TÉCNICO LISBOA

Real World Group Emotional Analytics Using Electrodermal Activity Signals

Gonçalo Filipe Duarte Salvador

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors: Prof. Hugo Humberto Plácido da Silva
Prof. Ana Luísa Nobre Fred

Examination Committee

Chairperson: Prof. João Miguel Raposo Sanches
Supervisor: Prof. Hugo Humberto Plácido da Silva
Member of the Committee: Prof. Hugo Alexandre Ferreira

November 2021

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.

Preface

The work presented in this thesis was performed at the Instituto de Telecomunicações (IT) and at the Department of Bioengineering of Instituto Superior Técnico (Lisbon, Portugal), in collaboration with the British Broadcasting Corporation Research and Development (BBC R&D), during the period of February 2021 to October 2021, being supervised by Professor Ana Luísa Nobre Fred and Professor Hugo Plácido da Silva, and co-supervised by Dr. Marco Torrado from the Faculty of Medicine (FM) from Universidade de Lisboa (UL). The supervision was also supported by Patricia Bota at IT, and by Vinoba Vinayagamoorthy and Joanna Rusznica at BBC R&D.

Acknowledgments

I would like to acknowledge my dissertation supervisors Professor Ana Luísa Nobre Fred and Professor Hugo Plácido da Silva for always pushing me towards greatness and supporting me in every step along the way. Furthermore, it was their insight and sharing of knowledge that has made this Thesis possible. I would also like to acknowledge Instituto de Telecomunicações (IT) for welcoming this project and experience of belonging to a team.

A special thank you to Vinoba, who supported and guided me throughout this work and was always available to address anything I needed. I would also like to extend this acknowledgement to Joanna and the rest of the BBC team, although this collaboration had to be done remotely due to the current pandemic situation it was still a constructive and enriching experience.

I also extended my acknowledgement to Patricia Bota, for all the support provided in helping me to get through the most difficult parts, and for providing me with several essential tools and knowledge which helped me bring this work to fruition. I would also like to thank Pedro Correia for conducting the experimental activities.

To all my friends who know that my way with words is not be the best, here are 80 pages of hard fought sentences written by me, which would not exist without all of you. Thank you for being there for me, thank you for being my inspiration. Thank you for all the good times we spend together, and know that I will always cherish the memories of these times until the end of my days. Where is a small token of my gratitude written in stone, with blood and tears which shall survive the most brutal test of all, time.

The person we become is the results of a combination of genetic and environmental factors throughout our lives. From this premise, the people which had the biggest influence in the construction of the person that I am today, the person who wrote this thesis, were my parents. There are no words in any dictionary which can express by deepest gratitude to the two people who brought me into this world. Mom, Dad, thank for for everything you gave up on for me, thank you for all the priceless life lessons you taught me, thank you for always helping me even when I did not know that I needed help, thank you...

To my brother, who I almost forgot, thank you for nagging me throughout all my life, but thanks also for helping me, even if all the help that I needed was to learn that I could to do it myself all along.

Abstract

The emotion recognition which almost all humans take for granted is a great challenge in the field of Human Computer Interaction.
The current work develops several tasks within the field of emotion recognition, such as the development of a new self-assessment annotations tool, an analysis of the minimum Sampling Frequency (SF) required for the acquisition of Electrodermal Activity (EDA) and the benchmarking of a new device to perform physiological data acquisition in group settings. Although, the most relevant task is the evaluation of collective group emotions while watching a long-duration uncalibrated audiovisual content in the wild, using EDA signals. The present work aims to analyze the similarities in simultaneous annotations across different participants, along with an analysis of the correspondent EDA signals and develop a new approach to identify time regions where the audience reacted with higher intensity based on the EDA data and unsupervised learning techniques.
The annotations performed by the participants did not follow the expected, revealing some limitations in the annotations phase. Furthermore, the evaluation of EDA data during simultaneous annotations revealed a tendency to increase over the period of the annotations. Although, the signals displayed few similarities during these time periods.
Regarding the application of clustering algorithms, the best performing methodology was hierarchical clustering with average linkage, providing a higher number of clusters, with more areas in which the audience had a more intense emotional reaction.

Keywords

Emotion recognition; Electrodermal activity; Emotional self-assessment; Valence-Arousal scale; Clustering algorithms.

