	-

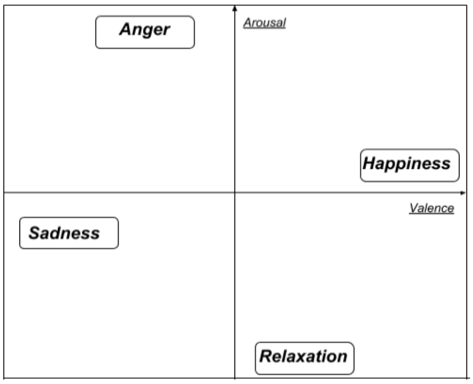
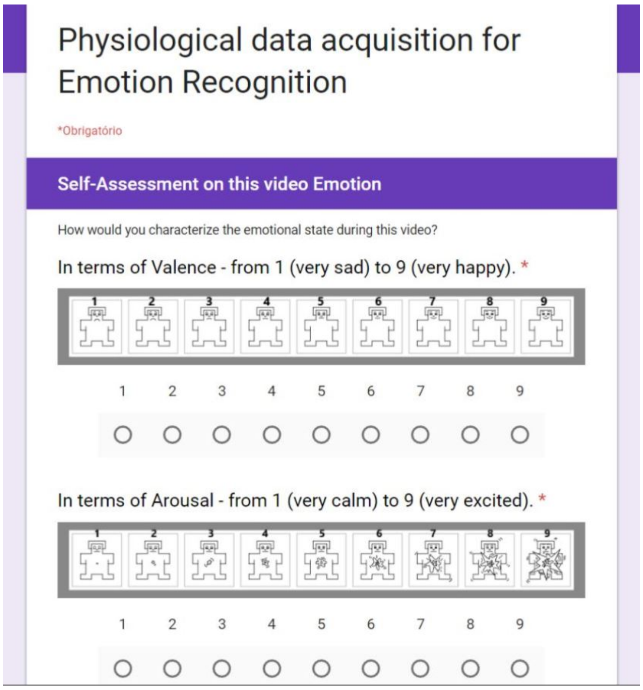


Figure 4.4: Representation of the target emotions of the calibration step in the two dimensional plane.

After being exposed to each emotion elicitation video, the participants were asked to self annotate their emotional state. Figure 4.5 illustrates the scales the participants were asked to fill out after each video visualization. Similarly to the procedure of [33], the self-assessment was initiated by the SAM Valence scale and then the SAM Arousal scale.



Physiological data acquisition for Emotion Recognition

*Obrigatório

Self-Assessment on this video Emotion

How would you characterize the emotional state during this video?

In terms of Valence - from 1 (very sad) to 9 (very happy). *

1	2	3	4	5	6	7	8	9
○	○	○	○	○	○	○	○	○

In terms of Arousal - from 1 (very calm) to 9 (very excited). *

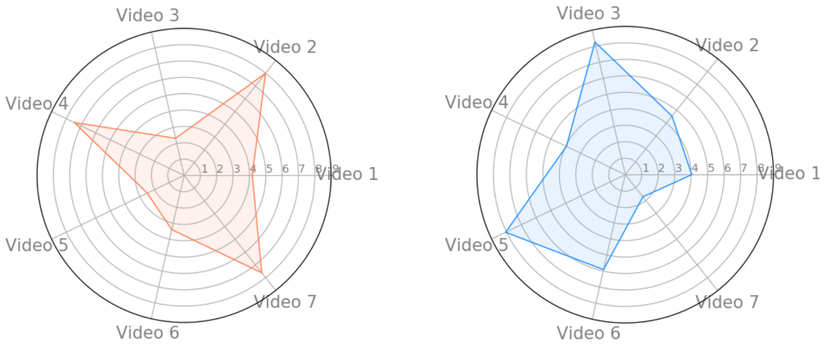
1	2	3	4	5	6	7	8	9
○	○	○	○	○	○	○	○	○

Figure 4.5: Self-Assessment Manikin 9-point scales presented after each video visualization, during the experiment.

The elicitation videos were selected based on two objectives, which were to obtain videos for a representative range of emotions, while presenting good quality in VR. An extension to those initial objectives took place when considering the extensive VR research by [33] and its respective VR videos database that, by providing each video's correspondence to valence and arousal, derived from their ratings collection, would further allow for a more objective comparison amongst the two studies.

The sequence of seven videos of the experiment target at elicitating seven distinct emotions, whose qualitative classification was: Boredom, Joyfulness, Panic/Fear, Interest, Anger, Sadness, and Relaxation. These emotions, when associated to their corresponding valence and arousal values, can be represented bythe radial plots of the specific Valence (a) and Arousal (b) values expected for the seven videos and respective emotions elicited, considering the values expected in [88] for videos 3 & 5, and the values in [33] for the remaining videos, as those where retrieved from these database.

Despite the fact that, as mentioned, during the protocol design it was given preference for the VR videos available in the [33] database, there were two emotions for which there were not corresponding videos. Those emotions were the Panic/Fear and Anger, for which those videos (3 & 5) were selected by independent research on YouTube. This way, the corresponding valence and arousal values for the seven target emotions can be seen in Figure 4.6, which were assigned considering their conceptualization along those dimensions conducted by [88] for videos 3 & 5, and the valence and arousal ratings collection made available by [33] for each of the remaining videos.



(a) Valence.

(b) Arousal.

Figure 4.6: Radial plots representing the Valence and Arousal ratings expected in each video [88].

4.2 Dataset Validation

The usage of methods for denoising and outlier removal takes special importance so as to accomplish a higher level of reliability for the biosignals, as it will be further discussed in the next section. Nevertheless, prior to that step, it matters to assess the overall dataset quality in order to validate it to proceed for analysis. In this work, for the particular case of the EDA and the Temperature biosignals there were further concerns besides the noise and outliers presence, such as signal abnormalities, for which the discussion will hereafter take place.

4.2.1 Electrodermal Activity (EDA)

According to the typical dynamics of an EDA signal, as described in Section 2.3.2, Figure 4.7 illustrates one reliable EDA signal recorded during this work. Nonetheless, by visual inspection, it was noticeable that not all the EDA signals of the participants appeared to be reliable. In fact, some of those presented abnormal curves, lacking physiological meaning. Figures 4.8 and 4.9 illustrate some instances of abnormal EDA signals, associated to three major problems identified and further described.

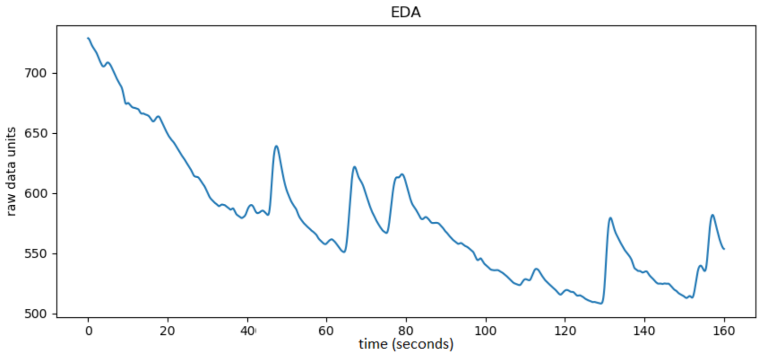


Figure 4.7: Reliable EDA signal (recorded on Participant 3, during Video 3).

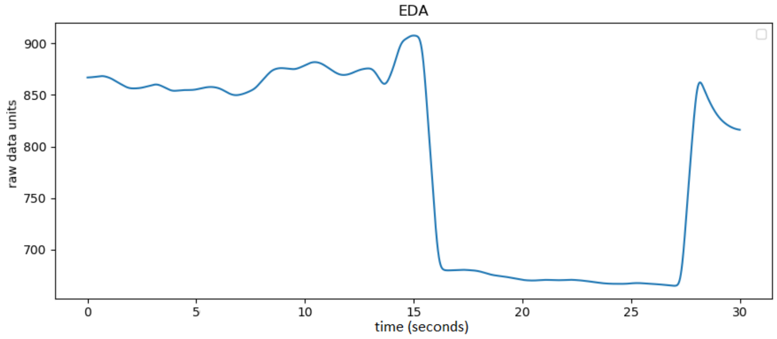
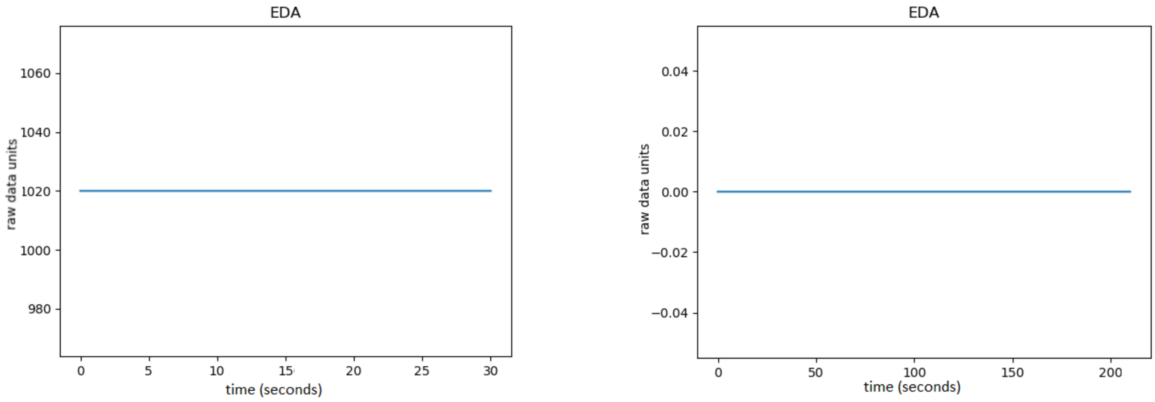


Figure 4.8: Abnormal EDA signal, with quick changes due to finger movement (recorded on Participant 7, during Calibration 1).

Broadly, corrupted EDA signals might be caused by physical movement, environmental factors such as changes in ambient temperature, and electrical noise [29, 89].

On one hand, on a visual inspection perspective, there are signals corrupted by quick changes (fall/rise), as illustrated in Figure 4.8. This is one of the most reported problems about EDA signal acquisition, and it is explained by the finger/hand movement, which can render the signal unusable [29]. In this case, one can infer the participant registered a slight hand a movement between the 15 and 28 seconds.

On the other hand, saturation anomalies were also identified, both at maximum value and at zero (cf. Figure 4.9 (a) and (b), respectively). The actual source of such anomalies cannot be fully explained, but one can point out some potential causes. Firstly, even though an effort was made in order to keep the lab settings from one session to the others, including the room temperature, it is likely that some experiments had an increased temperature due to the fact that part of the experiments were conducted during the summer months. This possibility is corroborated by the fact that, in some participants, the EDA abnormalities were noticed more towards the end of the protocol, which might indicate high temperature in room that lead the participant to sweat more. Moreover, some physiological conditions could also be the source of this early detachment of the electrodes, for instance in the case where the participant



(a) Abnormal EDA signal, due to saturation at the maximum ADC value (recorded on Participant 2, during Calibration 1).

(b) Abnormal EDA signal, due to permanence at zero (recorded on Participant 9, during Video 7).

Figure 4.9: Two abnormal EDA signals, caused by saturation.

suffers from hyperhidrosis, a disorder that involves hyperactivity of the sympathetic nervous system and leads to excessive sweating, including in the palms [90].

The fact that the acquisition setup included two EDA recording possibilities, the EDA acquired from the hand and the EDA acquired from the fingers, gave this work room to, in some cases, mitigate this acquisition anomaly. In fact, when both channels are working properly, the two physiological signals are, as expected, strongly similar.

In several works, the common approach for EDA signals validation, i.e. for the detection of movement artifacts, is to be carried out by visual inspection [41]. Nevertheless, visual inspection is time-consuming and prone to varying interpretations, as there are similarities between SCRs and artifacts [41, 89], for which some algorithms are already being explored towards their automatic detection. In this work, a module, developed in Python, is proposed for EDA artifacts automatic detection, and for EDA signals quality assessment, whose results will be presented and discussed in Section 5.3.

Artifact-free signals exclusively were taken into account for the analysis, having that in the cases where the two EDA signals of a participant were abnormal throughout the experiment, those were discarded from the analysis. The exclusion criteria was to present more than two corrupted signals. Moreover, for each of the participants, only one of the two EDA signals was selected for further analysis, considering the signal collected at the hand as default.

4.2.2 Temperature

The temperature signals collected from all the participants were generally very different from those physiologically expected, as well as anomalously distinct from one participant to the other.

Considering that the data was collected under the same circumstances, two possible causes can be identified. The source of the problem might have been originated at the hardware level (e.g. at the cables or sensor connection). Otherwise, and the most likely reason, consists of possible interference, for instance caused by the skin contact with exposed metal parts of the sensor.

Since this data validation step aimed at selecting only the reliable biosignals, it was decided to leave all the temperature data out of the remaining study.

4.2.3 General Considerations

Despite the efforts in order to keep the same lab and setup settings amongst sessions, as well as following a well-defined elicitation protocol that would facilitate the experimental control, there were unexpected anomalies in the dataset acquisition.

Namely, due to software or the acquisition module, there were sporadic failures to collect portions of data from some of the participants. This lead to failures during the acquisition of the last three and two videos of participants 2 and 3, respectively, video 4 and 7 of participant 23, having shorter acquisition durations than those expected. Complete lost of acquisition portions also occurred, in particular, the video 3 was not successfully recorded for participant 13, as well as the calibration 3 for participant 16, the calibration 4 for participant 22, and the video 3 for participant 23.

Some human errors took place as well, namely the wrong definition of the duration of an acquisition, as occurred in video 3 of participant 15, where 106 seconds were recorded instead of 160.

4.3 Signal Processing

The biosignals acquired during the emotion elicitation protocol present a more significant amount of noise in comparison to ideal acquisition conditions or even to medical contexts. Challenges in the physiological signals processing are further related to their subjective and complex nature, and the sensitivity to movement artifacts [30], which hardens the tasks of finding the ground truth from the raw physiological data. Those are, in fact, always contaminated with noises and other external interferences, as well as artifacts due to electrostatic devices and muscular movements [7].

This section addresses the processing of the raw signals collected from ECG, EDA, BVP, and respiration. Once some denoising techniques are applied, each biosignal is segmented according to their physiology, and a final step of outlier removal might take place.

The resultant pre-processed signals will then be ready to proceed to the next step in the processing workflow (cf. Figure 4.1), the feature extraction, as will be described in Section 4.4.

4.3.1 Filtering

4.3.1.1 Electrocardiogram (ECG)

It is broadly established that ECG signals are low frequency and low amplitude periodic signals. Nonetheless, these are susceptible to external electrical contamination as well as other artifacts [91]. The types of noise that mostly contaminate ECG signals correspond to power line frequency, mismatch of electrode impedance, wandering of the baseline signal, and motion artifacts [82, 92].

In this work, a finite impulse response (FIR) based filtering technique was used for the ECG signals [91] to suppress the high frequency noise, using a bandpass with an upper cutoff frequency of 45 Hz and a lower cut frequency of 3 Hz. Figure 4.10 illustrates one instance of a preprocessed ECG raw signal and the corresponding filtered signal.

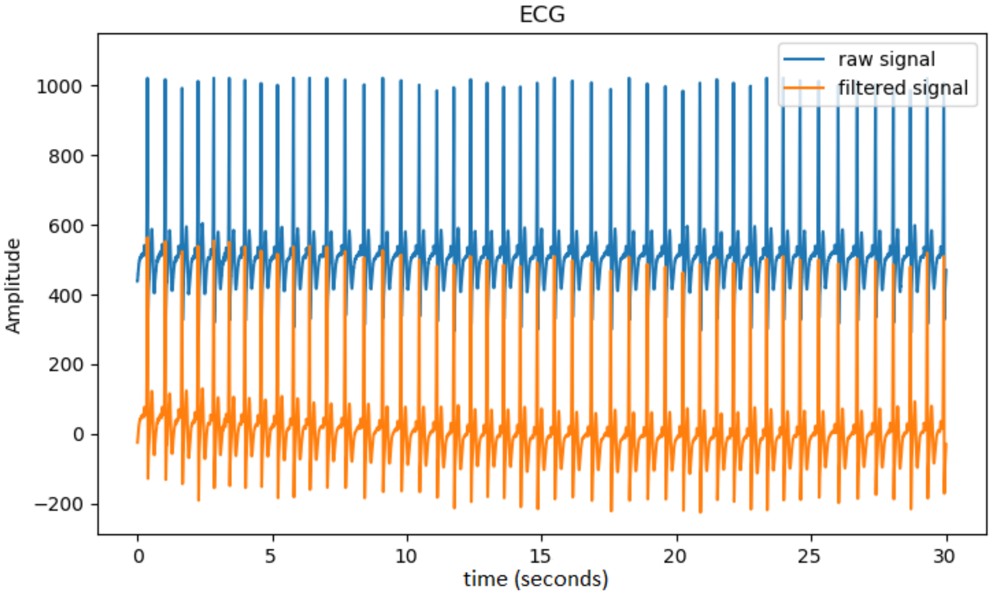


Figure 4.10: Example of a raw and filtered ECG signal.

4.3.1.2 Electrodermal Activity (EDA)

EDA signals were denoised following the method proposed by [44]. The frequency bandwidth of the EDA signal was thus filtered with a low pass Butterworth filter, using the forward and reverse digital filter technique[44, 93]. Filter cutoff frequencies vary widely study to study, ranging from 1 to 5 Hz, depending on the particular signals [89]. In this work, a cutoff frequency of 5 Hz was used.

Figure 4.11 illustrates one instance of a preprocessed raw EDA signal and the corresponding filtered signal, where it is observed that the method is successful in removing powerline interferences.

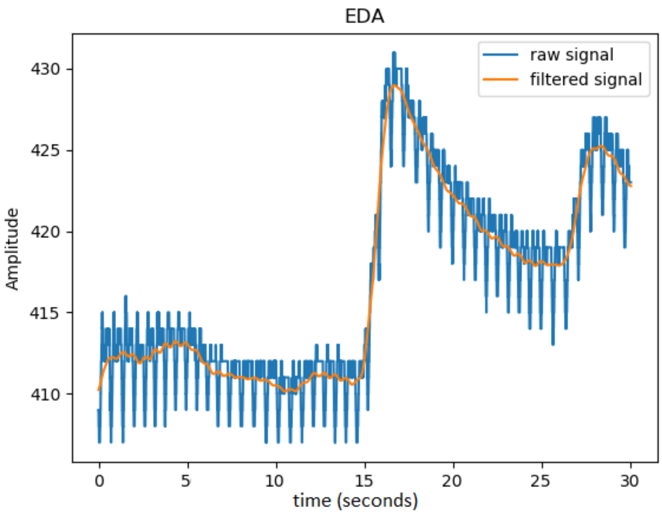


Figure 4.11: Example of a raw and filtered EDA signal.

4.3.1.3 Blood Volume Pulse (BVP)

Even though BVP sensors are usually highly susceptible to noise and artifacts, making it difficult to extract meaningful features from the BVP signals [94], this work used a BVP sensor developed by [45], whose plastic clip-on housing for placement on the finger, to house the light emitter and detector, allowed the minimization of interferences from external light sources.

The method used by [95] was applied so as to eliminate the bursts in the BVP signal, by using a 4th order Butterworth bandpass filter [38], with an upper cutoff frequency of 8 Hz and a lower cutoff frequency of 1 Hz. Figure 4.12 illustrates one instance of a preprocessed raw BVP signal and the corresponding filtered signal.