Resumo

O reconhecimento de emoções que quase todos os seres humanos consideram natural é um grande desafio no campo da Interação Homem-Computador.
O presente trabalho, desenvolve várias tarefas dentro do campo de reconhecimento de emoções, tais como, o desenvolvimento de uma nova ferramenta de autoavaliação emocional, uma análise da Sampling Frequency (SF) mínima necessária para a aquisição de Electrodermal Activity (EDA) e o benchmarking de um novo dispositivo para realizar aquisição de dados fisiológicos em configurações de grupo. No entanto, a tarefa mais relevante é a avaliação das emoções coletivas durante a visualização de um conteúdo audiovisual não calibrado de longa duração. O presente trabalho tem como objetivo analisar as semelhanças em anotações simultâneas entre diferentes participantes, juntamente com uma análise dos sinais EDA correspondentes e desenvolver uma nova abordagem para identificar regiões onde o público reagiu com maior intensidade com base no EDA e técnicas de aprendizagem automática. As anotações realizadas pelos participantes não vão de acordo com o esperado, revelando algumas limitações na fase de anotações. Além disso, a avaliação dos dados de EDA durante as anotações simultâneas revelou uma tendência para aumentar ao longo do período das anotações. Ainda assim, os sinais exibiram poucas semelhanças durante esses períodos de tempo.
Em relação à aplicação de algoritmos de clustering, a metodologia que revelou o melhor desempenho foi o clustering hierárquico com linkage médio, proporcionando um maior número de clusters, com mais áreas em que o público teve uma reação emocional mais intensa.

Palavras Chave

Reconhecimento de emoções; Atividade eletrodérmica; Autoavaliação emocional; Escala de valência-excitação; Algoritmos de clustering.

Contents

List of Figures

List of Tables

Acronyms

ANS	Autonomous Nervous System
Ag/AgCl	Silver–Silver Chloride
DS	Downsampled Frequency
ECG	Electrocardiogram
EDA	Electrodermal Activity
EEG	Electroencephalogram
EDL	Electrodermal Level
EDR	Electrodermal Response
EMA	Ecological Momentary Assessment
EMG	Electromyogram
ENS	Enteric Nervous System
FMCI	Future Media Convergence Institute
GAPED	Geneva affective picture database
GSR	Galvanic Skin Response
IAPS	International Affective Picture System
HR	Heart Rate
HRV	Heart Rate Variability
MAP	Mean Affective Profile
NA	Negative Affect
NASA-RTLX	NASA Raw Task Load Index
OSMA	One-step Matrix Annotation
PA	Positive Affect
PAM	Photo Affect Meter

PANAS	Positive Affect and Negative Affect Schedule
PCA	Principal Component Analysis
PCC	Pearson Correlation Coefficient
PNS	Parasympathetic Nervous System
PPG	Photoplethysmogram
SAM	Self-Assessment Manikins
SCL	Skin Conductance Level
SCR	Skin Conductance Response
SF	Sampling Frequency
SNS	Sympathetic Nervous System
STD	Standard Deviation
SUS	System Usability Scale
TSSA	Two-step Sequential Annotation

1

Introduction

Contents

1.1 Motivation

Humans can express emotions to each other based on body language, facial expressions and speech, therefore emotions can be recognized based on these traits which occur naturally. This is why face to face conversations are usually more effective, since they involve a direct interaction between the participants, in which individuals can express and recognize each other's expressions [10]. Nevertheless, these cues do not reliably represent emotions, since these can be influenced by factors such as environment, cultural background, personality or mood, and they can also be effortlessly faked. Fortunately, emotions also have physiological manifestations.

The Autonomous Nervous System (ANS) mediates the body response to internal or external stimulus, thus, modulating the physiological manifestations of the emotion felt [7]. The physiological manifestations are expressed in physiological signals such as the Electrocardiogram (ECG), Electroencephalogram (EEG), Electrodermal Activity (EDA), and others. The link between emotions and physiological responses has advantages for emotion recognition. These responses are detectable, easily collected using wearables in a non obtrusive way, and they can reflect human emotion more reliably, since they can not be controlled or faked [2, 11]. With this in mind, it was considered that physiological signals are the preferable method to perform emotion recognition [7].