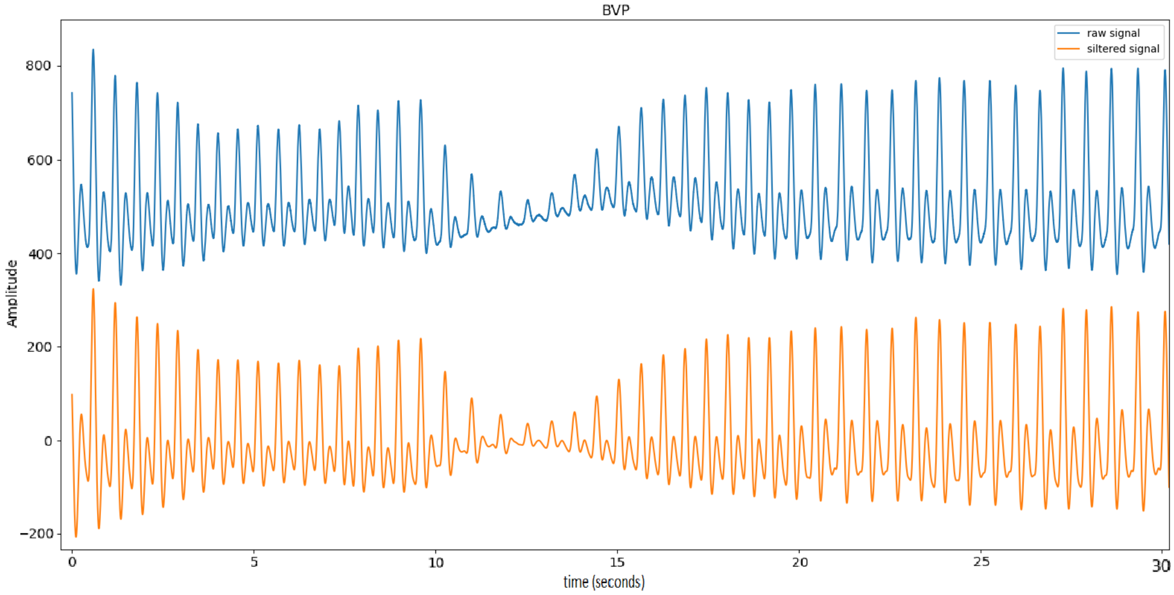


Figure 4.12: Example of a raw and filtered BVP signal.

4.3.1.4 Respiration

Similarly to the ECG signal, the respiration signal was also affected by movement or arm contact artifacts, as they were placed at the chest level. Moreover, some noise is present int the respiration signals acquired.

In order to reduce the noise in the respiration signals, this work followed the approach used by [96]. A FIR filter with cutoff frequency of 0,15 Hz, with order 30, was thus applied. The method was efficient in reducing the noise of the signal.

Figure 4.13 illustrates one instance of a preprocessed raw Respiration signal and the corresponding filtered signal.

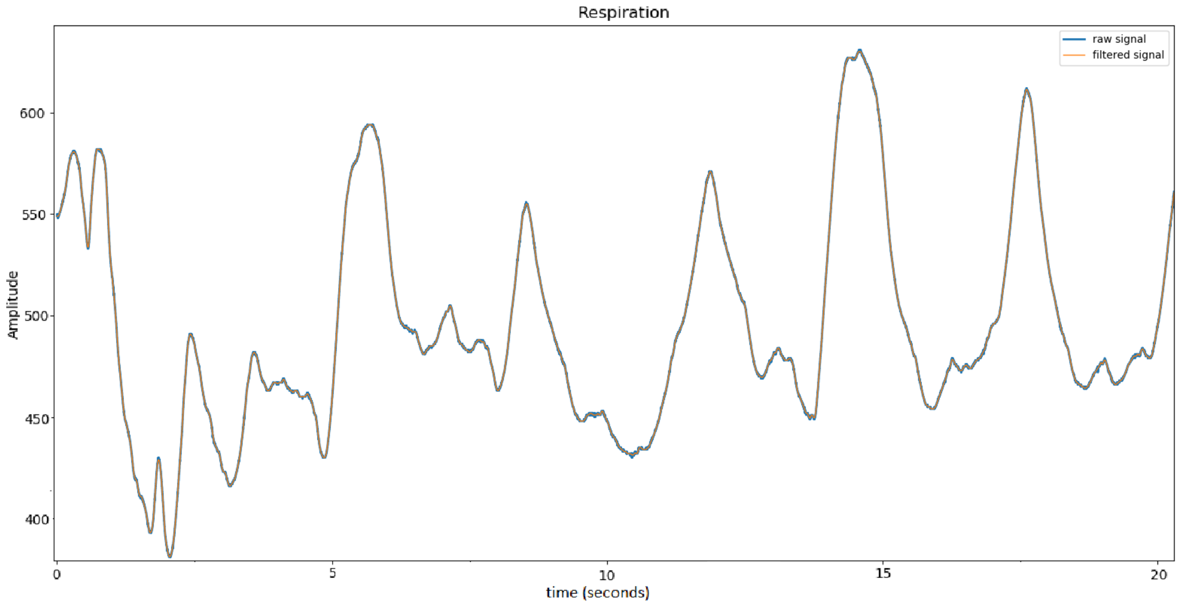


Figure 4.13: Example of a raw and filtered respiration signal.

4.3.2 Segmentation, Outlier Detection and Removal

The segmentation of the four biosignals is critical so as to obtain the respective set of templates, on which the feature extraction procedures will be applied. The computing of this step was mainly performed using the code resources of the BioSPPy library [20], selecting amongst the several algorithms available based on approaches described in the literature for ECG, EDA, and BVP, while for the respiration signal it was considered that higher accuracy could be obtained by implementing new methods described in another studies.

Taking into account the physiologic responses of this biosignals, basically two types of segmentation techniques will take place. In the case of the EDA signal, considering its non-periodicity the segmentation is performed based on the occurrence of relevant skin conductance responses, thus leading to the fact that different EDA templates will typically represent quite different signal durations. On the other hand, considering the periodicity that characterizes the signals of ECG, BVP and respiration, their segmentation was performed on the level of their wave cycles, which will typically present similar durations (within each biosignal), by detecting specific points, as will be further detailed in this section.

ECG and respiration were the biosignals most exposed to movement artifacts, either from the patients movement when looking from one side to the other or even from the accidental contact of their arm with the chestband. These factors made respiration the most vulnerable signal to the presence of outliers and artifacts, for which parts of it were excluded before the segmentation step. In all the periodic signals (ECG, BVP and respiration), a further exclusion step could take place after the cycle waves were identified.

4.3.2.1 Electrocardiogram (ECG)

To perform the segmentation of ECG signals, this work followed the approach proposed by Hamilton [97]. The algorithm yields the R-peak location indices, enabling the identification of the cycle waves (or templates) of the ECG. The basic steps include: using rectified signals; computing a set of filtering steps (low-pass filter with cutoff of 16 Hz, high-pass filter with cutoff of 8 Hz); applying a 80 ms averaging window; identifying potential peaks, and selecting those that represent real R-peaks [97]. This detection occurs by the following the rules: ignoring those that precede or follow larger peaks by less than 200 ms; if a peak occurs, checking whether the ECG signal contained both positive and negative slopes (if not, discard it); if the peak occurred within 360 ms of a previous detection and had a maximum slope less than half the maximum slope of the previous detection, assume it is a T-wave; If the peak is larger than the detection threshold it is considered as a R-peak, otherwise noise; if an interval equal to 1.5 times the average cycle period has elapsed since the previous detection, within that interval there was a peak that was larger than half the detection threshold, and the peak followed the preceding detection by at least 360 ms, it is classified as a R-peak [97].

Then, a correction method is computed to ensure slight corrections in the R-peak locations. In particular, these R-peak locations were corrected by finding the maximum of the signal within a correction tolerance of 0.05 seconds, i.e. within the window from 0.05 seconds before and after the R-peak identified initially.

This procedure yielded adequate segmentation results for the dataset collected in this work, and Figure 4.14 illustrates one ECG signal after this segmentation step, where 10 wave cycles (templates) were detected by identifying the respective R-peaks.

A final step was performed towards removing abnormal ECG templates, by applying an exclusion algorithm based on the physiology of ECG waves proposed in [98]. An acceptance range was computed considering the heart rate in each template, where the ECG templates that did not present heart rate in the interval [30, 200] bpm were considered as outliers and thus excluded. In terms of amplitude, the

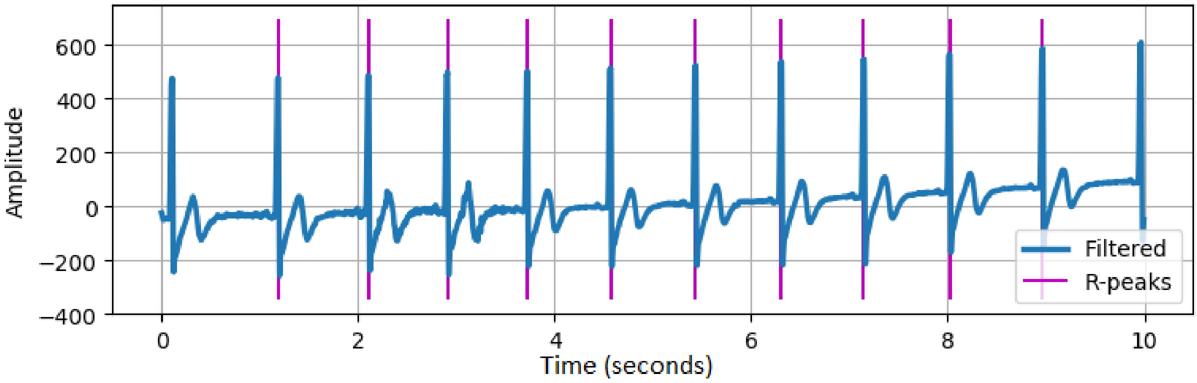


Figure 4.14: Segmentation step performed on one filtered ECG signal, through R-peak detection, resulting in the identification of 10 templates.

acceptable amplitude range for an ECG template to be considered valid was [-400, 550]. The morphology of the QRS complex was also taken into account by computing another exclusion step based on the distance to a template (determined as the mean template of the whole ECG signal), in this case using as metric the cosine distance. This way, the templates affected by contact noise and motion artifacts are removed from the proceeding analysis.

4.3.2.2 Electrodermal Activity (EDA)

For the segmentation of EDA signals, this work followed the method proposed by Kim et al. [30] to detect the occurrence of SCR. The filtered signal is differentiated, and a subsequent convolution with a 20-point Bartlett window is computed. The occurrence of the SCRs is detected by finding two consecutive zero-crossings, which correspond to the onsets of the response, and the SCR amplitude is determined at the peak locations. To consider only the relevant SCRs, the responses with an amplitude smaller that 10% of the maximum of the signal are excluded.

Figure 4.15 illustrates one EDA signal after this segmentation step, where five SCRs (templates) were detected, each associated with the respective onset and peak.

4.3.2.3 Blood Volume Pulse (BVP)

The segmentation of the BVP signals was based on the detection of the onsets of its pulses. This work followed the approach by [99], that converts the BVP waveform into a slope sum function (SSF) signal, in which the initial upslope of the blood pressure waveform is enhanced and the remainder is suppressed, leaving the location of the pulse onset unaltered, and easing the detection of the pulse onset [99].

A final exclusion step was performed in order to avoid the detection of false BVP peaks, using a method found empirically with this dataset for which the IBI was taken into account. Given a complete BVP

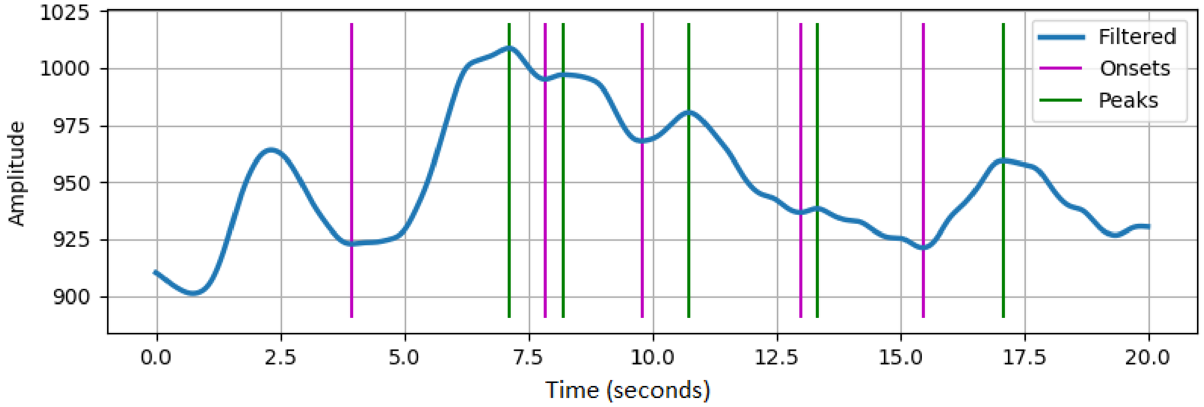


Figure 4.15: Segmentation step performed on one filtered EDA signal, through SCR detection, resulting in the identification of 5 templates.

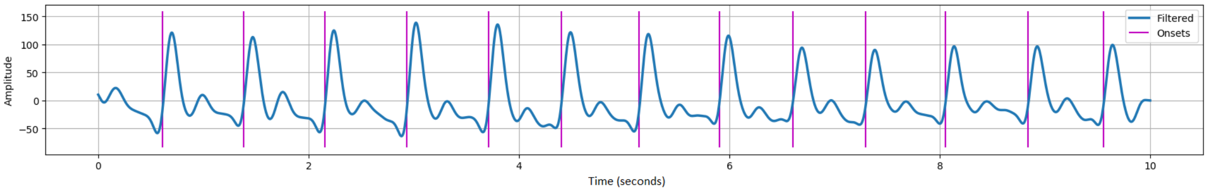


Figure 4.16: Segmentation step performed on one filtered BVP signal, through pulse onsets detection, resulting in the identification of 11 templates.

signal, the mean of the IBI of all the templates was computed and those templates with IBI differing in more than 0.1 seconds from that value were excluded.

4.3.2.4 Respiration

The assessment of the existence and predominance of outliers was used so as to exclude the unreliable parts of the signal, as it was, for some participants, affected by movement and contact artifacts. Hence respiration signals were subjected to outlier removal before the identification of their wave cycles.

The approach used was based on the methods proposed by Rahman et al. [100], that relies on the usage of quartiles for both effective outlier detection and segmentation for this biosignal, as those are less sensitive to spikes that may appear in respiration measurements collected in noisy field environments [100]. Hence, by arranging the data in ascending order, one can and split it in equal groups such that 25% of the observations are in each group [101]. The cutoff points are called quartiles, and there are three of them: the first or lowest quartile (25%);the second or median (50%); and third or upper quartile [100, 101]. Furthermore, the interquartile range (IQR) is determined as the difference between the upper (UQ) and lower quartiles (LQ). This way, outliers might be detected in case the point is smaller than $LQ - 1.5IQR$ or larger than $UQ + 1.5IQR$ [100].

Once the indices of the outliers are computed, the remaining parts of the signal were considered only if there was at least 5 seconds between consecutive outliers, to ensure the detection of at least one respiration cycle.

Regarding the segmentation of the signals, it was performed by detecting its peaks and thus identifying the respiration templates. The upper quartile was used as a lower threshold for a point to be considered a peak, and no more than one peak could be detected in a window of 1,5 seconds (as the duration of each respiration cycle cannot be as short as 1,5 seconds for an individual [100]).

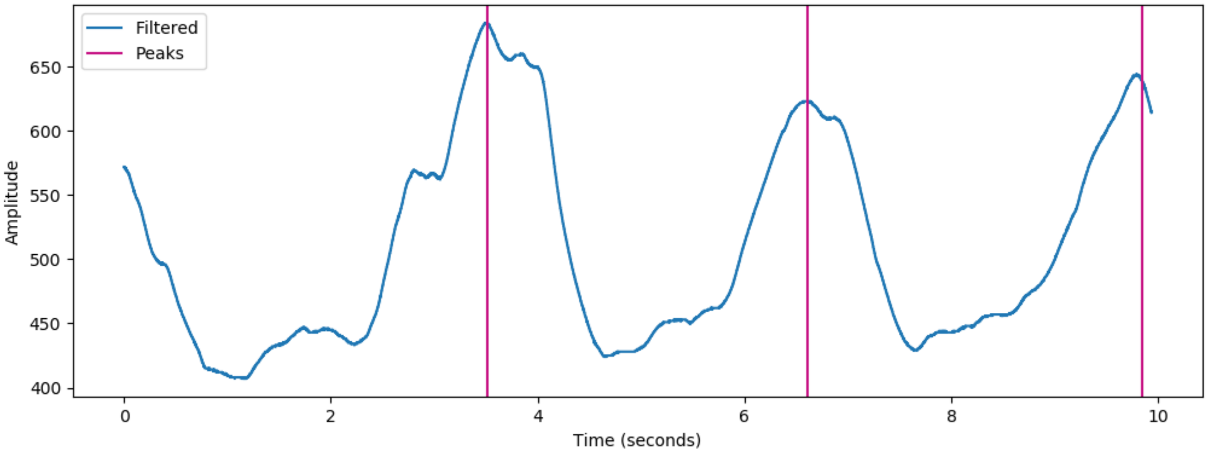


Figure 4.17: Segmentation step performed on one filtered Respiration signal, through peaks detection, resulting in the identification of 2 templates.

4.4 Feature Extraction

As previously stated, given the importance of choosing adequate features towards the purpose of emotion recognition, this section is focused on the description of the analysis and choice made regarding the most valuable features. These will correspond to a set of inputs, representative of each biosignal, to perform the classification tasks.

As proposed in Section 3.3.2.2, in this work each of the biosignals is associated to two types of features: (i) Physiological features and (ii) Statistical features. In this sense, hereafter we will explain the selection of the features along with their association to type (i) or (ii) for each biosignal, as summarized in Table 4.2, found in the end of the section.

4.4.1 Electrocardiogram (ECG)

The heart rate and the heart rate variability have a direct physiological meaning in the cardiac signal and were extracted from each R-wave. Furthermore, to characterize the ECG signals some parameters in the frequency domain were extracted, namely the low frequency band (LF) of the HRV, from 0,04 to

0,15 Hz, and the high frequency band (HF), from 0,15 to 0,4 Hz. From the LF and HF, further features were extracted from their relations, as their ratio LF/HF and their sum LF+HF.

Regarding the type (ii) of features, the statistical features extracted from ECG were the mean, the standard deviation (std), the absolute deviation (ad), and the kurtosis.

4.4.2 Electrodermal Activity (EDA)

Concerning the physiological attributes of EDA, several have been tested in the literature, as previously introduced. In this work, nearly all of them were tested and the ones selected to be extracted from the EDA collected were the rise time (r_time), the signal amplitude difference between the onset and peak (ampPO), the skin conductance response (SCR), the recovery time in which the signal decreases 63% from the peak amplitude (rec_time), and the total amplitude variation during the video (a_var). The mean and the standard deviation were the statistical features selected for this biosignal.

4.4.3 Blood Volume Pulse (BVP)

Physiologically, the information collected from the BVP signals was the inter-beat amplitude (IBI) and the heart rate. In terms of statiscal features, the mean, the median, the maximum amplitude (maxAmp), the kurtosis and the skew were the statistical features extracted from each BVP cycle.

4.4.4 Respiration

In terms of physiological features, the instantaneous respiration rate (IRR) was the only attribute collected from the respiration signals. Regarding statistical attributes, the ones selected were the mean and the median of the respiration cycles.

4.4.5 Overview

Preliminary tests were computed for all the biosignals, testing their performance with various combinations of the seven statistical features mentioned (mean, std, ad, median, maxAmp, kurtosis, and skewness), and the selection was based on those that yielded the most satisfactory results.

At the end of this feature extraction step, all the features of each biosignal are still in their natural units, i.e. the values are the physiological ones. Nonetheless, in the step of preparing them to input the machine learning algorithm, they will suffer a scaling process, as will be described in the next section.

Table 4.2: Set of features selected for extraction from each of the biosignals, split into two groups, the (i) Physiological features, and the (ii) Statistical features.

Biosignal	**(i) Physiological features**	**(ii) Statistical features**
ECG	heart rate, HRV, LF, HF, LF/HF, LF+HF	mean, std, ad, kurtosis
EDA	SCR, r_time, ampPO, rec_time, a_var	mean, std
BVP	IBI, heart rate	mean, median, maxAmp, kurtosis, skewness
Respiration	Instantaneous respiration rate	mean, median

4.5 Classification using Support Vector Machines

The library LIBSVM (A Library for Support Vector Machines), developed by Chih-Chung Chang and Chih-Jen Lin [47], consists of an integrated software for SVM classification, regression, and distribution estimation (one-class SVM). It also supports multi-class classification, which is a requirement for the present work. The library is available in an extensive range of languages, including Python. LIBSVM was thus considered suitable and used for the emotional classification approach followed in this work.

The general guidelines of the methodology proposed [57] were followed:

1. Transforming the data to the format of an SVM package
2. Conducting simple scaling of the data
3. Selecting the appropriate kernel
4. Computing cross-validation to perform model selection
5. Using the optimal model for the whole training set
6. Evaluating the performance

Regarding the classification approach, as stated, this work comprised two classification scenarios, an User-Dependent and an User-Independent.

Since most of the referred steps are very similar between the two scenarios, the methodology previously enumerated will be common to both approaches and hereafter explained, detailing the steps in which the user-dependent and user-independent mostly differ from each other.