To study emotions, a first step should be to understand the concept of emotion. Over the years, several definitions have been proposed, although no consensus has yet been reached. Generally, the most accepted concept is that emotions can be described according to two different models: discrete model and continuous (or affective) model. In the discrete model, the emotional experiences are described based on a list of words to label emotions into categories. However, the list of words used for this description is still widely debated. Some researchers propose 6 basic emotions [12] (Happy, Sad, Anger, Fear, Surprise and Disgust), others propose 8 basic emotions [13] (Joy, Trust, Fear, Surprise, Sadness, Anger, Disgust and Anticipation), but many other models exist [14].

This discretization of emotions can be difficult, since the distinction boundary between two emotions is often blurred, and the meaning of the chosen words are culturally dependent [7]. On the other hand, continuous models aim to describe emotions based on continuous scales addressing two factors: the correlation between distinct emotions (e.g. Grief and Sadness are more similar than Happiness and Sadness), and the quantification of a certain emotion (e.g. it should be possible to differentiate between Sad and Very Sad). Russel et al. [15] proposed a 2D Valence-Arousal space, in which valence describes how pleasant an emotion is, and arousal describes the intensity level [14]. Although other models such as the valence-arousal-dominance have been suggested [16], the valence-arousal model is the most widely accepted [3].

A factor which greatly influences emotions is the group effect. Humans are highly social beings that tend to live in complex social structures. Thus, many emotions are experienced in social contexts

where there can be several interactions between group members. The group effect can have both a positive or negative impact on the emotional states experienced by the participants. On the negative side, crowd psychology can shape individual emotions leading to a more extreme behaviour, on the other hand, this effect can also create bonds between the members leading to an increase in the group effectiveness [17]. Previous work on emotion recognition focuses mainly on the analysis of emotion in an individual setting and in controlled environments, ignoring important dimensions. Hence, to evaluate emotions experienced by the subjects in a group setting, it is necessary to collect the data simultaneously from all participants.

The present work aims to fill a gap in emotion analysis by evaluating the emotional states in a group setting using long-duration uncalibrated elicitation content, i.e. movies. Although most studies focus on analysing emotions in an individual setting, being in a group environment can have different effects on the experienced emotions. The analysis of the emotional states in group setting was carried out based on EDA data acquired simultaneously from all participants. In addition, this work also seeks to develop a emotion self-assessment tool for smartphones. This tool is developed based on the Valence-Arousal model, enabling the users to perform their emotional self-assessment in a group setting with any sort of elicitation material, with minimal distraction.

1.2 Objectives

The objectives of the current work are related to the analysis of emotions in group settings using long-duration uncalibrated elicitation materials:

- Study the emotional dynamics in a group setting in a real-world setting
- Design an experimental protocol to elicit emotions in a group setting using long-duration, uncalibrated video content;
- Collect EDA data and the corresponding emotional self-assessments;
- Implement the necessary steps to preprocess and detect the fiducial points in EDA data;
- Analyse emotional annotations, comparing simultaneous annotations and establishing a correspondence between the annotations and the elicitation which triggered them;
- Select and extract the most relevant features from the EDA signal;
- Use machine learning methods to categorize the parts of the elicitation content where the audience had a similar emotional reaction;
- Compare the main findings of this work with the approaches found in the literature.
- Develop a smartphone application to perform the emotional self-assessment in a real-world scenario (i.e. for both individual and group settings, for any type of elicitation material);

- Validate the performance of the developed application using the System Usability Scale (SUS) and the NASA Raw Task Load Index (NASA-RTLX);

1.3 Contributions

The current work provided the following contributions to the field of emotion recognition:

- Published and presented the paper entitled ”Smartphone-based Content Annotation for Ground Truth Collection in Affective Computing” at the 2021 ACM International Conference International Media Experiences [18];
- Abstract accepted and paper ”Impact of Sampling Rate and Interpolation on Photoplethysmography and Electrodermal Activity Signals Waveform Morphology and Feature Extraction” under revision in the Neural Computing and Applications Journal (Q1) under the Topic Collection ”Computational-based Biomarkers for Mental and Emotional Health”;
- Database with experimental data acquired in group settings using uncalibrated long duration elicitation content, including EDA signals and emotional annotations (Valence, Arousal and level of annotation uncertainty);
- Development of a list of elicitation content, consisting mainly of recent movies (from the last 3 years), and encompassing several genres to elicit a wide range of emotions;
- Implementation of EDA outlier removal and feature extraction methods, proposed to integrate the public biosignals processing library BioSPPy [19];
- Emotional group dynamics analysis based on the self-assessment annotation performed by the participants and evaluation of the individuals and audience reaction using EDA data and machine learning algorithms;
- Development of a smartphone application for emotional self-assessment in a group setting publicly available in the *Google PlayStore* [1, 2];
- Validation of BITalino R-IoT data quality against a benchmark device (BITalino (r)evolution)
- Analysis of the minimum Sampling Frequency (SF) required to acquire a quality EDA signal.

1.4 Thesis Outline

The present work is divided into seven chapters, organized in the following manner. Chapter 2 portrays the relevant theoretical background for this work, namely: 1) models used to define emotions; 2) assessment methods used to evaluate the participant's emotional state; 3) relation between emotion and

[1] https://play.google.com/store/apps/details?id=com.emoteu.app
[2] https://play.google.com/store/apps/details?id=com.emoteu2.app

the ANS, as well as the physiological manifestations of emotions; 4) description of the EDA signal, its characteristics and relation with emotions; and 5) different elicitation materials which can be used to trigger emotions.

The next four chapters describe the main problems addressed by this work. Each chapter starts with a motivation, followed by the state of the art of that topic, along with the methodology description, results and discussion of the developed work (except for Chapter 3 which does not have results nor discussion). Chapter 3 describes the methodology for experimental data acquisition in a group setting, presenting the data acquisition and emotional annotation tools, along with the setup used to present the elicitation content and store the acquired data. Chapter 4 summarizes the emotional analysis performed on the collective data. This chapter contains 2 different analysis: the first consists in evaluating the annotations given by the participants and the assessment of the similarities of synchronous annotations across different participants. The second analysis suggests a new approach to identify time regions where the participants and the audience reacted with higher intensity, based on EDA data and unsupervised machine learning techniques. Chapter 5 presents the development of the real-time collective emotional self-assessment tool, starting with a description of existing annotation tools, and followed by the description of the different steps taken in the development of this application, ending with real-world evaluation of this tool.

Throughout the development of the current work some limitation where identified, namely, the low acquisition frequency of the Future Media Convergence Institute (FMCI) Xinhua Net device, and limitations in the emotion recognition performance when using a single physiological sensor (EDA). Taking this into consideration, Chapter 6 presents an analysis of the minimum SF required for the acquisition of the EDA signal, along with a benchmarking of a new device, the BITalino R-IoT, capable of acquiring EDA and Photoplethysmogram (PPG) data simultaneously across several participants. Finally, Chapter 7 draws the main conclusions obtained throughout the current work.

2

Background

Contents

This chapter contains the relevant theoretical background regarding the field of emotion recognition. Throughout the years emotion have been described using different theoretical models, thus Section 2.1 presents an overview of the most common theoretical models to describe emotions.

Then, considering the different manifestations of emotions, Section 2.2 introduces the different modalities typically used to assess them. Afterwards, the relation between emotions and the ANS is described in Section 2.3, particularly the connection between emotion and physiological signals. A detailed presentation of the EDA signal can be seen in Section 2.4, containing the relevant characteristics of this signals for the current work. Lastly, Section 2.5 provides an overview of the different elicitation material used to trigger emotions.

2.1 Emotion Models

The need and desire to understand human emotions dates back to the philosophers in Ancient Greece, such as Plato and Aristotle [20]. In the Roman Empire, Cicero and Graver created an emotional model discretizing emotions under four categories: Fear, Pain, Lust, and Pleasure [1, 7]. Hundreds years later, Wundt [21] proposed a completely different model of emotions describing all emotional states as a single point in a three-dimension space: pleasure-displeasure, excitement-inhibition, tension-relaxation [1]. These represent the first discrete and continuous models of emotion, respectively.

Over the years, several other definitions have been proposed, although no consensus has yet been reached. Generally, the most accepted concept is that emotions can be described according to discrete or continuous models. Other models exist, such as the Appraisal Theory, which describes emotions as a process rather than a state. However, these models are harder to integrate into machine learning algorithms to identify emotions [22, 23], reason for which the discrete and continuous models are still the most widely used. [1].