4.5.1 Transforming the data to the format of an SVM package

4.5.1.1 User-Dependent

This first approach aims at classifying the emotional state of each participant, using only their own biosignals as training set. In particular, for each individual with fully reliable datasets, 56 classifications take place (4x7x2), as there are four biosignals and seven datasets to classify (corresponding to the seven videos watched), in terms of valence and arousal. For the emotion classification of each video the leave-one-out technique is used. Hence, the training set includes data from the four calibrations and the remaining 6 videos, and corresponding labels, while only the data from the video to classify is left out to classify.

The whole dataset of each biosignal was firstly placed into a matrix with shape [Total of samples x Number of features], where the samples are the segmented templates (from the data acquired from both calibrations and videos) and the number of features varied according to the biosignal (as previously detailed in the Table 4.2). A vector with the correspondent labels, with size [Total of samples] was also composed. These matrices with the whole data will be used for the model selection, as will be seen in the step computed in Section 4.5.3.

The labeling of each sample is done according to the self-assessments collected during the experimental protocol. Despite the emotional self-assessment after each video was performed through the SAM-scale with a range of 9 options for both valence and arousal, the quantity of training data available was not enough for representing the whole range of 9 classes, for which it was decided to compute a simple mapping of the rating into a coarser scale. It was chosen to map the 1-9 ratings into a 3-point scale, as Figure 4.18 illustrates.

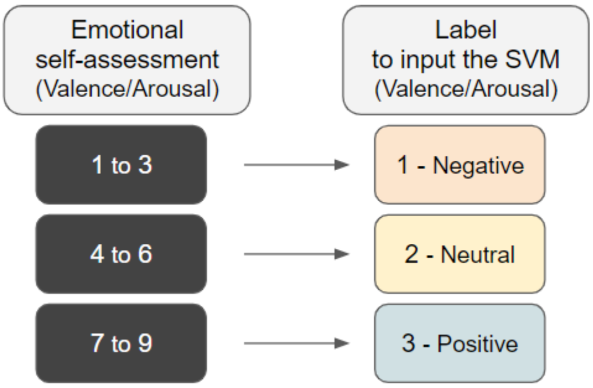

Emotional self-assessment (Valence/Arousal)
Label to input the SVM (Valence/Arousal)
1 to 3
1 - Negative
4 to 6
2 - Neutral
7 to 9
3 - Positive

Figure 4.18: Mapping of the self-assessed ratings of the emotional states from a 9-point scale into a 3-point scale.

Then, for each video testing, the testing set is composed of a matrix [Video samples x Number of features], with the samples of the video to classify, and a training matrix that includes the remaining samples, i.e. the samples from all the remaining videos and calibrations, [Remaining samples x Number of features], and a vector of labels with shape [Remaining samples].

4.5.1.2 User-Independent

As one can derive from the name, the aim of this second approach is to classify the emotional state of a given individual using now the datasets of the whole population as training set.

Regarding the organization of the datasets into a suitable package for the whole population, the procedure is very similar to that described for the user-dependent approach, with the difference that here all the data of each participant is taken into account, generating matrices with much larger dimensions. Nonetheless, unlike the prior case, where both calibrations and videos information was considered for the training set, in this approach only the videos were taken into account, as their information is, in principle, more reliable, and here the amount of data is larger. The testing set is similar to that in the first approach, as one still aims to classify the matrix with the samples of a given video data.

One critical difference in this second scenario concerns the labelling and number of classes under evaluation. Considering that this user-independent approach is more ambitious in terms of difficulty of the classification task, as remarked by the fewer number of studies that explore this scenario (cf. Section 3.3.2.1), it was decided to divide the data in only two classes for each emotion dimension (instead of three classes, as was computed for the previous scenario). This way, Figure 4.19 illustrates the simple mapping of the ratings (in a 9-point scale) into a coarser 2-point scale, correspondent to a negative and positive class. Note that the median of the initial scale, i.e. 5, was included in the negative class, as it ensured more balance in the number the samples in each class.

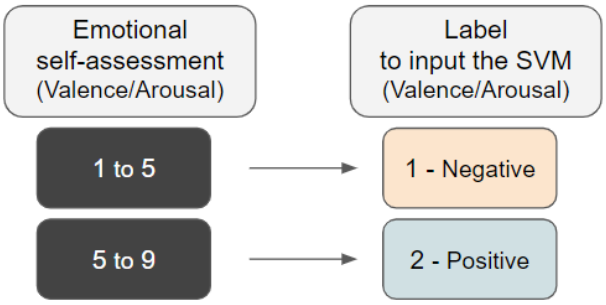

Emotional self-assessment (Valence/Arousal)
Label to input the SVM (Valence/Arousal)
1 to 5
5 to 9
1 - Negative
2 - Positive

Figure 4.19: Mapping of the self-assessed ratings of the emotional states from a 9-point scale into a 2-point scale.

4.5.2 Conducting simple scaling of the data

Performing simple scaling of the data is a step of great importance, as its major advantage is to avoid attributes in greater numeric ranges that could dominate those in smaller numeric ranges; moreover, the own computation becomes easier, avoiding numerical difficulties during the calculation [57].

This step corresponded to the linear scaling of each attribute or feature to the range $[-1, +1]$. Hence, for each of the four biosignals, two vectors were computed from the total dataset matrix, one of them containing the maximum values for each of the features and the other with the minimum values for each of the features. Then, the former would be associated with a vector of 1s and the latter with a vector of -1s, these with same length as the number of features. Then, all the features of the datasets are converted into numbers between -1 and 1, through linear scaling computation.

The scaling of the data is conducted in an individual basis, *i.e.* the scaling of the data is performed amongst all the samples from each participant, instead of scaling the features taking into account all the data from the population. This is done taking into account that there is a relevant variability amongst the physiological features of different individuals, and in this way one can ensure a comparability between the participants.

4.5.3 Selecting the appropriate kernel

Regarding the model selection for the classifier to be used on each of the ECG, EDA, BVP and Respiration datasets, this decision took into consideration the four most common kernels and their characteristics [48, 52, 57]. These kernels are:

- Linear: $K(x_i, x_j) = x_i^T x_j$
- Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$
- Radial basis function (RBF): $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0$
- Sigmoid: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$

where γ, *r*, and *d* are kernel parameters.

Furthermore, the model selection takes into account two main questions: the behavior of the kernel in the cases where the relation between class labels and features is nonlinear; and the number of hyperparameters, i.e. its complexity.

Starting to distinguish between linear and nonlinear classifiers, one has that while the latter map data to a higher dimensional space, linear classifiers work directly on data in the original input space [102]. In terms of the accuracy the usage of linear kernels is comparable, in some applications, to nonlinear kernels, while providing much faster training and testing speed [102]. Nevertheless, and even though the linear kernel has lower complexity, it does not handle the cases where the relation between class labels and attributes is nonlinear. On the contrary, RBF is able to tackle that problem, since it nonlinearly maps samples into a higher dimensional space, for which it is referred as a reasonable first choice [57].

In terms of complexity, the polynomial kernel is the model with the largest number of hyperparameters, which increases its complexity against the other models [57]. Similarly, the sigmoid model has two kernel parameters (i.e. still more complex than the linear and RBF), and has presented some validity issues [57]. Furthermore, it has a scaling factor that affects the quality of the approximation, as by choosing it one has to make a tradeoff between quality of approximation and the rate of convergence [49]. Another key advantage of the RBF kernel in comparison to the polynomial, is that the former has fewer numerical difficulties, as its values are comprised between 0 and 1, while the latter can range from 0 to ∞ [57].

Both linear and nonlinear classifiers are useful under different circumstances [102]. In this work, taking the stated aspects into account, the polynomial and sigmoid kernels were excluded from the further analysis. Even though the RBF kernel appears to be the most advised and suitable, there are also some situations where the RBF kernel is not suitable, for instance when the number of features is very large, in which case the linear kernel might be the best option [57]. In effect, even though linear classifiers fail to handle some inseparable data, they may be sufficient for data in a rich dimensional space [102]. For these reasons, the search over specified parameter values for the optimal estimator that will be described in the following section, besides computing the optimal SVM hyperparameters, also determine the optimal kernel between linear and RBF, as it will be hereafter described.

4.5.4 Computing cross-validation to perform model selection

Unbiased and robust performance evaluation is a key basis of machine learning research, both for finding reliable indication of the relative performance of competing algorithms and avoiding undesirable optimistic bias that can arise due to overfitting in model selection [103]. Model selection strategies for machine learning algorithms typically involve the numerical optimisation of an appropriate model selection criterion, often achieved based on an estimator of generalisation performance, such as K-fold cross-validation (CV) [48, 103, 104]. In K-fold CV, the data set is uniformly partitioned, at random, into K folds of similar size. Then, a classifier is trained using $K-1$ folds, and the performance is determined by testing the classifier in the remaining fold [104].

The two procedures most typically used for tuning the hyperparameters via CV are the (i) non-nested cross-validation (or flat cross-validation [105]) and the (ii) nested cross-validation (or double-cross validation [106]):

- **Non-nested cross-validation:** In this approach, the hyperparameters of each model are tuned to minimise an estimate of generalisation performance, based on CV. This performance estimate, evaluated for those optimal hyperparameter values, is then used to select the best model [106]. Despite this procedure being computationally inexpensive, its main drawback is the optimistic bias that is introduced into the performance estimate, since it has been directly optimised in tuning the hyperparameters [106]. A possible consequence is selecting a model that has not a genuinely higher performance, or a sub-optimal biased model.
- **Nested cross-validation:** This approach fits the model iteratively using a pair of nested loops [103], computing the hyperparameter selection in the inner CV, while the outer loop computes an unbiased estimate of the expected accuracy of the algorithm with CV based hyperparameter tuning [105]. In the inner loop (model fitting/training) there is thus an optimization of a training criterion, while in the outer loop (model selection) the hyperparameters are adjusted to optimise a model selection criterion [103]; in each fold of the outer CV, the hyperparameters of the model are

> tuned independently to minimise an inner CV estimate of generalisation performance [106]. This eliminates the bias introduced by the non-nested CV procedure as the test data in each iteration of the outer CV has not been used to optimise the performance of the model in any way, and may therefore provide a more reliable criterion for choosing the best model. This robust unbiased performance evaluation requires more rigorous and computationally intensive protocols [103, 106].

While some authors argue that the nested CV procedure is overzealous for most of the applications [105], others claim that over-fitting in model selection represents a pitfall in the practical application of machine learning algorithms and in empirical comparisons [103], for which nested CV must be preferred over the non-nested approach. In this work, it was chosen to compute both non-nested and nested CV, to compare their accuracy and associated bias, and then selecting the nested CV results, aiming a robust model with stronger SVM generalization performance, i.e. smaller error rates on test sets [52].

The inner and outer loops were computed using a 4-fold CV, over 30 random trials. The approaches were implemented using several tools available from the class model_selection of the sklearn module [107]. In the inner loop, an exhaustive search over specified parameter values is performed for the estimator. The goal is to determine which parameters suit better the model, and those are optimized by cross-validated grid-search over a parameter grid [107]. It allows not only the search of optimal parameters amongst a given set of C or γ, but also compares accuracy between different kernels. This thus corresponds to the inner loop, the non-nested parameter search and scoring. Then, the nested CV is computed in the outer loop, with parameter optimization by evaluating the scores obtained.

The scores of both non-nested and nested CV are stored in a matrix with the size of the number of trials. In the end, in order to analyze the actual impact of computing the two strategies on this dataset, a comparison takes place by determining the difference between their scores, their average difference and standard deviation. The summarized results are present in the Section 5.6.1. Hence, by finding the highest score index after the nested CV, one can identify the optimal parameters.

At this step, model selection is performed to find the optimal model by testing different combinations over the following set:

- *Kernel*: [RBF, Linear]
- *C* (regularization) parameter: [1, 5, 10, 50, 100]
- γ (Gamma) parameter: [0,01, 0,001, 0,0001] (in case the kernel used is the RBF)

4.5.5 Using the optimal model for the whole training set

Model selection was computed twice for each biosignal of the dataset (for valence and arousal classifications). The total of models was thus eight (2x4) per participant in the user-dependent approach, and eight times for the whole population in the user-independent scenario. These were further implemented for the testing step, by defining the model hyperparameters accordingly.

4.5.6 Evaluating the performance

The final step of both classification tasks is their testing and performance evaluation where, as stated, the testing set of each modality is composed of a matrix with the samples of the video to classify. Then each template is attributed a class, and the most occurred class is given as the classification of that video. In order to compute the fusion of the information of the four modalities, the accuracies obtained in the nested CV for all participants are taken into consideration to establish a weighting for the fusion of the individual biosignals classification information into the final multimodal classifier, as will be exemplified in Sections 5.6.3 and 5.7.3.

The classification performance is analyzed per video and per participant. The results of the user-dependent classifications are detailed in the Section 5.6, and the results for the user-independent case are detailed in the Section 5.7.

Chapter 5

Results

This chapter will include the most relevant results regarding the objectives stated for the present work. Firstly, Section 5.1 provides an overview of the participants included in the study, and Section 5.2 gives some general considerations regarding the visual inspection of biosignals associated with different target emotions. The EDA dataset quality-assessment module, as described in Section 4.2.1, was tested and its results are shown in Section 5.3. Moreover, concerning the emotion elicitation assessment, two analysis took place, both the influence of the day's emotion baseline of the participants on the elicitation outcomes, in Section 5.4, and a comparison between the participants valence-arousal ratings and those expected, in Section 5.5. Finally, the purpose of emotion recognition itself was tested using the SVM and multimodalities fusion, in Sections 5.6 and 5.7, whose results are shown for the user-dependent and user-independent scenarios.

5.1 Sample Characteristics

23 individuals participated in this study, from whom 43,5% were female. Moreover, due to previous evidence of a dedifferentiation of emotional processing in old age [108] and to possible difference in the perceiving of emotions amongst adults in different age ranges, the participants age was limited among 18-40 years-old. The average age of the participants was 23 years-old, with a standard deviation of 3,7.

Only individuals with no history of psychological or neurological conditions were admitted, and none of the participants were reportedly taking any medication that would affect the cardiovascular, respiratory, or central nervous system.

Subjects participated as volunteers in the experiment, and consented to the use of the collected data for the scientific purpose of this work.

5.2 Preliminary Visual Inspection of the Biosignals

Despite this work is focused on the automatic feature extraction and detection of emotional patterns, prior to that more numerical and automatic analysis, a preliminary visual inspection was performed on some signals associated with different target emotions. This visual inspection was useful in the sense of validating whether the characteristics of this dataset were consistent with the most common features described in the literature, as seen in Section 3.3.2.2, which is important to open some general considerations with respect to the differences in the physiological responses within the range of emotions and amongst different individuals.

5.2.1 Arousal

The differentiation in the biosignals due to variations along the arousal dimension might be observed when comparing the physiological responses associated to distinct self-assessed arousal ratings. This section presents examples of portions of ECG, EDA, BVP, and respiration signals retrieved from the same participant while feeling emotions with negative and positive arousal (in order to keep the analysis unbiased and focused on the arousal influence on the signals, the pairs of signals had similar valence ratings, with at most one point difference). The characteristic attributes of arousal manifestation, which were found in the literature and presented in Section 2.3 will thus be hereafter analyzed.

The heart rate information, yielded by both ECG and BVP signals, has been correlated with states of arousal, for which Figures 5.1 and 5.2 illustrate two pairs of ECG and BVP signals associated to opposite arousal by the same participant. One can observe that, in fact, the heart rate found to be larger in the signals associated to higher arousal: in the ECG example, 14 and 13 cycles are found in the high and low arousal signals, respectively, and in the BVP example, 11 and 10 cycles are found in the high and low arousal signals, respectively.

Concerning EDA signals, it was expected that emotions rated with high arousal would be characterized by high skin conductance, for which Figure 5.3 illustrates one instance where that is verified, though not in a considerably strong manner. Nevertheless, in another instance, depicted in Figure 5.4, the contrary is shown, raising questions about the validity of either generalizing this feature or of taking into account the self-assessed ratings by the participants as a fully reliable ground truth, as it might be subjective and biased by their thought on what their answer should be. The skin conductivity responses are also a key characteristic of EDA that manifests arousal states, and Figure 5.5 successfully illustrates an instance in which the participant, while experiencing a high arousal emotional state (8 points in the 9-point SAM scale), had a considerable activity seen in the peak around 38 seconds, while one can see that the EDA was quite stable during a low arousal emotional state (2 points in the 9-point SAM scale).

Finally, the respiration rate tends to reflect arousal [14, 42], for which Figure 5.6 illustrates respiration signals associated to the arousals of 9 and 1. Although the characteristics are controversial and considerably dataset-dependent, fast and deep breathing has been reported to indicate high arousal, while slow and shallow breathing can indicate states of low arousal, e.g. calm happiness [42]. In this example the signal associated to higher arousal appears to have, in fact, a deeper and faster respiration rate.

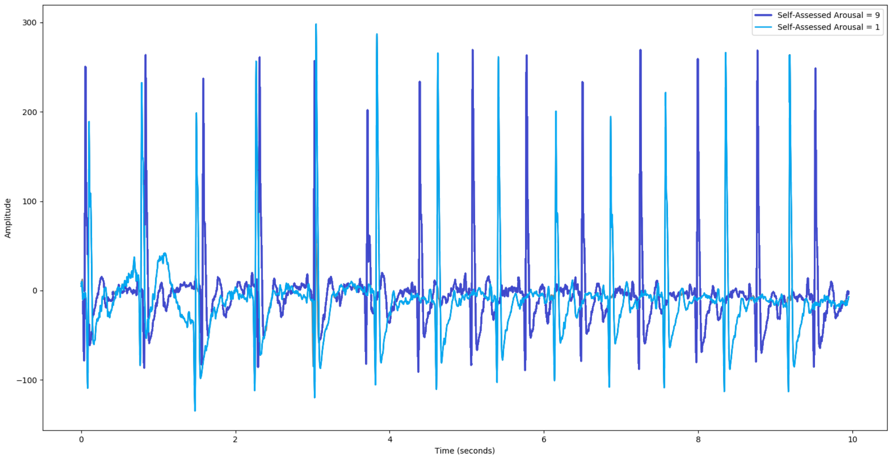


Figure 5.1: Two ECG signals, during emotional states of high and low arousal (9 and 1, respectively).

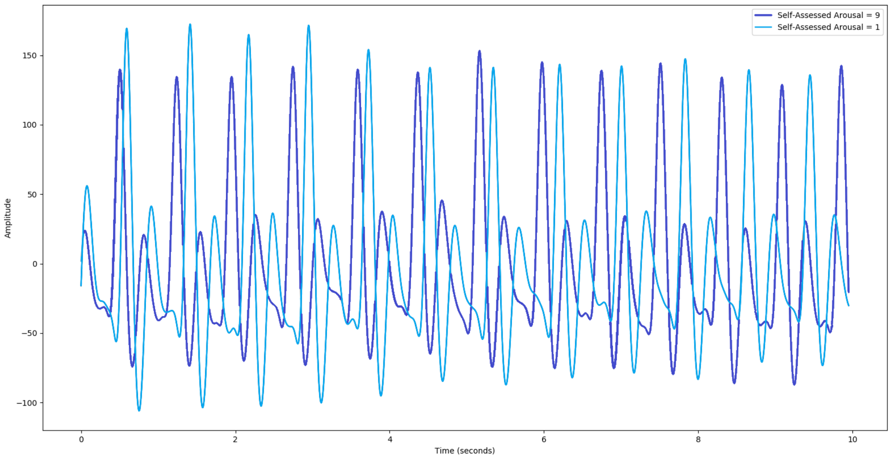


Figure 5.2: Two BVP signals, during emotional states of high and low arousal (9 and 1, respectively).

5.2.2 Valence

In order to assess differentiation in the biosignals due to variations along the valence dimension, one example takes place hereafter. Contrarily to the existence of several attributes of the biosignals that have been consistently correlated with arousal, and as seen in Section 2.3, none of the features ex-

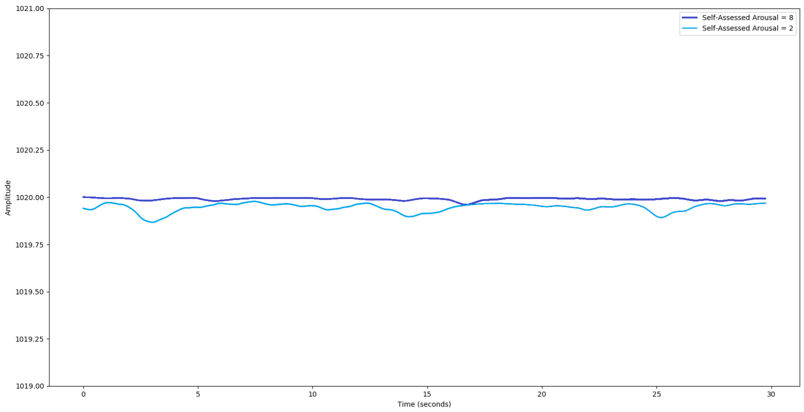


Figure 5.3: Two EDA signals, during emotional states of high and low arousal (8 and 2, respectively).