Discrete Emotion Model

In the discrete model, emotions are categorised. Categories are represented by words, which are associated with a certain significance in expressing emotions, and several taxonomies have been proposed throughout the years. Ekman [12] argues that emotions are shared between cultures and, as such, they can be universally recognised. In his view, emotions arise from evolutionary physiological and communicative functions. The manifestations of emotions could help in life or death scenarios, e.g. a facial expression of fear could imply a situation of danger and warn people nearby. Furthermore, Ekman listed six discrete basic emotions, each having a distinct physiological pattern: Joy, Sadness, Anger, Fear, Disgust, and Surprise [12]

Figure 2.1: Plutchik's color wheel of emotions. [1]

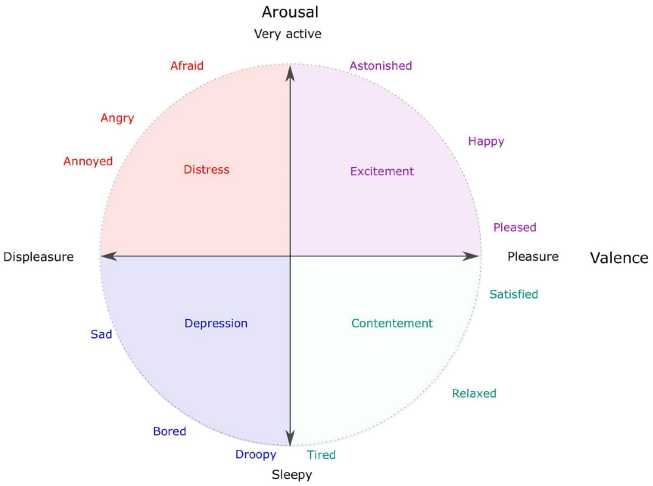


Figure 2.2: Russel's circumplex model of emotions. [2]

Plutchik [13] created a taxonomy to classify discrete emotions, known as 'wheel of emotions' [1], since it was based on a wheel incorporating eight primary emotions: Grief, Amazement, Terror, Admiration, Ecstasy, Vigilance, Rage, and Loathing. Furthermore, emotions can have different intensity levels, thus, the primary emotions are located in the center of the wheel, while weaker emotions occupy the extremities as shown in Figure 2.1. In his taxonomy, emotions could be mixed to form new and more complex emotions.

These models are based on describing emotions with a single word. However, this discretization can be difficult to perform, since the distinction boundary between emotions can be blurred, and complex mixed emotion scan be difficult to label into a single word. Furthermore, the meanings of the chosen words are culturally dependent, so similar emotions could be described using different labels [7, 24]. As such, it can be hard to implement this model, since it is necessary to reduce a wide range of emotions to a finite amount of labels and, even in cases where the model is applied, it may not produce completely reliable results due to the cultural dependence in the meaning of such labels.

Continuous Emotion Model

To overcome the difficulties found in the discrete models of emotion, a new concept of describing emotion emerged. This concept consisted of mapping emotion into a multidimensional space [1]. These dimensional spaces must address two factors: the correlation between distinct emotions, and the quantification of a certain emotion.

Similarly to the discrete models, several different dimensions have been proposed to measure emotions. Russel's two-dimension model is a popular approach [15], in which emotions are characterized as a discrete point in space composed of two axes, valence and arousal. The Arousal axis describes emotions in terms of intensity (e.g. how energised one feels), while Valence axis portrays emotion in

terms of how positive or negative an emotion is. In Figure 2.2 it is possible to see a schematic representation of this model. The Valence-Arousal model was later extended by Mehrabian [16], adding a new dimension called Dominance to better describe the consciousness of emotion. This dimension aided in the distinction between emotions such as Anger and Fear.

In emotion recognition, the Valence-Arousal is the model which is most frequently used [1, 7] due to several reasons. From a machine learning point of view, this model has low complexity and allows for different classification problems (e.g. multiclass classification with low/medium/high arousal or valence, or a classification based on the four quadrants of emotion) [1]. Another reason for the popularity of these models is the simplicity of emotion assessment, since it is already integrated in validated and vastly used questionnaires, such as the Self-Assessment Manikins (SAM), and it can also be easily integrated into a new forms of questionnaire that can be understood across cultures [3].

2.2 Assessment Methods

For performance assessment when working with machine learning models, a ground truth value is required. To acquire the ground truth emotional value, one common practice is to annotate the individual's emotions, which can be done through internal annotation and/or external annotation. Internal annotation methods, also called self-assessment methods, involve asking the participants to report the emotional state felt during the experiment.

Although this method can be easy to implement and replicate, the annotations acquired may not be completely accurate due to the influence of subjective factors during the annotations. Namely, (1) participants can have difficulties expressing their emotions into words and/or scales, (2) they may hesitate to give honest answers when such answers are socially undesirable, and (3) the rationalizations of their answers may affect the perceived emotion [25]. Furthermore, for some participants this method may be intrusive, causing the subject to unreliably report their emotion to preserve their privacy; this can be surpassed with the implementation of data protection measures reassuring the participants' privacy [26]. On the other hand, in external annotation methods, also called implicit assessment methods, an external subject assesses the subject affective state based on the analysis of observable factors. In this method, the external subject can be easily deceived by faking observable manifestations of emotion [27].

Both methods have been shown to have a significant correlation between the two [27], however, the self-assessment method is still the most applied emotional annotation approach, usually based on a questionnaire presented to the subject [7, 28]. Several questionnaires can be used for this annotation, such as the SAM, Positive Affect and Negative Affect Schedule (PANAS), or the Affect Grid described below. Despite being the most common, others like the Photo Affect Meter (PAM) [29], Ecological Momentary Assessment (EMA) [30], are also frequently found in literature.

Self-Assessment Manikin

The SAM is a questionnaire that provides a non-verbal, graphical representation for cross-cultural measurement of emotional response [7]. In the representation of this scale (Figure 2.3), it is possible to see three rows with five images, each row represents a different dimension (Valence; Arousal and Dominance); the questionnaire is answered by placing an 'x' over any of the five images, or between any two figures, which results in a nine-point scale. The Valence axis ranges from a smiling figure to a frowning figure; in the Arousal axis the manikins range from an excited figure to a sleepy figure; and, lastly, in the Dominance axis, the manikins range from a small figure to a very large figure, representing the range of feelings between being controlled or submissive, to being in control or a powerful feeling [3]. Furthermore, this questionnaire can be adapted by using different scales and/or figures which allows the acquisition of annotations without being skewed by cultural perceptions of emotions [31]. Although this questionnaire includes the annotation of the Dominance dimension, such dimension is not commonly used.

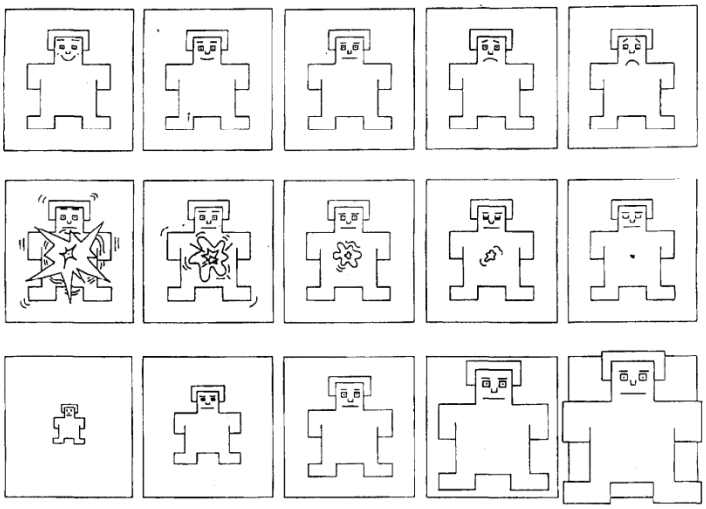


Figure 2.3: The Self-Assessment Manikin (SAM) used to annotate the affective dimensions of Valence (first row), Arousal (second row) and Dominance (third row) [3].

Additional Questionnaires

The PANAS is a questionnaire with two distinct axes, the Positive Affect (PA) measures the level of Enthusiasm, Activeness and Alertness of the participant, thus a high level of PA characterizes a state of high energy, full concentration and pleasurable engagement, in opposition a low level of PA characterizes a state of Sadness and Lethargy. On the other hand, the Negative Affect (NA) measures subjective distress and unpleasurable engagement, which include different states, such as Anger, Disgust, Guilt, Fear and Nervousness, thus a low level of NA represents a state of calmness and serenity [32]. This

questionnaire is represented by 20 terms (10 per axis), representing different feelings and emotions, and the participant must indicate the association between his emotional state and each term with an integer level from 1 to 5 (1 - Not at all, to 5 - Extremely) [27].