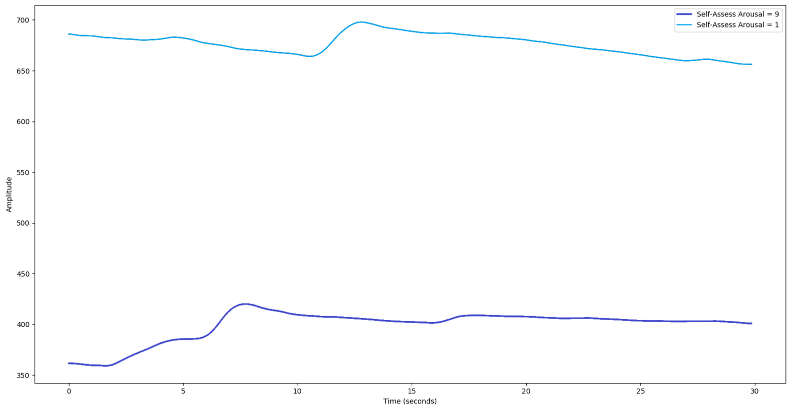


Figure 5.4: Two EDA signals, during emotional states of high and low arousal (9 and 1, respectively).

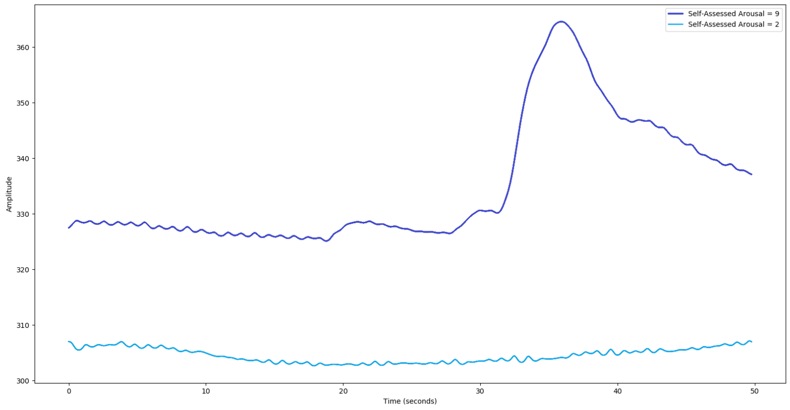


Figure 5.5: Two EDA signals, during emotional states of high and low arousal (9 and 2, respectively).

tracted in the four biosignals of this study is unanimously associated to variation in valence levels across the literature. There are rather different interpretations that can strongly depend on the dataset of the

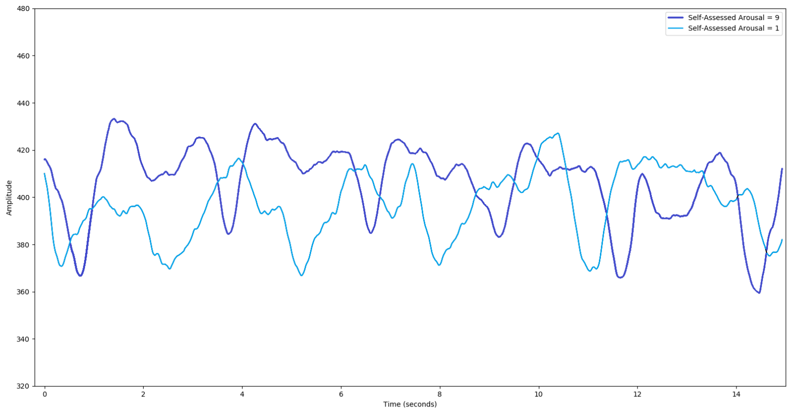


Figure 5.6: Two respiration signals, during emotional states of high and low arousal (9 and 1, respectively).

studies. For instance, positive emotions have been associated with both low and high EDA levels [31]. Similarly, considering respiration information, joy has been associated to both fast and deep breathing [42], and deep and slow breathing [31].

Hence no representative association with valence was found by visual inspection of the dataset used in this work, except for one possible characteristics in the BVP signals. Nonetheless, considering the dataset of this study, it appears that the amplitude of the peak of the BVP signals might have an inverse correlation with valence. However this has not been referenced as a significant indicator of valence variations in any of the literature reviewed, for which the matter would need further study to evaluate the validity of the hypothesis. Figure 5.7 illustrates one instance in which it is verified that the peaks of the BVP waveforms are almost always larger in the signal rated with lower valence (2 points in the 9-point SAM scale) than in the signal rated with higher valence (7 points in the 9-point SAM scale).

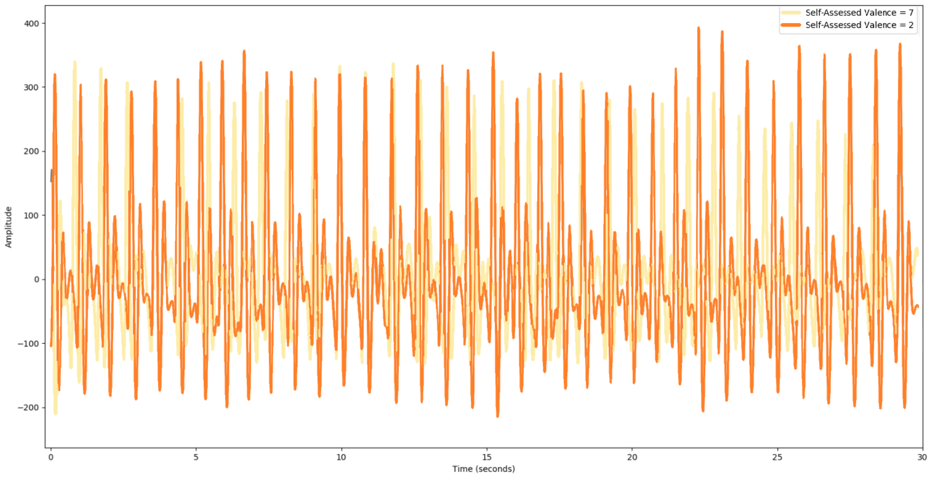


Figure 5.7: Two BVP signals, during emotional states of positive and negative valence (7 and 2, respectively).

5.3 EDA Dataset Validation with Automatic Quality Assessment

Concerning the three types of anomalies identified in the EDA signals of this dataset, within the 490 EDA time-series acquired, the module was able to identify 96 saturated signals, 11 EDA time-series saturated at zero value, and 22 EDA time-series corrupted by fast changes in the signal due to finger/hand movements, as detailed in Table 5.1.

Although the latter is one of the most reported problems in the literature [29] concerning EDA signal acquisition, one can explain its low prevalence in this dataset by the cautionary indication that was given to participants at the beginning of the experiment, preventing them to move the finger/hand and thus rendering the signal unusable.

Table 5.1: EDA quality assessed by the EDA dataset validation module proposed in Section 4.2.1.
* signal excluded due to saturation level
** signal excluded due to saturation at zero
*** signal excluded due to finger movements artifacts

	EDA	Calibration 1	Calibration 2	Calibration 3	Calibration 4	Video 1	Video 2	Video 3	Video 4	Video 5	Video 6	Video 7
Participant 1	Hand	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
	Fingers	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
Participant 2	Hand	*	*	*	*	*	*	*	*	*	*	*
	Fingers	OK	OK	OK	OK	***	OK	OK	OK	***	OK	OK
Participant 3	Hand	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
	Fingers	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
Participant 4	Hand	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
	Fingers	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
Participant 5	Hand	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
	Fingers	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
Participant 6	Hand	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
	Fingers	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
Participant 7	Hand	***	***	***	***	*	*	*	*	*	OK	***
	Fingers	***	***	***	OK	OK	OK	OK	**	**	OK	OK
Participant 8	Hand	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
	Fingers	OK	OK	*	OK	OK	OK	OK	OK	OK	OK	OK
Participant 9	Hand	*	*	*	*	*	*	*	*	*	*	**
	Fingers	OK	*	*	*	OK	OK	OK	OK	OK	OK	**
Participant 10	Hand	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
	Fingers	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
Participant 11	Hand	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
	Fingers	*	OK	OK	OK	*	OK	OK	OK	OK	OK	OK
Participant 12	Hand	*	*	*	OK	OK	OK	OK	OK	**	**	**
	Fingers	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
Participant 13	Hand	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
	Fingers	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
Participant 14	Hand	OK	OK	OK	OK	OK	***	OK	1 fall	***	***	1 fall
	Fingers	OK	OK	OK	OK	OK	***	OK	OK	***	***	***
Participant 15	Hand	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
	Fingers	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
Participant 16	Hand	OK	OK		***	OK	OK	OK	***	OK	OK	***
	Fingers	*	*		*	OK	**	*	*	*	OK	OK
Participant 17	Hand	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
	Fingers	*	*	*	*	*	*	*	*	*	*	*
Participant 18	Hand	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
	Fingers	1 fall	*	*	*	*	*	OK	***	**	**	**
Participant 19	Hand	OK	OK	*	*	*	OK	OK	OK	OK	OK	OK
	Fingers	*	*	*	OK	*	OK	*	*	*	*	*
Participant 20	Hand	OK	OK	OK	OK	OK	OK	*	OK	OK	OK	OK
	Fingers	OK	OK	OK	OK	OK	OK	*	OK	OK	OK	OK
Participant 21	Hand	OK	OK	OK	OK	OK	OK	OK	OK	OK	***	OK
	Fingers	OK	OK	OK	OK	*	*	*	*	*	*	*
Participant 22	Hand	*	OK	OK		*	*	*	*	*	*	*
	Fingers	OK	OK	OK		OK	OK	OK	OK	OK	OK	OK
Participant 23	Hand	*	*	*	*	*	*		*	*	*	*
	Fingers	OK	OK	OK	OK	OK	OK		OK	OK	OK	OK

Table 5.2 summarizes these results, and describes the portion of EDA signals corrupted due to each anomaly, thus unusable for further analysis. Recalling the exclusion criteria defined in Section 4.2.1, it was decided to exclude the EDA datasets of five participants.

- Data from participants 2, 7, 9, 16 and 19 was excluded due to its unreliability throughout the

Table 5.2: Sources of EDA artifacts and percentage of corrupted signals.

EDA artifacts/anomalies	Percentage of corrupted signals
Saturation	19,59%
Permanence at zero	2,24%
Movement artifacts	4,49%

dataset.

- EDA recorded at the hand was used by default for the remaining participants, excepting the participants 12 and 23, since the EDA recorded at the fingers showed higher quality (cf. Table 5.1).

Only artifact-free signals were considered for further analysis, thus including data from 18 participants.

5.4 Influence of Day's Emotion Baseline in the Elicitation Outcome

It was observed that the mean emotional baseline, felt by the participants at the day of the experiment, was 7 (in the 9-points SAM scale) in terms of valence, with standard deviation of 1,096; and 5 for arousal, with standard deviation of 1,702.

To investigate possible influence of the initial emotional state of a subject into the efficacy of the further elicitated emotions, we can start by observing the scatter plots of Figures 5.8-5.21, where the self-assessment, for each video, is plotted against the respective day's baseline, with grey circles (darker circles represent overlaid samples).

In an ideal scenario, assuming that there is no subjectiveness in the self-assessment, one could expect that there would not exist any influence of the emotion baseline if the grey circles would be represented in an horizontal line, meaning that every participant, regardless of their particular emotional baseline, would self-assess each video with the same rating. On the other hand, a highest possible influence of the emotion baseline would be revealed for grey circles aligned along the diagonal (i.e. the elicitation outcomes would be proportional to the emotion baseline of the day of the experiment). The existence of grey circles along the vertical direction is due to subjectiveness of the emotion assessment, not necessarily the influence of the emotion baseline of the participant during that day.

By visual inspection, one can notice that there might be a larger influence regarding the arousal baseline, as these plots tend to present grey circles more spread across the plane and with some sign of correlation between emotion baseline and the emotion self-assessed (elicitation outcomes).

To evaluate the validity of this qualitative inspection, the Pearson's Correlation Coefficient is used to statistically evaluate the relationship between each dimension (valence or arousal) baseline and the elicitation outcome in each video. This acknowledged statistic metric has previously been used in the

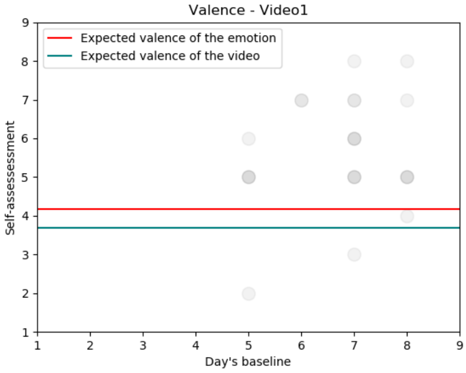


Figure 5.8: Distribution of the valence ratings of Video 1 against day's baseline.

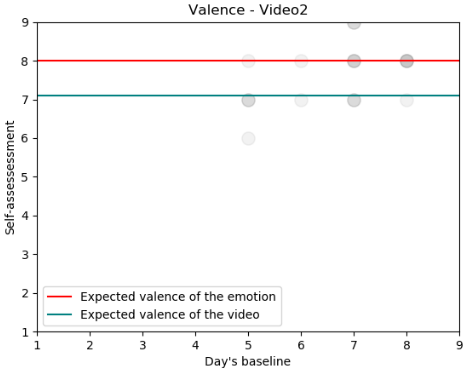


Figure 5.9: Distribution of the valence ratings of Video 2 against day's baseline.

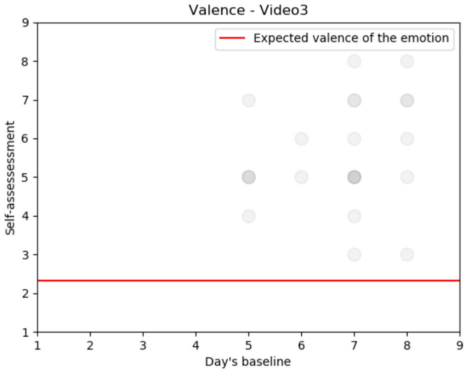


Figure 5.10: Distribution of the valence ratings of Video 3 against day's baseline.

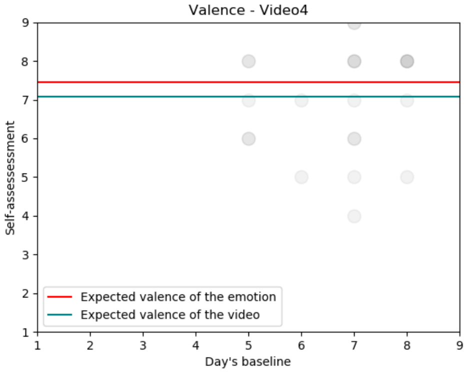


Figure 5.11: Distribution of the valence ratings of Video 4 against day's baseline.

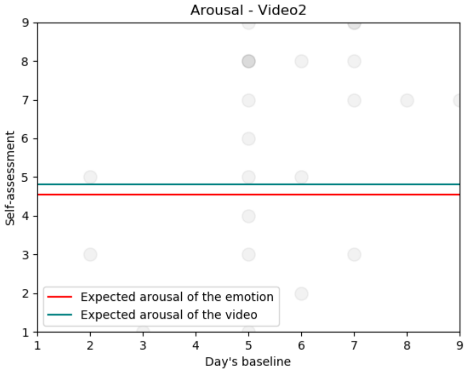


Figure 5.12: Distribution of the valence ratings of Video 5 against day's baseline.

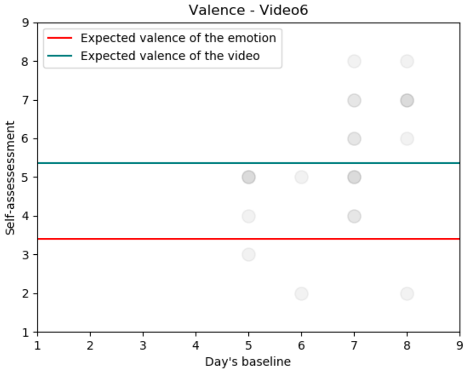


Figure 5.13: Distribution of the valence ratings of Video 6 against day's baseline.

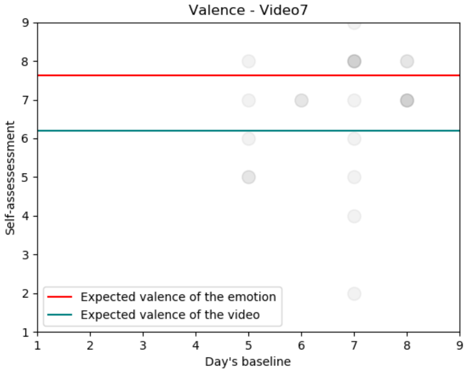


Figure 5.14: Distribution of the valence ratings of Video 7 against day's baseline.

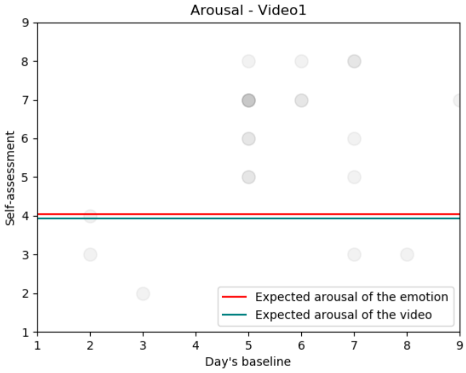


Figure 5.15: Distribution of the arousal ratings of Video 1 against day's baseline.

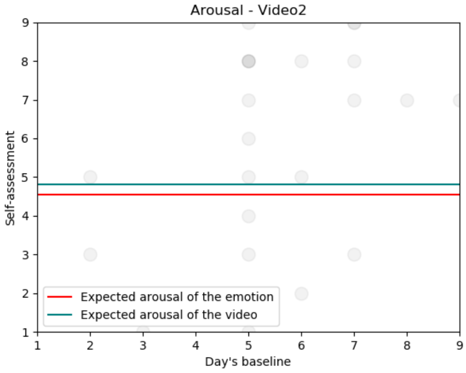


Figure 5.16: Distribution of the arousal ratings of Video 2 against day's baseline.

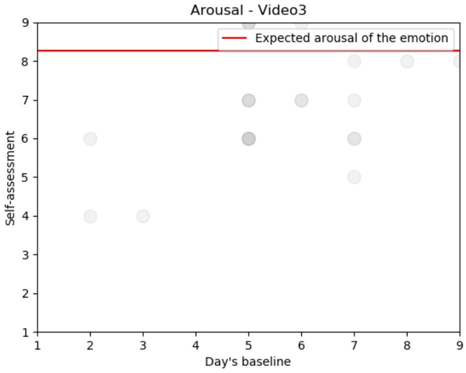


Figure 5.17: Distribution of the arousal ratings of Video 3 against day's baseline.

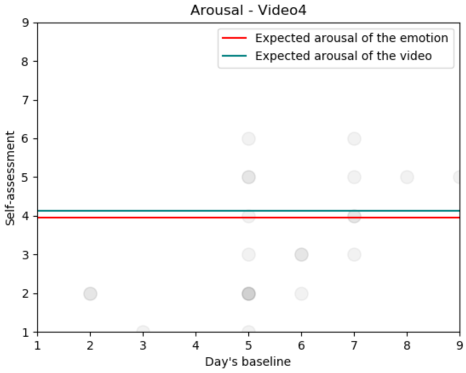


Figure 5.18: Distribution of the arousal ratings of Video 4 against day's baseline.

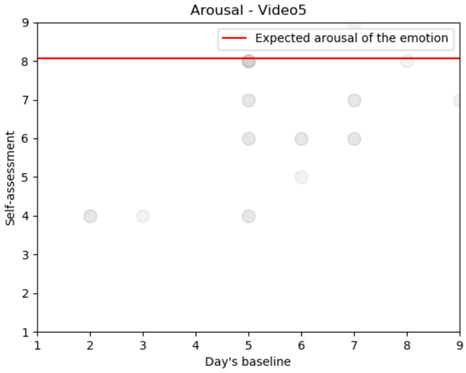


Figure 5.19: Distribution of the arousal ratings of Video 5 against day's baseline.

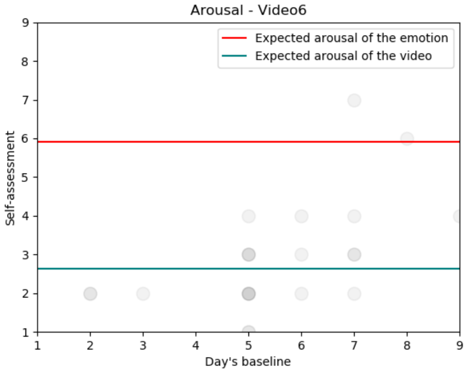


Figure 5.20: Distribution of the arousal ratings of Video 6 against day's baseline.

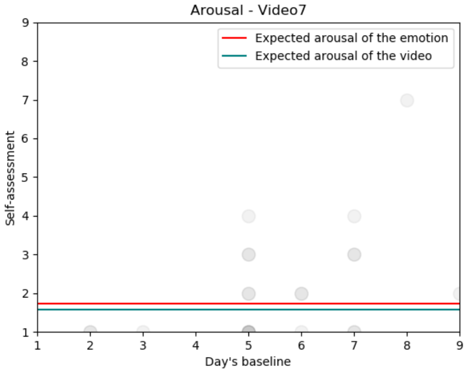


Figure 5.21: Distribution of the arousal ratings of Video 7 against day's baseline.

context of emotion assessment, for instance to evaluate the relation between emotion and head movement data [33]. In this case, the Pearson's correlation coefficient are determined between the emotion and emotion baseline during the day of the experiment, as detailed in Tables 5.3 and 5.4.

Table 5.3: Pearson's correlation coefficients obtained between the day's baseline of valence and each video's self-assessed valence. Correlations marked with * presented p $<$ 0,05.

Valence	**Video 1**	**Video 2**	**Video 3**	**Video 4**	**Video 5**	**Video 6**	**Video 7**
Correlation	0,216	0,443*	0,188	0,120	0,316	0,442*	0,211
P-value	0,322	0,034	0,390	0,585	0,141	0,035	0,334

Table 5.4: Pearson's correlation coefficients obtained between the day's baseline of arousal and each video's self-assessed arousal. Correlations marked with * presented p $<$ 0,05.

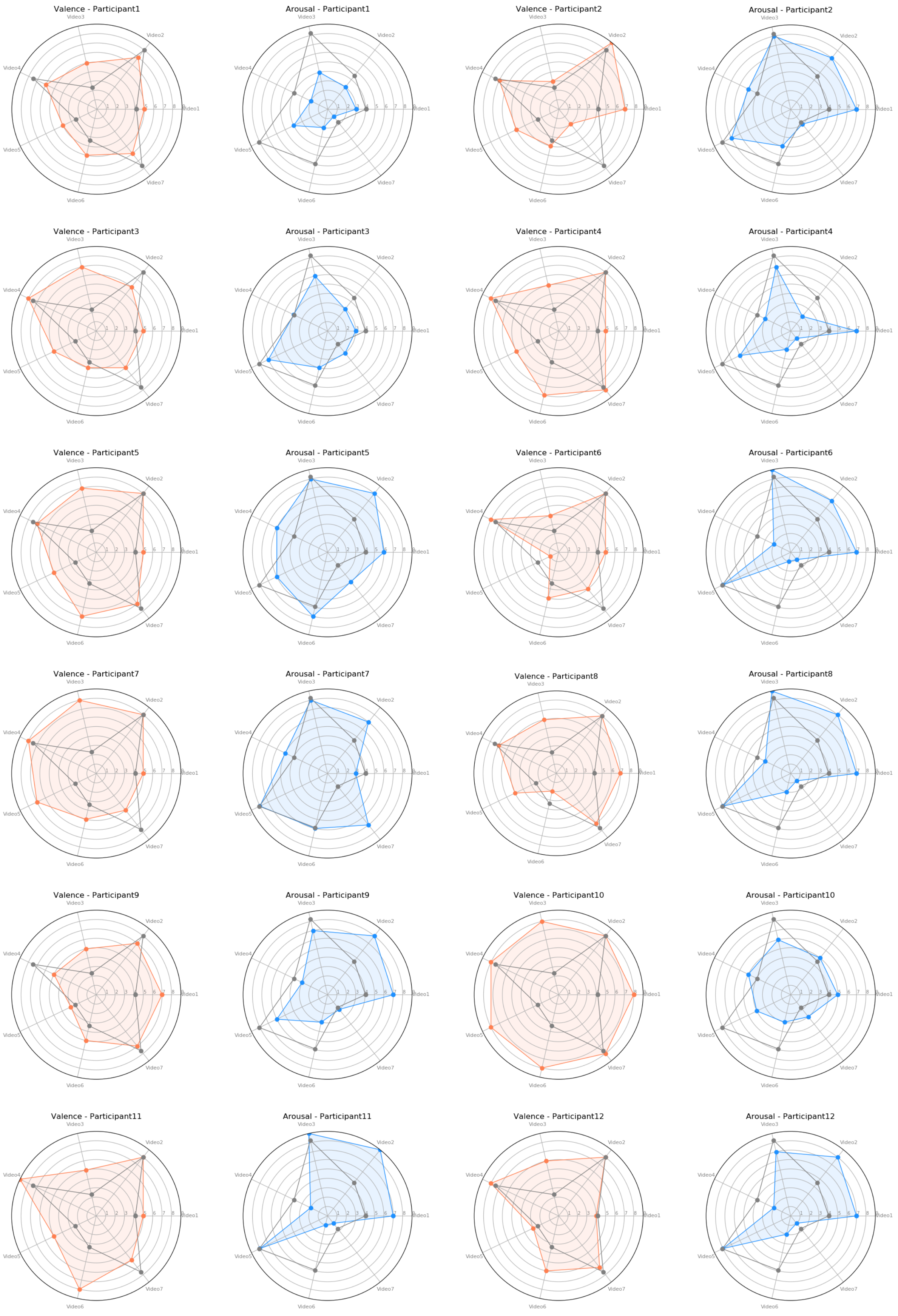
Arousal	**Video 1**	**Video 2**	**Video 3**	**Video 4**	**Video 5**	**Video 6**	**Video 7**
Correlation	0,336	0,415*	0,420*	0,564*	0,540*	0,556*	0,452*
P-value	0,117	0,049	0,046	0,005	0,008	0,006	0,030

As expected, these tables show that the arousal baseline has slightly higher influence on the elicitation outcomes than the valence baseline. In fact, the Pearson's correlation obtained with respect to the arousal baseline influence was larger than that in valence, in all the videos. Furthermore, the p-value associated to those correlations was significant for six out of seven videos in the arousal analysis, whereas only two videos presented statistical significance (p-value$<$0.05).

5.5 Expected vs. Self-assessed Emotions

In order to evaluate the elicitation success obtained with each video, the self-assessed ratings are hereafter compared to their target emotion, in terms of valence and arousal. Figures 5.22 and 5.23 illustrate, for each participant, the valence (orange) and arousal (blue) ratings self-assessed for the seven videos, plotted along with their expected values (in grey) [88]. The average of the ratings and respective standard deviation have been calculated for each video, to summarize the self-assessments given by the participants, whose results are plotted in Figure 5.24.

Moreover, one can compare the ratings reported by the participants of this work with those obtained from the populations of previous studies. In particular, two comparisons are performed: the first one is with respect to the target emotions of the videos and thus analyzing the results of this study against to their conceptual labelings in the literature [88]; and the second takes into account that most of the VR videos used were retrieved from the database by Li et al. [33], which are associated to the ratings obtained for the specific video in the population of their study. The study by Hepach et al. [88] had a population of 100 participants (50% feminine), with mean age of 37 years-old and standard deviation of 12, whereas the study conducted by Li et al. [33] had 95 participants (56% feminine) with age in the range of 18-24 years-old. for the five videos retrieved from the database by Li et al. [33], one can analyze the difference between the ratings obtained in the populations of the two studies, comprised in Table 5.5.



Valence - Participant1

Arousal - Participant1

Valence - Participant2

Arousal - Participant2

Valence - Participant3

Arousal - Participant3

Valence - Participant4

Arousal - Participant4

Valence - Participant5

Arousal - Participant5

Valence - Participant6

Arousal - Participant6

Valence - Participant7

Arousal - Participant7

Valence - Participant8

Arousal - Participant8

Valence - Participant9

Arousal - Participant9

Valence - Participant10

Arousal - Participant10

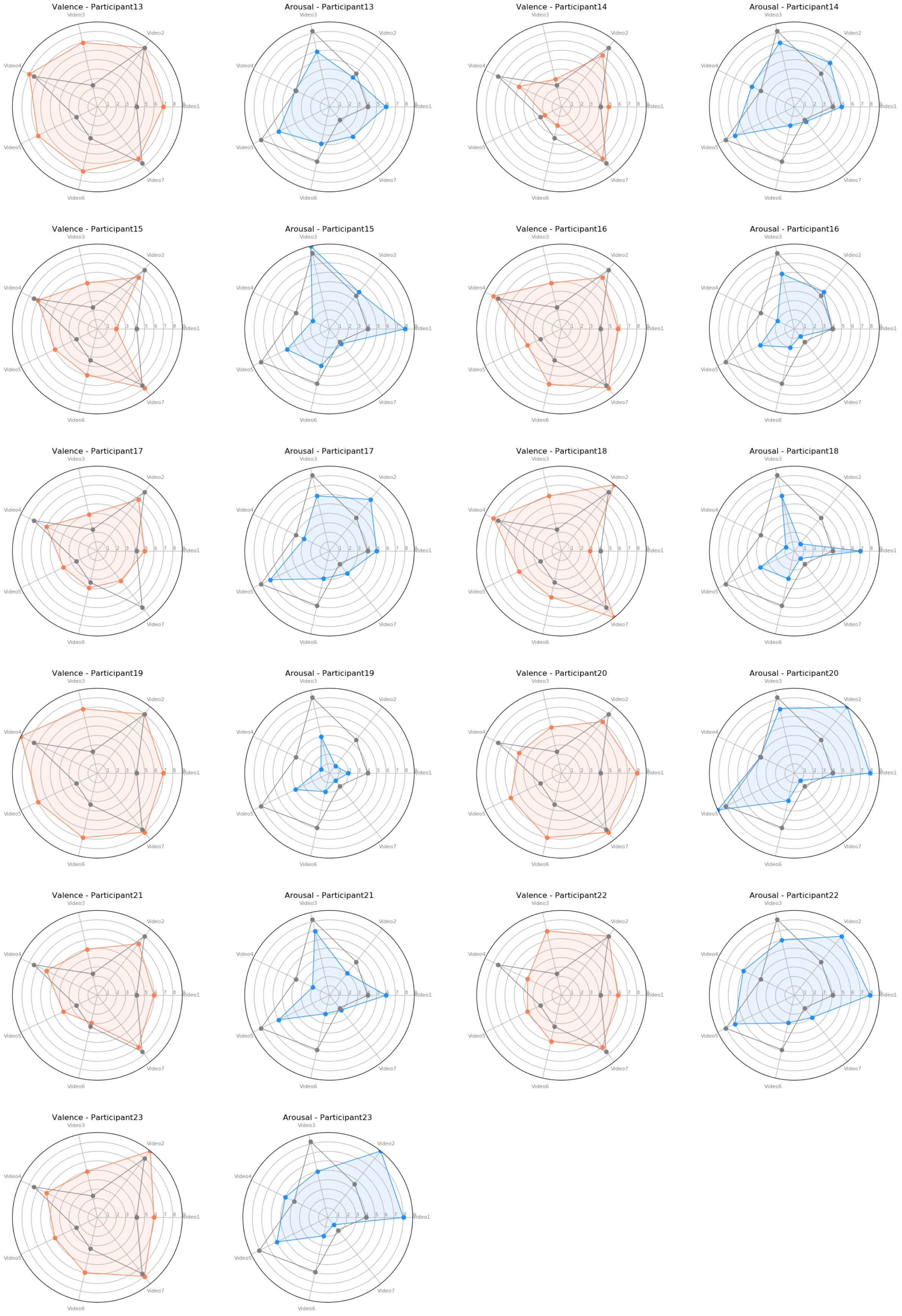
Valence - Participant11

Arousal - Participant11

Valence - Participant12

Arousal - Participant12

Figure 5.22: Radial plots representing the Valence (orange) and Arousal (blue) ratings self-assessed by each participant after each video, as well as the expected values (grey). (1/2)



Valence - Participant13

Arousal - Participant13

Valence - Participant14

Arousal - Participant14

Valence - Participant15

Arousal - Participant15

Valence - Participant16

Arousal - Participant16

Valence - Participant17

Arousal - Participant17

Valence - Participant18

Arousal - Participant18

Valence - Participant19

Arousal - Participant19

Valence - Participant20

Arousal - Participant20

Valence - Participant21

Arousal - Participant21

Valence - Participant22

Arousal - Participant22

Valence - Participant23

Arousal - Participant23

Figure 5.23: Radial plots representing the Valence (orange) and Arousal (blue) ratings self-assessed by each participant after each video, as well as the expected values (grey). (2/2)

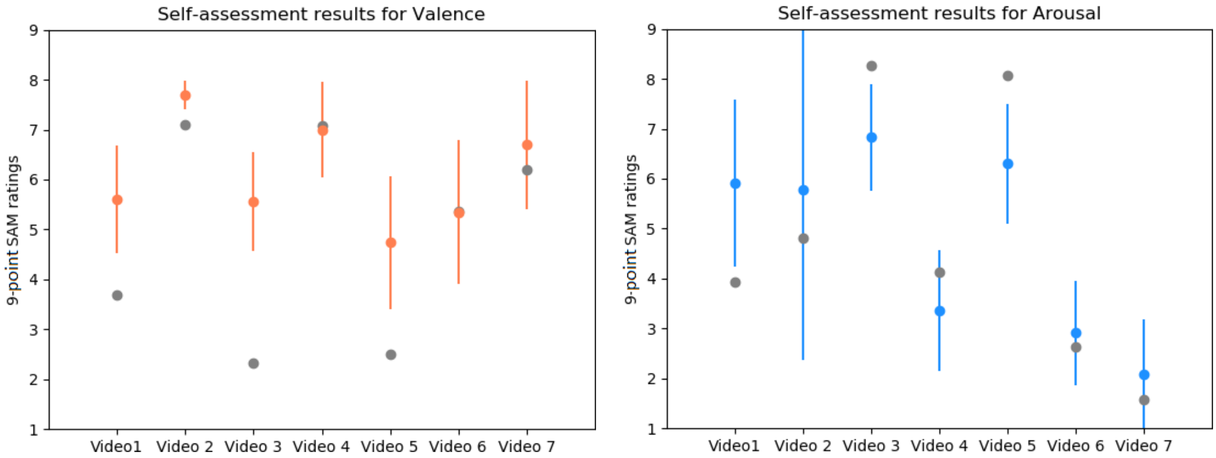


Figure 5.24: Average of the self-assessment ratings for valence and arousal, obtained for each video, along with the respective standard deviation. The gray points represent the expected ratings.

5.6 User-Dependent Classification

5.6.1 Non-nested vs. Nested Cross-Validation Bias Assessment

In this subsection the results of the non-nested CV procedure are compared against those of the nested CV computed on the four modalities of the work, analyzing the accuracies and bias obtained for each. As introduced in the Section 4.5, both techniques are commonly used, having that in principle the nested CV provides more realistic estimates of the performance of the model, as in this case the model selection step is performed separately in each fold of the CV [103].

Table 5.6 shows a comparison of the 4-fold CV estimates using the non-nested and nested model selection procedures, by presenting the average difference between their scores. One can observe that the non-nested CV procedure consistently yields to optimistic bias with respect to the more rigorous internal CV procedure. Nevertheless, for the datasets of this work, the bias has shown not to be statistically significant, since its average was not twice the standard error of the estimates in any of the cases.

For this reason, despite this work following the most cautionary option of selecting the model through a nested procedure, the error of using a non-nested procedure would not have been statistically significant.

Table 5.5: Summary of the ratings obtained during the emotional elicitation of this work, and comparison (differences of mean ratings) with the ratings reported in two previous studies in the literature [33, 88].

		Results from this Work				Reported in Previous Studies [33, 88]				Differences of the mean Ratings			
	Target	Valence		Arousal		Valence	Arousal	Valence	Arousal	Study in [33]		Study in [88]	
	Emotion	Mean	SD	Mean	SD	Mean	Mean	Mean	Mean	Valence	Arousal	Valence	Arousal
Video 1	Boredom	5,61	1,08	5,91	1,68	3,69	3,94	4,17	4,04	1,92	1,97	1,44	1,87
Video 2	Joy	7,74	0,24	5,78	3,41	7,10	4,80	8,00	4,55	0,64	0,98	-0,26	1,23
Video 3	Fear	5,57	0,99	6,83	1,08			2,33	8,26			3,24	-1,43
Video 4	Interest	7,00	0,96	3,35	1,21	7,07	4,13	7,46	3,96	-0,07	-0,78	-0,46	-0,61
Video 5	Anger	4,74	1,33	6,30	1,20			2,50	8,06			2,24	-1,76
Video 6	Sadness	5,35	1,44	2,91	1,04	5,36	2,64	3,40	5,91	-0,01	0,27	1,95	-3,00
Video 7	Relaxation	6,70	1,29	2,09	1,09	6,19	1,57	7,63	1,72	0,51	0,52	-0,930	0,37

Table 5.6: Comparison of non-nested and nested cross-validation procedures in terms of bias. The relative bias of each dataset are presented in the form of the average error rate over 30 trials, *i.e.* the mean difference between the nested and non-nested scores, along with the associated standard error.

		Average Bias	**Standard Deviation**
ECG	Valence	0,0031	0,0020
	Arousal	0,0032	0,0019
EDA	Valence	0,0242	0,0230
	Arousal	0,0318	0,0320
BVP	Valence	0,0050	0,0033
	Arousal	0,0043	0,0029
Respiration	Valence	0,0200	0,0311
	Arousal	0,0279	0,0247

5.6.2 Model Selection and Performance in each Modality

Regarding the model selection itself, one might observe that the decision of selecting the model in terms of both kernel and respective hyper parameters was appropriate. In fact, for each of the four biosignals, none of the kernels (RBF and Linear) was predominantly selected over the other, as well as a wide range of C and γ parameters were selected. More specifically, the selection of the RBF kernel (instead of the linear kernel) occurred in 30,5% of the cases for ECG, 50% for EDA, 22,2% for BVP and 52,8% for Respiration, with different C and γ hyperparameters.Considering the number of features used in each biosignal, which is 10 for ECG, 7 for EDA and BVP, and 3 for Respiration, and the kernel selection guidelines found in the literature and discussed in Section 4.5, it was expected that the linear kernel would be more preferable for biosignals with larger number of features. In this sense, these results could suggest a slight relation between the number of features extracted in each biosignal and the optimal type of kernel, as the biosignal with less features (Respiration) presented the lower predominance of the linear kernel, and the one with larger number of features (ECG) presented a high preference for the linear kernel. However, the fact that BVP, not being the biosignal with larger number of features, was the one with highest prevalence of the linear kernel, and that the amount of data and number of features is not so large in any of the datasets, do not allow to take conclusions on this matter.

In terms of accuracy, the nested CV scores are detailed in the Table 5.7, for the model selection of both valence and arousal classifications for all the individuals. Taking into consideration all those results, the average accuracy and respective standard deviation are calculated at each column, so as to compare, for each dimension (valence/arousal), the accuracy provided by each of the four biosignals.

5.6.3 Multimodal Classifier

The performance capability of each modality in the nested CV is translated into the average accuracy obtained for each biosignal, detailed in the Table 5.7. The fusion of the information from the four biosignals into one multimodal classifier was done by weighting each modality accordingly. Thus, the relative accuracy was determined (present at the last row of the table) and used to establish a weighting for each

Table 5.7: Accuracy scores of the models selected through nested cross-validation, for each emotion dimension (valence and arousal) and biosignal, per participant. The average accuracy and standard deviation are determined per dimension and biosignal (column), and result in the weighting to be assigned to each biosignal in the classification of each dimension.

	ECG		**EDA**		**BVP**		**Respiration**	
Participants	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal
1	0,760	0,751	0,761	0,696	0,859	0,720	0,806	0,642
3	0,718	0,745	0,627	0,638	0,607	0,679	0,571	0,487
4	0,694	0,651	0,674	0,594	0,708	0,598	0,581	0,541
5	0,871	0,800	0,893	0,697	0,822	0,566	0,869	0,702
6	0,809	0,686	0,758	0,722	0,609	0,713	0,604	0,642
8	0,716	0,850	0,690	0,613	0,645	0,710	0,582	0,709
10	0,929	0,707	0,964	0,537	0,926	0,621	0,902	0,703
11	0,839	0,789	0,680	0,866	0,732	0,668	0,605	0,761
12	0,720	0,703	0,792	0,680	0,589	0,712	0,569	0,792
13	0,960	0,976	0,828	0,883	0,951	0,944	0,894	0,885
14	0,584	0,714	0,671	0,762	0,595	0,631	0,493	0,432
15	0,640	0,684	0,756	0,641	0,754	0,658	0,578	0,590
17	0,787	0,661	0,577	0,534	0,754	0,586	0,705	0,517
18	0,631	0,804	0,662	0,653	0,646	0,680	0,527	0,559
20	0,794	0,511	0,808	0,779	0,760	0,666	0,739	0,618
21	0,732	0,654	0,455	0,628	0,594	0,697	0,725	0,514
22	0,853	0,726	0,549	0,516	0,737	0,547	0,808	0,525
23	0,812	0,827	0,699	0,782	0,669	0,589	0,890	0,648
Average accuracy	0,769	0,735	0,695	0,678	0,726	0,665	0,692	0,626
Standard deviation	0,098	0,097	0,108	0,106	0,109	0,088	0,136	0,115
Weighting	**0,27**	**0,27**	**0,24**	**0,25**	**0,25**	**0,25**	**0,24**	**0,23**

modality into the final classifier.

The resulting models, for the classification of valence and arousal, are illustrated in Figure 5.25, fusing the classification performed by each of the biosignals.

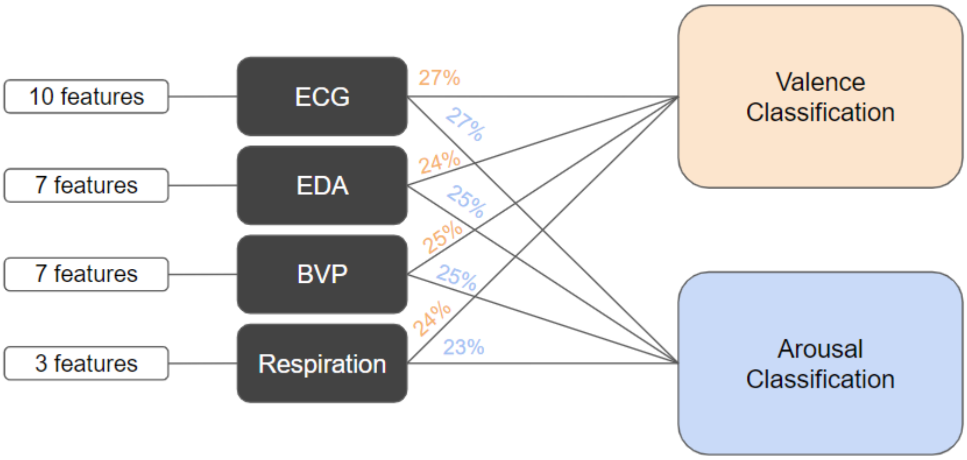


Figure 5.25: Weighting assigned to the information of each biosignal in the multimodal user-dependent classification of Valence and Arousal.

5.6.4 Emotion Recognition

Overall performance was quantified as the percentage of correctly classified emotional states per video, Figure 5.27, and per participant, Figure 5.26.

Taking into account the small amount of data available within each participant, it was decided to calculate the recognition performances accepting a penalty for some misclassified emotions. Thus, it was considered half of the performance punctuation for the classes that were ”neighbors” of the actual label expected. For instance, within a given participant analysis, each video could be rated with up to 1/7 points, and if for one given video the expected class was 3, and the classified class was 3, it would be given 1/7 points; if it was classified with the class 2 (”neighbor” of the class 3), it would be given 1/14 points; otherwise it was considered completely misclassified and thus classified with 0 points. Then the scores calculated for each of the seven videos were summed per participant and the resulting accuracies, in percentage, were obtained.

Under these assumptions, the average recognition performance was 67,68% for valence and 51,07% for arousal.

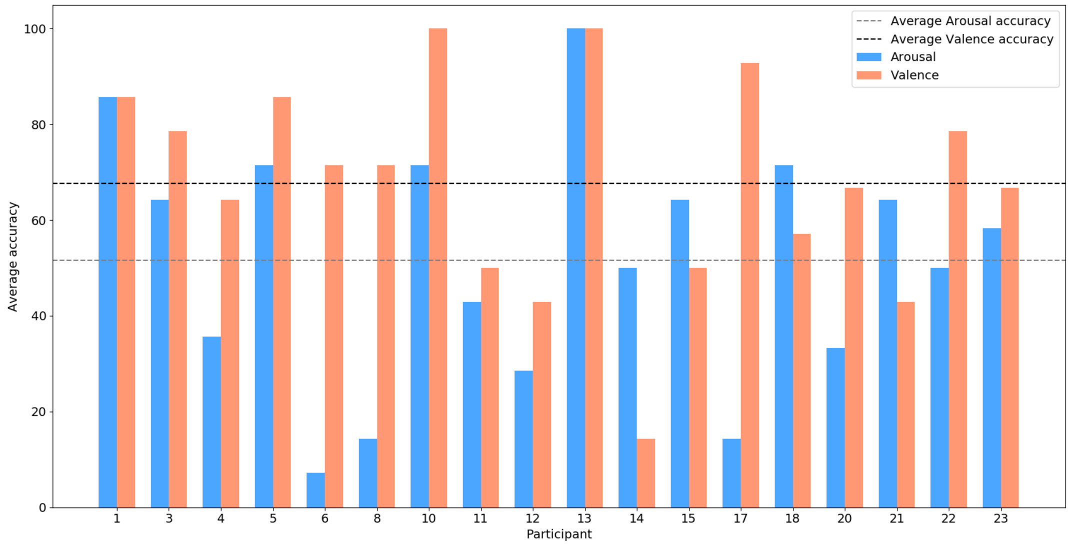


Figure 5.26: Accuracy obtained in the user-dependent approach, for each participant.

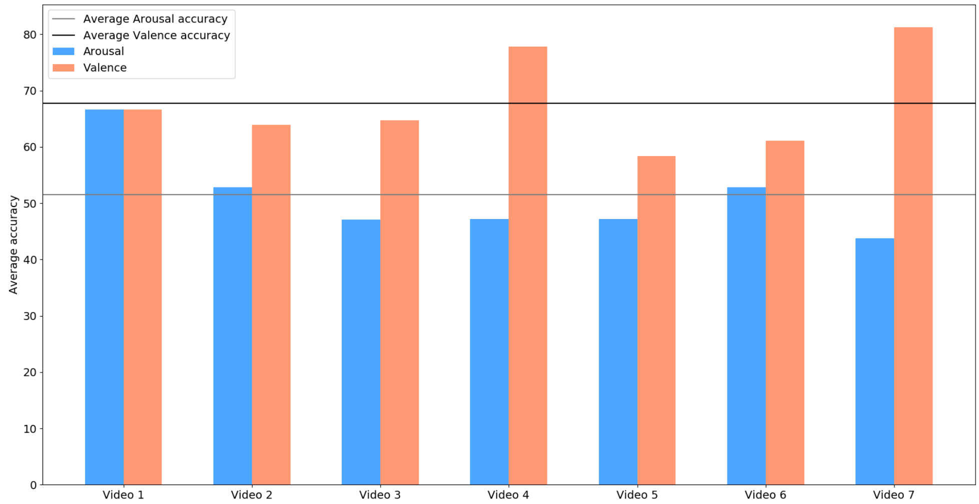


Figure 5.27: Accuracy obtained in the user-dependent approach, for each video.

5.7 User-Independent Classification

5.7.1 Non-nested vs. Nested Cross-Validation Bias Assessment

In this section the results for the non-nested CV are compared against those of the nested CV computed on the four modalities of the work, analyzing the accuracies and bias obtained for each. Since this user-independent approach uses the data of the whole population as training set, only eight models are selected in this step (i.e. one for valence and another for arousal classifications, for each of the four biosignals).

Table 5.8 shows a comparison of the 4-fold CV estimates using the non-nested and nested model selection procedures, by presenting the average difference between their scores and respective standard deviation. Unlike what has been observed in the first scenario, here the non-nested CV procedure did not yield consistently to optimistic bias, as there were five models in which the average difference was rounded to zero. In the three cases where the bias was verified, it was only significant for the model of EDA for arousal classification, as the average bias was larger than twice its standard deviation.

Concerning the kernels and hyperparameters selected for each model, one can note that all of them had the kernel RBF as optimal kernel, not varying in terms of the γ obtained, which was $\gamma = 0,01$ for all the cases. The regularization parameter C selected was either 1 or 100. One can thus infer that the models selected for the valence classification through BVP and Respiration, and for the arousal classification through respiration signals, having $C = 1$, had a softer margin and less error penalty on the training data, while the other models, having $C = 100$, had larger error penalty (noting that $C = \infty$ would lead to a hard margin).

Table 5.8: Comparison of non-nested and nested cross-validation procedures in terms of bias. The relative bias of each dataset are presented in the form of the average error rate over 30 trials, *i.e.* the mean difference between the nested and non-nested scores, along with the associated standard error.

		Average Bias	**Standard Deviation**
ECG	Valence	0,0032	0,0039
	Arousal	0,0000	0,0000
EDA	Valence	0,0007	0,0007
	Arousal	0,0061	0,0007
BVP	Valence	0,0000	0,0000
	Arousal	0,0000	0,0000
Respiration	Valence	0,0000	0,0000
	Arousal	0,0000	0,0000

5.7.2 Model Selection and Performance in each Modality

The nested CV scores are detailed in the Table 5.9, for the model selection of both valence and arousal classifications for the whole population.

Table 5.9: Accuracy scores of the models selected through nested cross-validation, for each emotion dimension (valence and arousal) and biosignal, per participant. The average accuracy and standard deviation are determined per dimension and biosignal (column), and result in the weighting to be assigned to each biosignal in the classification of each dimension.

	ECG		**EDA**		**BVP**		**Respiration**	
Population	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal
Accuracy	0,700	0,656	0,721	0,572	0,695	0,660	0,629	0,585
Weighting	**0,26**	**0,27**	**0,26**	**0,23**	**0,25**	**0,27**	**0,23**	**0,24**
Model Selected	C=100 kernel=RBF γ=0,01	C=100 kernel=RBF γ=0,01	C=100 kernel=RBF γ=0,01	C=100 kernel=RBF γ=0,01	C=1 kernel=RBF γ=0,01	C=100 kernel=RBF γ=0,01	C=1 kernel=RBF γ=0,01	C=1 kernel=RBF γ=0,01

5.7.3 Multimodal Classifier

Similarly to the previous scenario, the performance capability of each modality in the nested CV is translated into the average accuracy obtained for each biosignal, detailed in the Table 5.9, and the fusion into one multimodal classifier was done by weighting each modality accordingly.

The resulting models, for the classification of valence and arousal, are illustrated in Figure 5.28, fusing the classification performed by each of the biosignals.

5.7.4 Emotion Recognition

Overall performance was quantified as the percentage of correctly classified emotional states per video, Figure 5.30, and per participant, Figure 5.29.

The average recognition performance, considering all the videos, was 57,12% for valence and 58,11% for arousal. By excluding the videos in which the absolute difference between the average and expected

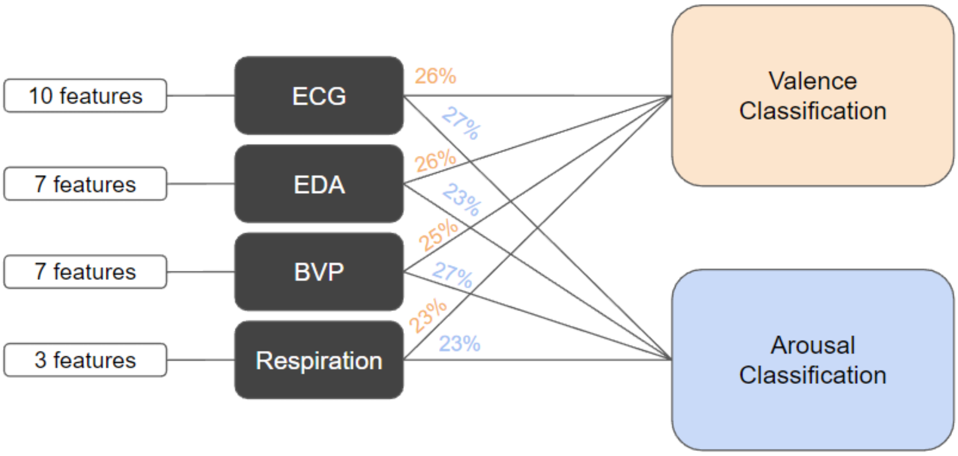


Figure 5.28: Weighting assigned to the information of each biosignal in the multimodal user-independent classification of Valence and Arousal.

ratings was larger than the standard deviation (i.e. considering only videos 4, 6 & 7 for valence, and videos 2, 4, 6 & 7 for arousal) yielded a performance of 65,74% for valence and 61,98% for arousal (cf. (**) in Figure 5.30). Excluding the two videos with worse performance in each dimension (i.e. videos 1 & 5 for valence, and videos 2 & 5 for arousal) the resulting accuracies were 67,75% for valence and 69,13% for arousal (cf. (*) in Figure 5.30).

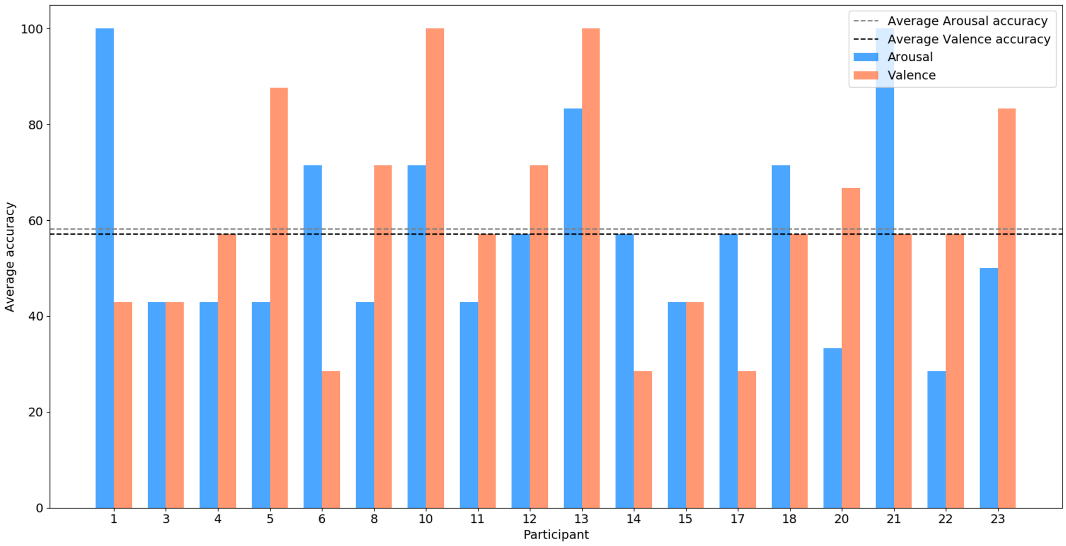


Figure 5.29: Accuracy obtained in the user-independent approach, for each participant.

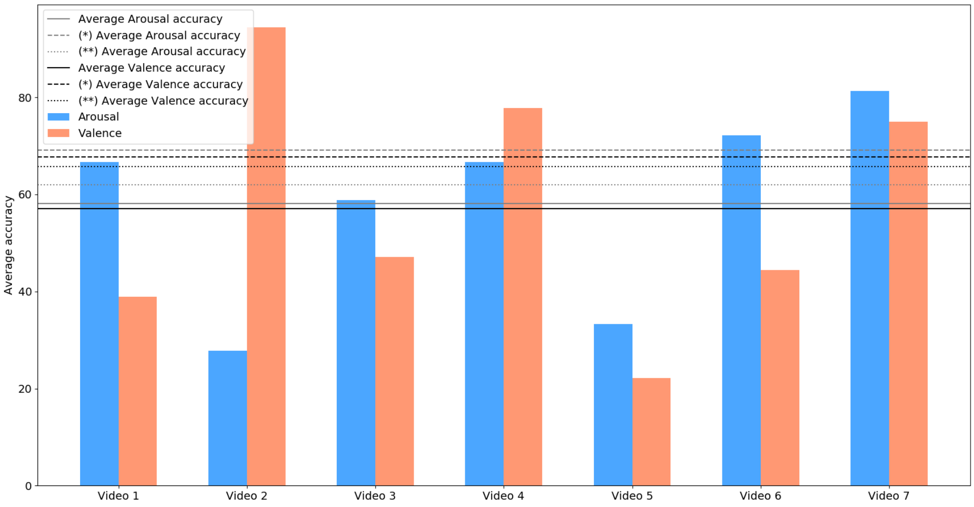


Figure 5.30: Accuracy obtained in the user-independent approach, for each video. Solid line: considering all videos; (*) excluding videos 2 & 5 in arousal, and 1 & 5 in valence; (**) considering only the videos whose absolute difference between the average and expected ratings was smaller than the standard deviation.

Chapter 6

Discussion and Future Directions

The first section of this chapter presents a discussion over the elicitation outcomes of this study, regarding the influence of the emotion baseline of the participants in the day of the experiment, and the comparison of the self-assessed emotions and the expected ones.

A discussion over the emotion recognition results is then conducted, concerning the user-dependent and user-independent scenarios. For both cases, average accuracies were shown in bar plots depicting the performance obtained per participant (cf. Figures 5.26 and 5.29) and video (cf. Figures 5.27 and 5.30).

General concernings regarding the ambiguous and subjective nature of emotion self-assessment and scaling are addressed at the end of the chapter, comprising also a discussion of several technical steps performed in this workflow, whilst possible future directions are pointed throughout the sections.

6.1 Elicitation Outcomes

6.1.1 Influence of Day's Emotion Baseline in the Elicitation Outcome

The results suggested that the emotion baseline felt by the participant during the day of the experiment influenced the elicitation outcomes. Particularly, despite the maximum correlations found were 0,443 for valence and 0,564 for arousal, this has been statistically significant for six videos in terms of arousal, and two videos in terms of valence. These results reinforce the influence of the emotional baseline of the participant that can be seen as context effect and source of bias in subjective judgement [109], for which this factor should be taken into account in future studies, as further discussed in this chapter.

6.1.2 Expected vs. Self-assessed Emotions

By visual inspection of the shapes represented in the radial plots (cf. Figures 5.22 and 5.23), one can start by noticing a more ”circular” tendency for the valence ratings, which might suggest a greater resistance to influence or elicitate emotion variation in this dimension. It is likely that a subject is more susceptible to calmness-arousal changes than along the sadness-happiness dimension. In fact, the radial shape of the arousal ratings appear to be more irregular, suggesting bigger efficacy in elicitating changes in this dimension.

Inspecting the comparison between the emotion ratings of this work and other studies, outlined in Table 5.5, one can notice that, not surprisingly, all of the mean ratings obtained in this work were relatively similar to those obtained for the videos retrieved from the database [33], presenting less than 1-point (out of 9) difference, except for Video 1, which presented both higher valence and arousal than expected. One can point out at least two reasons to explain this discrepancy: for being the first video of the protocol, even though targeting a boredom state, there could exist some excitement in the participants to initiate this VR experiment, inflating both valence and arousal; furthermore, due to its neutral nature, and once again for being the first video watched, there might have been a tendency for the participant not to decrease largely their initial valence state, taking into account that the day’s baseline of the valence was, as seen in the previous section, of 7 in the SAM 9-points scale.

Comparing the ratings obtained in this work with those conceptually expected from the literature [88], one can observe higher mean differences, as the last column of the table details. This is not largely surprising, as it is natural that even though each video targets one specific emotion, the ratings of that specific video end up being slightly different from the ones of the conceptual emotion itself, as a video might be considered more complex. The videos that most largely deviated from their conceptual target emotions were Videos 3, 5 and 6, for fear, anger and sadness, respectively. The three videos were expected to elicitate lower valence than what was verified, which suggests, again, that there was difficulty to elicitate variations in the valence dimension. Thus video 6 elicitated a melancholic state rather than a sad feeling itself, while videos 3 and 4, even though with some suspense and violent events, respectively, were not successful enough in elicitating fear and anger emotional states, both in terms of valence and arousal.

6.2 Emotion Recognition Rates

6.2.1 User-Dependent

Primarily, one can notice that the user-dependent approach yielded better recognition accuracy in terms of valence than arousal, contrarily to expected, as the majority of related work reported slightly better results for the arousal dimension [12, 31, 84]. In this scenario the uniform random accuracy of classifying three classes would be 33,33%, which was overcome. Recalling the accuracies obtained for the dataset of the work, 51,07% and 67,68%, for arousal and valence, respectively, one can suggest that the accuracy obtained for arousal is below its potential, as the system could obtain an accuracy for valence that is similar to those in the literature review [84] or slightly better [76]. One can point at least two reasons to explain it; firstly, the set of inputs for the predictor, i.e. the features extracted in this system, could have been more appropriate for valence classification than for arousal classification, as the same set of features were extracted for the classification of both dimensions. The other hypothesis, and possibly the main reason to explain this lower recognition rate obtained for arousal, is associated to the low amount of data available from each participant, which led to, in some cases, few training data to represent each of three classes and thus to enable the classifier to identify the class of the testing data correctly. It is thus believed that the existence of a larger dataset of multimodal information for each participant would yield considerably better results, with potential to enhance the overall recognition accuracy, both in valence and arousal dimensions.

In this classification, the recognition rate did not appear to vary strongly across videos, as seen in Figure 5.27, since the accuracies obtained for each emotion dimension were within a small variation range, with standard deviation of 7,62% and 8,52% for arousal and valence, respectively, considered admissible.

On the other hand, comparing the performance across participants, in Figure 5.26, it appears to present a stronger variance from one participant to another. This might have to do with what was referred by Chanel et al. [58] as the variability in the interest features, which might differ from one user to another, and are very sensitive to day to day variations as well as to the context of the emotion induction [58]. This might suggest that in a user-dependent scenario, relying only on the data from one individual, it could be advantageous to automatically select the subset from the whole available features that is more appropriate for the physiological responses of that participant.

Moreover, it might be natural to assume a parallelism between the problems of emotion patterns classification and the biometric patterns classification, where there has been evidence that different users have different recognition performances within a system [110]. Yager and Dunstone [110] proposed a new framework for the evaluation of biometric systems based on the biometric menagerie, as opposed to collective statistics, taking into account the existence of various animal groups which differ in terms of behavior or recognition performance. In the case of biometrics, the assigning of a user to each menagerie group as to do with a number of factors, including enrollment procedures, feature extraction and matching algorithms, data quality, and intrinsic properties of the user population [110], which are relatable to the process involved in an emotion system. This way, it is admissible to infer the presence of some users with responses that have complex characteristics, explaining the existence of participants with considerably worse performances (namely participants 6, 8, 12 and 17 for arousal, and participants 12 and 14 for valence), which worsened the overall recognition rates obtained. One can also notice that the users with the lowest performance capabilities did not necessarily perform badly for both emotion dimensions, rather it was more common for arousal. Such fact might be related to either different capabilities of self-assessing the two emotion dimensions, as the error might be caused by wrong labels in the first place (this topic will be further discussed hereafter), or to more complex physiologic-related characteristics that make certain users more difficult to automatically assess by the system [31, 58]. Once again, it is believed that the availability of a larger dataset of multimodal data for each of the participants would significantly enhance the performance of the system and mitigate the strong variability amongst individuals that is seen here. In fact, Soleymani et al. [111] claim that 20 videos, even though with shorter durations than those used in this work, would not be a number of samples enough to perform user-dependent emotion recognition of three classes per dimension.

6.2.2 User-Independent

The recognition rates obtained in the user-independent approach can be considered in more conformity with the state-of-the-art [30, 31], possibly due to the larger amount of training data to train the classifier. Moreover, the accuracy obtained for arousal was slightly better than for valence, as expected [12, 31, 84], and the uniform random accuracy of classifying two classes (50,00%) was overcome.

The results across videos have shown a noteworthy variability, which suggested a further analysis to understand that discrepancy and calculate the accuracies obtained when certain videos were excluded. As opposite to what was seen in the previous classification scenario, here some videos definitely yielded worse recognition accuracies, namely videos 2 & 5 for arousal, and videos 1 & 5 for valence. Several hypothesis could explain this underperformance, namely the complex and subjective particularities of self-assessing emotion through scales, and the own emotion elicitation capabilities of some videos might be questionable, leading to possible ambiguity in the emotion felt by the participant. To enlighten the latter topic, one can analyze Figure 5.24, assessing the possibility of any relation between larger intra-video variability or larger discrepancy between the obtained and the expected ratings, and the resulting worse recognition accuracies for the videos mentioned. Firstly, the arousal ratings of video 2 definitely presented a large variability around its average, which might mean that the arousal elicitated by the video was unclear and thus different participants rated very differently the arousal of the same emotion, indicative of an ambiguous perception of the content across participants and thus the reliability of the ratings retrieved might be questionable. In this case, the discrepancy between the emotions classified by the system are not necessarily wrong, rather discrepant from the self-assessed ones. Considering the poor results obtained for the arousal classification in video 5, and valence classification in videos 1 and 5, one can observe that they did not present a large variability as in the previous case, but all presented a considerable difference between the average and the expected ratings, larger than the own standard deviation values. Although not fully conclusive, one might suggest that the performance assessment may be negatively biased by the difficulty of the participants to rate their real emotions, cultural factors, desensitizing or other variability sources that make the content perception differ from what was expected, and thus yielding false-wrong classifications. Moreover, one should notice that the best recognition rate was accomplished for the valence classification of video 2, which was the one with the smallest standard deviation from all, somehow corroborating this analysis. Hence it was decided to compute the recognition rate excluding videos 2 & 5 in arousal, and videos 1 & 5 in valence, which yielded more encouraging results, 69,13% for arousal and 67,75% for valence ((*) in Figure 5.30). Finally, for a systematic point of view, the recognition rate was also assessed when excluding all the videos in which the absolute difference between the average and expected ratings was larger than the standard deviation, thus considering only videos 4, 6 and 7 for valence, and videos 2, 4, 6 and 7 for arousal, which yielded the results of 61,98% for arousal and 65,74% for valence ((**) in Figure 5.30). The results were, not surprisingly, worse than in the previous analysis, as some of the best recognition rates were as well excluded.

6.3 General Considerations

The ambiguous and subjective nature of emotion self-assessment, as well as of the scaling process, are a general concern one must into account in the scope of emotion recognition and in the analysis of the results of the present work. Citing Alvarado [112], ”subjectivity is inescapable in science as in everyday life”. In making use of subjective response data, it matters to analyse how the conditions of data collection and its perceived purpose may influence the interpretation of the task and the related subjective experience [112], identifying the most common sources of emotion bias, and working towards self-assessment scales able to mitigate them, both intra- and inter- individuals (e.g. in judgements of internal sensations, the use of descriptors whose meaning is clear and shared amongst different individuals is valuable [109]).

Annett [109] addressed some of the fundamental problems concerning the use of subjective evidence in the practice of ergonomics. Namely, the problem of subjectivity, questions about the nature of scale and whether subjective data comply with the logical requirements of true scientific measurement [109], in this case for the purpose of training a machine learning algorithm for emotion recognition. Dissociations between subjective and objective measures are referred as often, being caused by various sources. For instance, context effects have been widely recognized as a common source of bias in subjective judgement [109], for which the analysis of the impact of the emotional baseline of each participant, in the day of the experiment, took place in the previous chapter.

In fact, whereas the quantitative nature of some physically measurable attribute, e.g. length, is almost intuitively obvious, the appropriate scaling structure in psychological attributes is less clear. This way, attributes such as intelligence and emotion, although assigned values on a numerical scale, lack an identifiable, completely objective, unit of measurement [112]. Furthermore, specifically for self-assessment of emotions, Alvarado [112] claims that, even though there is evidence that justifies the assumption of an ordinal scale type during data analysis of the emotional response of an individual, there is no evidence that the subjective distances between adjacent numbers on every portion of the scale are equal [112]. Thus aggregation of data and comparisons amongst the ratings of a population of individuals are problematic because it is unclear how individual differences in emotional response are related to individual differences in the use of rating scales, as well as the distances between numbers have not been shown to correspond to the same subjective differences in response for each individual in a study [112]

Regarding the multimodal fusion of the information of the four biosignals, while this work followed an approach where each modality yielded a classification process and only then a multimodal classifier considered the four classifications and weighted them in order to retrieve a global classification. Another possibility would be gathering all the features from the four biosignals and performing the classification only once. It was chosen to compute the former as a means to respect the different durations of the templates of each biosignal, and to avoid losing information when generalizing the features in a fixed time duration, as it would be required to average the several features for the number of samples existent in each biosignal in that given duration. Moreover, that would result in a large number of features that could lead to the problem known as the ”curse of dimensionality”, a term introduced by Bellman [113] to describe the problem caused by the exponential increase in volume associated with adding extra dimensions to the Euclidean space, having that in high dimensions all feasible training samples sparsely populate the input space [56]. Nonetheless, it could also be a good approach to further explore.

One should also address the downscaling process that was performed in this work to convert the 1-9 ratings into 1-3 and 1-2 ratings, for the user-dependent and user-independent scenarios, respectively. On one hand, it is legitimate to identify the drawbacks of this approach; considering the above-mentioned problems concerning self-assessment through scales, the computing of this linear scaling might have yielded some conceptual errors with respect to the subjective interpretation of those numbers by the participant, specially for the emotions rated with the numbers in the ”borders” of each class (e.g. depending on the subjective interpretation of the scale by the participant, a 6 in the valence SAM could represent a positive emotion, whereas through this linear scaling it was considered as neutral, in the user-dependent case; for another participant, a 7 in the valence SAM could represent a neutral emotion, whereas in this case it would have been considered as positive). On the other hand, the usage of the 9-points scale was advantageous in terms of conformity and comparability with the ratings in another studies, as it is the most commonly used scale. Future work should perhaps collect, besides the 9-points self-assessments, ratings in scales with the number of points of classes one aims to classify, in order to avoid that subjective scaling and consecutive generalization issues.

6.4 Practical Guidelines for Future Work

Further guidelines regarding the extension or acquisition of a larger database include keeping in mind the tradeoff between collecting more samples from each participant, while not having long sessions that would make them tired and unable to feel the target emotions [111]. Thus, it would be advisable to conduct an experiment with several sessions, in order to collect a considerable amount of intra-participant data, while the duration of each session should remain relatively short (around 30 minutes of physiological data acquisition).

Regarding the duration of the immersive video clips itself, it should be, as selected in this work, as similar as possible between the several videos and relatively short, i.e. around 1 minute to avoid elicitation of multiple emotions while allowing the physiological occurrence. The videos selected should be able to elicit one and only one emotion.

As stated, future work might collect both 9-points self-assessments and ratings in scales with the same number of points as the number of classes one aims to classify, in order to avoid subjective scaling and related issues associated with the scaling of the ratings.

To ensure that the participant gains the right knowledge for filling the SAM scales accurately, it is advisable to specifically enlighten the meaning of the parameters assessed in each scale (e.g. valence, arousal) during the preparation procedure of the experiment.

Finally, Soleymani et al. [111], considering the nature of affective experiments, refer the importance of motivating the participants, stating rewards as an effective tool.

Chapter 7

Conclusions

The results suggested that the emotion baseline felt by the participant during the day of the experiment influenced the elicitation outcomes, which appeared to be more significant in terms of arousal. It should thus be seen as a relevant context effect and source of bias in subjective judgement in future studies. Moreover, the self-assessments suggested a larger resistance to elicitate emotion variation in the valence dimension, i.e. a subject is more susceptible to calmness-arousal variations.

The emotion ratings obtained in this work were compared with two related studies. As expected, the valence and arousal ratings were very similar to those obtained for the videos retrieved from the database [33], presenting less than 1-point (in the 9-point SAM scale) difference, except for Video 1. Larger mean differences were observed when comparing with the conceptual values of the emotions found in [88].

Overall, the results for the user-dependent scenario have shown a underperformance with respect to related work, which is explained by the small database used in this study, whilst the results for the user-independent approach are in conformity with the state-of-the-art, and can be considered promising.

As argued, the appropriate scaling structure in emotion is not fully understood, being unclear how individual differences in emotional response are related to self-rating differences. Future work will focus on increasing the database size and considering the subjective factors involved in emotion interpretation.

7.1 Achievements

The present study fulfilled the objectives drawn at the beginning of the work. The results were in accordance with the state-of-the-art, and some new information was introduced in the field. Two major achievements were the acceptance of an abstract at the 6th IEEE ENBENG 2019, and of a full contributed paper at the 41st IEEE EMBC 2019.

Bibliography

[1] M. Paleari, R. Benmokhtar, and B. Huet. Evidence theory-based multimodal emotion recognition. In Benoit Huet, Alan Smeaton, Ketan Mayer-Patel, and Yannis Avrithis, editors, *Advances in Multimedia Modeling*, pages 435–446, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[2] S. Thushara and S. Veni. A multimodal emotion recognition system from video. pages 1–5, March 2016.

[3] G. Rigas, C. D. Katsis, G. Ganiatsas, and D. I. Fotiadis. A user independent, biosignal based, emotion recognition method. In Cristina Conati, Kathleen McCoy, and Georgios Paliouras, editors, *User Modeling 2007*, pages 314–318, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[4] H. P. Martinez, Y. Bengio, and G. N. Yannakakis. Learning deep physiological models of affect. *IEEE Computational Intelligence Magazine*, 8(2):20–33, May 2013.

[5] P. Ekman. Basic emotions. In Tim Dalgleish and M. J. Powers, editors, *Handbook of Cognition and Emotion*, pages 4–5. Wiley, 1999.

[6] P. J. Lang. The emotion probe: Studies of motivation and attention. *The American psychologist.*, 50:372–85, 1995.

[7] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan. Physiological signals based human emotion recognition: a review. In *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*, pages 410–415, March 2011.

[8] H. P. Martinez. *Advancing Affect Modeling via Preference Learning and Unsupervised Feature Extraction*. PhD thesis, Denmark, 2013.

[9] H. A. Osman and T. H. Falk. Multimodal affect recognition: Current approaches and challenges. In Seyyed Abed Hosseini, editor, *Emotion and Attention Recognition Based on Biological Signals and Images*, chapter 5. IntechOpen, Rijeka, 2017.

[10] Q. Li, Z. Yang, S. Liu, Z. Dai, and Y. Liu. The study of emotion recognition from physiological signals. In *2015 Seventh International Conference on Advanced Computational Intelligence (ICACI)*, pages 378–382, March 2015.

[11] H. Silva, A. Fred, S. Eusebio, M. Torrado, and S. Ouakinin. Feature Extraction for Psychophysiological Load Assessment in Unconstrained Scenarios. In *2012 ANNUAL INTERNATIONAL CONFERENCE OF THE IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY (EMBC)*, volume 2012 of *IEEE Engineering in Medicine and Biology Society Conference Proceedings*, pages 4784–4787, 2012. Citations: scopus, wos.

[12] A. Haag, S. Goronzy, P. Schaich, and J. Williams. Emotion recognition using bio-sensors: First steps towards an automatic system. In Elisabeth André, Laila Dybkjær, Wolfgang Minker, and Paul Heisterkamp, editors, *Affective Dialogue Systems*, pages 36–48, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[13] H. P. Silva, C. Carreiras, A. Lourenco, A. Fred, R. Neves, and R. Ferreira. Off-the-person electrocardiography: performance assessment and clinical correlation. *Health and Technology*, 4, 04 2015.

[14] A. Alhargan, N. Cooke, and T. Binjammaz. Multimodal affect recognition in an interactive gaming environment using eye tracking and speech signals. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI 2017, pages 479–486, New York, NY, USA, 2017. ACM.

[15] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari. Audio-visual emotion recognition in video clips. *IEEE Transactions on Affective Computing*, pages 1–1, 2017.

[16] J. Chen, B. Hu, L. Xu, P. Moore, and Y. Su. Feature-level fusion of multimodal physiological signals for emotion recognition. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 395–399, Nov 2015.

[17] S. Zhalehpour, Z. Akhtar, and C. Eroglu Erdem. Multimodal emotion recognition with automatic peak frame selection. In *2014 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA) Proceedings*, pages 116–121, June 2014.

[18] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schüller. Enhanced semi-supervised learning for multimodal emotion recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5185–5189, March 2016.

[19] Abuhashish F., Zraqou J., Alkhodour W., Sunar M., and Kolivand H. Emotion interaction with virtual reality using hybrid emotion classification technique toward brain signals. *International Journal of Computer Science Information Technology (IJCSIT)*, 7, 2015.

[20] C. Carreiras, A. Alves, A. Lourenço, F. Canento, H. Silva, A. Fred, et al. BioSPPy: Biosignal processing in Python, 2015–. [Online; accessed ¡today¿].

[21] A. Kelava, M. Muma, M. Deja, J. Dagdagan, and A. M. Zoubir. A new approach for the quantification of synchrony of multivariate non-stationary psychophysiological variables during emotion eliciting stimuli. In *Front. Psychol.*, 2014.

[22] L. F. Barrett and E. Bliss-Moreau. Affect as a psychological primitive. *Advances in experimental social psychology*, 41:167–218, 2009.

[23] T. Eerola and J. K. Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49, 2011.

[24] E. Harmon-Jones, C. Harmon-Jones, and E. Summerell. On the importance of both dimensional and discrete models of emotion. *Behavioral Sciences*, 7(4), 2017.

[25] J. Kim and E. Andre. Emotion recognition based on physiological changes in music listening. *IEEE transactions on pattern analysis and machine intelligence*, 30:2067–83, 01 2009.

[26] U. Schimmack and A. Grob. Dimensional models of core affect: A quantitative comparison by means of structural equation modeling. *European Journal of Personality*, 14, 07 2000.

[27] J. Fontaine, K. Scherer, E. Roesch, and P. Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18:1050–7, 01 2008.

[28] H. A. Elfenbein and N. Ambady. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2):203—235, March 2002.

[29] J. Lorenz C. Arnrich B. Tröster G. Setz, C. Schumm. Combining worthless sensor data.

[30] S. Bang K. Kim and S. Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, 42:419–427, 2004.

[31] J. Wagner, J. Kim, and E. Andre. From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In *2005 IEEE International Conference on Multimedia and Expo*, pages 940–943, July 2005.

[32] M. Ali, A. H. Mosa, F. A. Machot, and K. Kyamakya. *Emotion Recognition Involving Physiological and Speech Signals: A Comprehensive Review*, pages 287–302. Springer International Publishing, Cham, 2018.

[33] B. J. Li, J. N. Bailenson, A. Pines, W. J. Greenleaf, and L. M. Williams. A public database of immersive vr videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures. *Frontiers in Psychology*, 8:2116, 2017.

[34] E. H. Jang, B. J. Park, S. H. Kim, Y. Eum, and J. H. Sohn. Identification of the optimal emotion recognition algorithm using physiological signals. In *2011 2nd International Conference on Engineering and Industries (ICEI)*, pages 1–6, Nov 2011.

[35] Y. Dai, X. Wang, X. Li, and P. Zhang. Reputation-driven multimodal emotion recognition in wearable biosensor network. In *2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings*, pages 1747–1752, May 2015.

[36] D. E Becker. Fundamentals of electrocardiography interpretation. *Anesthesia progress.*, 53:53–63, 2006.

[37] Edward J. Ciaccio. Biomedical signal and image processing, second edition, review of biomedical signal and image processing, crc press, taylor & francis group, boca raton, review by edward j. ciaccio,. *BioMedical Engineering OnLine*, 12(1):88, Sep 2013.

[38] Y. Chu, X. Zhao, J. Han, and Y. Su. Physiological signal-based method for measurement of pain intensity. *Frontiers in Neuroscience*, 11:279, 2017.

[39] J. Malmivuo and R. Plonsey. *Bioelectromagnetism. 6. The Heart*, pages 119–130. 01 1995.

[40] R. Priya Muthusamy. Seminar paper: Emotion recognition from physiological signals using biosensors. 05 2018.

[41] A. Greco, G. Valenza, and E.P. Scilingo. *Advances in Electrodermal Activity Processing with Applications for Mental Health: From Heuristic Methods to Convex Optimization*. Springer International Publishing, 2016.

[42] E. Jang, B. Park, M. Park, S. Kim, and J. Sohn. Analysis of physiological signals for recognition of boredom, pain, and surprise emotions. *Journal of Physiological Anthropology*, 34(1):25, Jun 2015.

[43] H. Silva, A. L. N. Fred, and A. Lourenço. Electrodermal response propagation time as a potential psychophysiological marker. In *International Conf. of the IEEE Engineering in Medicine and Biology Society - EMBC*, pages 6756–6760, August 2012.

[44] A. Greco, G. Valenza, and E. Scilingo. Modeling for the analysis of the eda, 11 2016.

[45] plux — wireless biosignas. Blood Volume Pulse (BVP): Sensors, 2015–. [Online; accessed ¡21/10/2018¿].

[46] A. Kushki, J. Fairley, S. Merja, G. King, and T. Chau. Comparison of blood volume pulse and skin conductance responses to mental and affective stimuli at different anatomical sites. *Physiological Measurement*, 32(10):1529–1539, aug 2011.

[47] C. Chang and C. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[48] A. Shmilovici. *Support Vector Machines*, pages 257–276. Springer US, Boston, MA, 2005.

[49] V. N. Vapnik. An overview of statistical learning theory. *Trans. Neur. Netw.*, 10(5):988–999, September 1999.

[50] A. Y. Chervonenkis. *Early History of Support Vector Machines*, pages 13–20. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[51] M. Pal. Multiclass approaches for support vector machine based land cover classification. *CoRR*, abs/0802.2411, 2008.

[52] C. J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, Jun 1998.

[53] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152. ACM Press, 1992.

[54] M. Schels, M. Glodek, S. Meudt, S. Scherer, M. Schmidt, G. Layher, S. Tschechne, T. Brosch, D. Hrabal, S. Walter, H. Traue, G. Palm, H. Neumann, and F. Schwenker. *Multi-Modal Classifier-Fusion for the Recognition of Emotions*, pages 73–97. 10 2013.

[55] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995.

[56] Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. *Ann. Statist.*, 26(2):451–471, 04 1998.

[57] C. Hsu, C. Chang, and C. Lin. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2003.

[58] G. Chanel, J. J.M. Kierkels, M. Soleymani, and T. Pun. Short-term emotion assessment in a recall paradigm. *International Journal of Human-Computer Studies*, 67(8):607 – 627, 2009.

[59] Bradley M.M. Lang, P.J. and B.N. Cuthbert. International affective picture system (iaps): Affective ratings of pictures and instruction manual. technical report a-8. *University of Florida, Gainesville, Fl.*, 2008.

[60] P. J. Lang. Behavioral treatment and bio-behavioral assessment: Computer applications. In J. B. Sidowski, J. H. Johnson, and T. A. Williams, editors, *Technology in mental health care delivery systems*, pages 119 – 137. Norwood, NJ: Ablex, 1980.

[61] M. M. Bradley and P. J. Lang. The international affective digitized sounds (2nd edition; iads-2): Affective ratings of sounds and instruction manual. technical report b-3. *University of Florida, Gainesville, Fl.*, 2007.

[62] J. Selvaraj, M. Murugappan, K. Wan, and S. Yaacob. Classification of emotional states from electrocardiogram signals: a non-linear approach based on hurst. *BioMedical Engineering OnLine*, 12(1):44, May 2013.

[63] L. Li and J. Chen. Emotion recognition using physiological signals from multiple subjects. In *2006 International Conference on Intelligent Information Hiding and Multimedia*, pages 355–358, Dec 2006.

[64] G. Riva, F. Mantovani, C. S. Capideville, A. Preziosa, F. Morganti, D. Villani, A. Gaggioli, C. Botella, and M. Alcañiz. Affective interactions using virtual reality: The link between presence and emotions. *CyberPsychology & Behavior*, 10(1):45–56, 2007. PMID: 17305448.

[65] J. Diemer, G. W. Alpers, H. M. Peperkorn, Y. Shiban, and A. Mühlberger. The impact of perception and presence on emotional reactions: a review of research in virtual reality. In *Front. Psychol.*, 2015.

[66] J. Blascovich, J. Loomis, A. C. Beall, K. R. Swinth, C. L. Hoyt, and J. N. Bailenson. Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry*, 13(2):103–124, 2002.

[67] G Riva and J A Waterworth. Presence and the self: a cognitive neuroscience approach. *Presence-Connect*, 3, 2003.

[68] C. Coelho, J. Tichon, T. J. Hine, G. Wallis, and G. Riva. Media presence and inner presence : The sense of presence in virtual reality technologies. 2006.

[69] S. A. Hosseini. Classification of brain activity in emotional states using hos analysis. 4, 02 2012.

[70] H. Irtel. The pxlab self-assessment-manikin scales.

[71] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *Trans. Evol. Comp*, 1(1):67–82, April 1997.

[72] M. Fernandez-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? 15:3133–3181, 10 2014.

[73] E. Jang, B. Park, S. Kim, and Jin-Hun Sohn. Emotion classification by machine learning algorithm using physiological signals. 2012.

[74] C. Liu, P. Rani, and N. Sarkar. An empirical study of machine learning techniques for affect recognition in human-robot interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2662–2667, Aug 2005.

[75] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, Mar 1998.

[76] M. B. H. Wiem and Z. Lachiri. Emotion assessing using valence-arousal evaluation based on peripheral physiological signals and support vector machine. In *2016 4th International Conference on Control Engineering Information Technology (CEIT)*, pages 1–5, Dec 2016.

[77] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[78] Af. Eduardo, H. Aidos, and A. Fred. Ecg-based biometrics using a deep autoencoder for feature learning - an empirical study on transferability, 01 2017.

[79] B. Forster-Heinlein and P. Massopust. *Four short courses on harmonic analysis. Wavelets, frames, time-frequency methods, and applications to signal and image analysis*. 11 2009.

[80] D. Batista and A. Fred. Spectral and time domain parameters for the classification of atrial fibrillation, 2015.

[81] W. Rosenberg, T. Chanwimalueang, T. Adjei, U. Jaffer, V. Goverdovsky, and D. P. Mandic. Resolving ambiguities in the lf/hf ratio: Lf-hf scatter plots for the categorization of mental and physical stress from hrv. In *Front. Physiol.*, 2017.

[82] M. Murugappan, S. Murugappan, and B. S. Zheng. Frequency band analysis of electrocardiogram (ecg) signals for human emotional state classification using discrete wavelet transform (dwt). In *Journal of physical therapy science*, 2013.

[83] R. Gupta, M. K. Abadi, J. A. Cárdenes Cabré, F. Morreale, T. H. Falk, and N. Sebe. A quality adaptive multimodal affect recognition system for user-centric multimedia indexing. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ICMR '16, pages 317–320, New York, NY, USA, 2016. ACM.

[84] D. Kulic and E. A. Croft. Affective state estimation for human–robot interaction. *IEEE Transactions on Robotics*, 23(5):991–1000, Oct 2007.

[85] H. P. da Silva, A. Fred, and R. Martins. Biosignals for everyone. *IEEE Pervasive Computing*, 13(4):64–71, Oct 2014.

[86] D. Batista, H. P. Silva, A. Fred, C. Moreira, M. Reis, and H. Ferreira. Benchmarking of the bitalino biomedical toolkit against an established gold standard. *Healthcare Technology Letters*.

[87] Hp windows mixed reality headset with motion controllers. `https://www.microsoft.com/en-us/p/hp-windows-mixed-reality-headset-with-motion-controllers/8n5g0j1qf74b/5jp3?cid=msft_web_collection&activetab=pivot:overviewtab`. Accessed: 2018-10-19.

[88] R. Hepach, D. Kliemann, S. Grüneisen, H. Heekeren, and I. Dziobek. Conceptualizing emotions along the dimensions of valence, arousal, and communicative frequency – implications for social-cognitive tests and training tools. 2:266, 10 2011.

[89] M. Kelsey, R. V. Palumbo, A. Urbaneja, M. Akçakaya, J. Huang, I. R. Kleckner, L. F. Barrett, K. S. Quigley, E. Sejdic, and M. S. Goodwin. Artifact detection in electrodermal activity using sparse recovery. 2017.

[90] F. R. Romero, G. R. Haddad, H. Amante Miot, and D. C. Cataneo. Palmar hyperhidrosis: clinical, pathophysiological, diagnostic and therapeutic aspects*. In *Anais brasileiros de dermatologia*, 2016.

[91] S. Sharma and R. P. Narwaria. Performance evaluation of various window techniques for noise cancellation from ecg signal. *International Journal of Computer Applications*, 93:1–5, 05 2014.

[92] R. Uplane M. Chavan, M. Agarwala. Suppression of baseline wander and power line interference in ecg using digital iir filter. *International Journal of Circuits, Systems and Signal Processing.*, 2, 2008.

[93] F. Gustafsson. Determining the initial states in forward-backward filtering. *IEEE Transactions on Signal Processing*, 44(4):988–992, April 1996.

[94] S. Luo, J. Zhou, H. B. Duh, and F. Chen. Bvp feature signal analysis for intelligent user interface. In *CHI Extended Abstracts*, 2017.

[95] S. Ouakinin M. Santos, A. Fred. Biometrical and psychophysiological assessment through biosensors, 2012.

[96] Kemalasari and P. S. Wardana. Processing of respiration signals using fir filter for analyze the condition of lung. In *2017 International Electronics Symposium on Engineering Technology and Applications (IES-ETA)*, pages 229–233, Sept 2017.

[97] P. Hamilton. Open source ecg analysis. In *Computers in Cardiology*, pages 101–104, Sep. 2002.

[98] A. Lourenço, H. Silva, C. Carreiras, and A. Fred. Outlier detection in non-intrusive ecg biometric system. In Mohamed Kamel and Aurélio Campilho, editors, *Image Analysis and Recognition*, pages 43–52, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[99] W. Zong, T. Heldt, G. B. Moody, and R. G. Mark. An open-source algorithm to detect onset of arterial blood pressure pulses. In *Computers in Cardiology, 2003*, pages 259–262, Sep. 2003.

[100] Md. M. Rahman, A. A. Ali, K. Plarre, M. al'Absi, E. Ertin, and S. Kumar. mconverse: inferring conversation episodes from respiratory measurements collected in the field. In *Wireless Health*, 2011.

[101] D. Altman and J. M. Bland. Statistics notes: Quartiles, quintiles, centiles, and other quantiles. *BMJ*, 309:996 – 996, 10 1994.

[102] G. Yuan, C. Ho, and C. Lin. Recent advances of large-scale linear classification. *Proceedings of the IEEE*, 100(9):2584–2603, Sept 2012.

[103] G. C. Cawley and N. L.C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.*, 11:2079–2107, August 2010.

[104] J. D. Rodriguez, A. Perez, and J. A. Lozano. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):569–575, March 2010.

[105] J. Wainer and G. Cawley. Nested cross-validation when selecting classifiers is overzealous for most practical applications, 09 2018.

[106] J. Wainer. Comparison of 14 different families of classification algorithms on 115 binary datasets, 2016.

[107] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and

E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[108] D. Grühn and S. Scheibe. Age-related differences in valence and arousal ratings of pictures from the international affective picture system (iaps): Do ratings become more extreme with age? *Behavior Research Methods*, 40(2):512–521, May 2008.

[109] J. Annett. Subjective rating scales: science or art? *Ergonomics*, 45(14):966–987, 2002. PMID: 12569049.

[110] N. Yager and T. Dunstone. The biometric menagerie. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(2):220–230, February 2010.

[111] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Trans. Affect. Comput.*, 3(1):42–55, January 2012.

[112] N. Alvarado. Arousal and valence in the direct scaling of emotional response to film clips. *Motivation and Emotion*, 21(4):323–348, Dec 1997.

[113] R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edition, 1957.

Appendix A

Abstract accepted to ENBENG 2019

EXPLORING PHYSIOLOGICAL MULTIMODALITY FOR EMOTIONAL RECOGNITION THROUGH VIRTUAL REALITY ELICITATION

Pinto, Joana[1]; Fred, Ana[2]; Silva, Hugo[2]

[1] Instituto Superior Técnico, Portugal, joana.f.pinto@tecnico.ulisboa.pt
[2] Instituto de Telecomunicações and Instituto Superior Técnico, Portugal, {afred, hsilva}@lx.it.pt

KEYWORDS: *Emotion Recognition, Biosignals, Virtual Reality*

ABSTRACT: The mental response of emotion combines subjective feeling and cognitive processes expressed by both motor and physiological responses.

Various models have been proposed to theorize emotion. The most common are the discrete model proposed by Ekman and the two dimensional model proposed by Lang, comprising valence (pleasantness) and arousal (activation level). Moreover, it is key to effectively design methodologies able to collect and label emotion. Previous works have a wide range of emotion elicitation tools, including pictures, movie clips, music, sounds, or recall paradigm. Data sources range from EEG, brain imaging, and multimodal physiological signals [1], to facial images. Regarding the machine learning algorithms used for emotion recognition purposes, both supervised and non-supervised techniques have been proposed.

In this work we explore several biosignals for emotion assessment, in an immersive Virtual Reality (VR) elicitation setup. The experimental protocol was applied on 23 participants (18 to 40 years-old). VR videos were used to elicit reliable emotions, through their immersive experiences [2]. The response of four physiological signals was collected by two modules built upon BITalino devices, one placed on the wrist (EDA & BVP), and the other on the chest (ECG & Respiration). Participants reported their emotional state of the day (baseline), by filling the Self-Assessment Manikin (SAM) 9-point scale, for valence and arousal; data from 11 tasks was collected, comprising four calibrations and seven VR videos, targeting different emotions. The participants filled the SAM scales with respect to the emotion of each video. Physiological and statistic attributes were used to validate, process, and extract features from the signals. Emotion assessment was performed with SVM, as it is considered one of the most promising classifiers in the field [3].

The emotion baseline of the day was observed to impact the elicitation success, particularly in terms of arousal. The average ratings per video were compared against those expected in the literature (cf. Figure 1). Ongoing work explores the fusion of the multimodality information, both at the feature and the classifier levels, using population- and individual-based approaches.

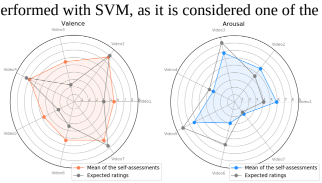


Figure 1 - Comparison between the average self-assessed emotions for each video and the respective expected ratings, for valence (left) and arousal (right).

REFERENCES

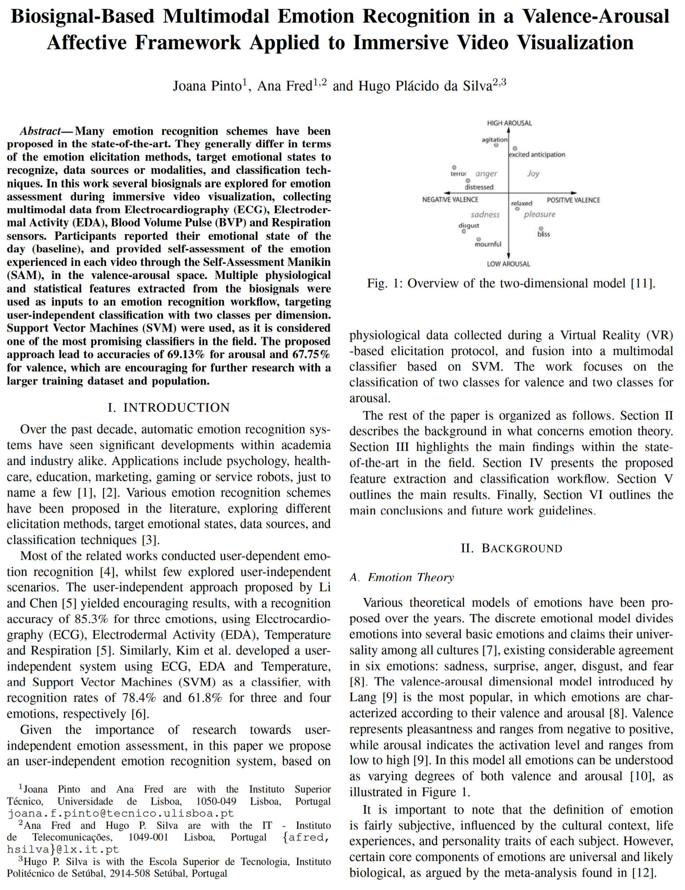
[1] Jerritta S, Murugappan M, Nagarajan R, Wan K. Physiological signals based human emotion recognition: a review. IEEE 7th International Colloquium on Signal Processing and Its Applications. 2011; 410-415.

[2] Riva G, Mantovani F, Capideville C, Preziosa A, Morganti F, Villani D, Gaggioli A, Botella C, Alcaniz M. Affective interactions using virtual reality: The link between presence and emotions. CyberPsychology & Behavior. 2007; 10(1):45–56.

[3] Jang E, Park B, Kim S, Sohn J.Emotion Classification by Machine Learning Algorithm using Physiological Signals. Proceedings of Computer Science and Information Technology, Singapore. 2018; 25: 1-5.

Appendix B

Title Page of the Paper accepted to EMBC 2019



Biosignal-Based Multimodal Emotion Recognition in a Valence-Arousal Affective Framework Applied to Immersive Video Visualization

Joana Pinto[1], Ana Fred[1,2] and Hugo Plácido da Silva[2,3]

***Abstract*— Many emotion recognition schemes have been proposed in the state-of-the-art. They generally differ in terms of the emotion elicitation methods, target emotional states to recognize, data sources or modalities, and classification techniques. In this work several biosignals are explored for emotion assessment during immersive video visualization, collecting multimodal data from Electrocardiography (ECG), Electrodermal Activity (EDA), Blood Volume Pulse (BVP) and Respiration sensors. Participants reported their emotional state of the day (baseline), and provided self-assessment of the emotion experienced in each video through the Self-Assessment Manikin (SAM), in the valence-arousal space. Multiple physiological and statistical features extracted from the biosignals were used as inputs to an emotion recognition workflow, targeting user-independent classification with two classes per dimension. Support Vector Machines (SVM) were used, as it is considered one of the most promising classifiers in the field. The proposed approach lead to accuracies of 69.13% for arousal and 67.75% for valence, which are encouraging for further research with a larger training dataset and population.**

I. INTRODUCTION

Over the past decade, automatic emotion recognition systems have seen significant developments within academia and industry alike. Applications include psychology, healthcare, education, marketing, gaming or service robots, just to name a few [1], [2]. Various emotion recognition schemes have been proposed in the literature, exploring different elicitation methods, target emotional states, data sources, and classification techniques [3].

Most of the related works conducted user-dependent emotion recognition [4], whilst few explored user-independent scenarios. The user-independent approach proposed by Li and Chen [5] yielded encouraging results, with a recognition accuracy of 85.3% for threc cmotions, using Elcctrocardiography (ECG), Electrodermal Activity (EDA), Temperature and Respiration [5]. Similarly, Kim et al. developed a user-independent system using ECG, EDA and Temperature, and Support Vector Machines (SVM) as a classifier, with recognition rates of 78.4% and 61.8% for three and four emotions, respectively [6].

Given the importance of research towards user-independent emotion assessment, in this paper we propose an user-independent emotion recognition system, based on physiological data collected during a Virtual Reality (VR) -based elicitation protocol, and fusion into a multimodal classifier based on SVM. The work focuses on the classification of two classes for valence and two classes for arousal.

The rest of the paper is organized as follows. Section II describes the background in what concerns emotion theory. Section III highlights the main findings within the state-of-the-art in the field. Section IV presents the proposed feature extraction and classification workflow. Section V outlines the main results. Finally, Section VI outlines the main conclusions and future work guidelines.

Fig. 1: Overview of the two-dimensional model [11].

II. BACKGROUND

A. *Emotion Theory*

Various theoretical models of emotions have been proposed over the years. The discrete emotional model divides emotions into several basic emotions and claims their universality among all cultures [7], existing considerable agreement in six emotions: sadness, surprise, anger, disgust, and fear [8]. The valence-arousal dimensional model introduced by Lang [9] is the most popular, in which emotions are characterized according to their valence and arousal [8]. Valence represents pleasantness and ranges from negative to positive, while arousal indicates the activation level and ranges from low to high [9]. In this model all emotions can be understood as varying degrees of both valence and arousal [10], as illustrated in Figure 1.

It is important to note that the definition of emotion is fairly subjective, influenced by the cultural context, life experiences, and personality traits of each subject. However, certain core components of emotions are universal and likely biological, as argued by the meta-analysis found in [12].

[1] Joana Pinto and Ana Fred are with the Instituto Superior Técnico, Universidade de Lisboa, 1050-049 Lisboa, Portugal joana.f.pinto@tecnico.ulisboa.pt

[2] Ana Fred and Hugo P. Silva are with the IT - Instituto de Telecomunicações, 1049-001 Lisboa, Portugal {afred, hsilva}@lx.it.pt

[3] Hugo P. Silva is with the Escola Superior de Tecnologia, Instituto Politécnico de Setúbal, 2914-508 Setúbal, Portugal