A different approach, proposed by Cowie et. al [33], has the users annotating their emotional state by moving the cursor on a computer screen displaying a 2D circular space. This display has verbal landmarks at the periphery and within the the circle, to ensure that participants could easily relate the position to a categorical description of emotion. Furthermore, the cursor is also colour coded, following the colour proposed in Plutchik's color wheel of emotions(Figure 2.1) [13]. Another approach for emotion annotation was proposed by Lopes et. al [34]; this annotation technique is based on a "wheel-like" hardware which the participants can use to increase or decrease the intensity of a single emotion dimension. This approach uses an unbounded annotation, which means that the annotation has no upper or lower limit, allowing users to adopt a range of values as broad as they wish. This revealed to achieve better results than with bounded annotations in [35]. These two methods enable a continuous annotations that, in turn, allow the analysis of the evolution of the emotional state throughout the elicitation content.

Assessment based on Physiological Signals

Unlike the previous methods that require a subject to annotate emotions on a certain questionnaire, emotions can also be assessed based on body and facial expressions. We conduct our social interactions based on the behavioural traits that the people around us express, either through body language, facial expression and speech [24], which are the simplest methods to express and identify emotions in daily life. However, these demonstrations of emotion can be faked, hence not representing emotions reliably.

Emotions can have three different types of manifestations: behavioural, e.g. the smile a person expresses when told a joke; physiological, the sweating and dry mouth experienced when being anxious or afraid [1]; and chemical, in stressful experiences there is an increase in the production of hormones such as cortisol, oxytocin and progesterone [36]. As it was previously described in Section 2.2, assessing emotion through their behavioural and chemicals manifestations can be unreliable and/or intrusive methods since external manifestations of emotion can be faked, and analysing chemicals manifestations requires drawing blood or saliva. So, the preferred method to assess emotions is through physiological manifestations.

A viable alternative to assess one's emotional state is to use physiological signals since, as presented in Section 2.3, emotions are linked to physiological changes, one can also use physiological signals to assess emotions. The use of these signals allows the implementation of non-intrusive methods to assess emotions. Their acquisition only requires the application of electrode on the skin of the participant, which,

although may restrict certain movements, is still a manageable and reliable method; the manifestation of emotions in physiological signals is an unconscious process that is difficult to manipulate [37].

2.3 Autonomous Nervous System

Emotions are reflected by responses of the ANS and the limbic system (in particular the amygdala). The amygdala is the region of the limbic system which is primarily responsible for the regulation of the perception and reaction to aggression and fear. It has connections with facial muscles, which perceive and express emotions, and other bodily systems related to emotions; it also regulates the release of neurotransmitters related to stress and aggression [38].

The ANS is divided into three branches: the Sympathetic Nervous System (SNS), the Parasympathetic Nervous System (PNS), and the Enteric Nervous System (ENS). The SNS is associated with the "fight-or-flight" response, thus being activated during physically or mentally stressful situations. This system is responsible for increasing Heart Rate (HR) and strength of cardiac contractions, increase in respiration rate and bronchial tube dilatation, pupils dilatation, decreased salivation and digestion, and, lastly, adrenaline and glucose release. During high arousal states, like an angry situation, one usually experiences the heart pounding, troubles breathing, a sick feeling in the stomach, etc.; this is the result of the activation of the SNS [38]. On the other hand, the PNS is associated with the "rest-and-digest" functions, being responsible for the homeostasis of the body. This system is responsible for slowing the HR, decrease blood pressure, increase salivation and digestive system activity, muscle relaxation, pupils constriction and increase in urinary output [1, 7]. Lastly, the ENS is responsible for the digestive functions of muscle contraction/relaxation, secretion/absorption, and intestinal blood flow (e.g. this system is responsible for the movement of water and electrolytes across the intestinal wall) [39], hence not very related to the response of the ANS to emotions.

Although the SNS and PNS branches have opposite functions, they both work in harmony, for example, in a potential threat, the SNS raises the HR, and after the threat is over the PNS brings the HR back to a rest state. Therefore, a good way to measure the activity of these two branches is to measure the Heart Rate Variability (HRV), a high HRV implies an increased activity of the SNS and a low HRV implies an increased activity of the PNS. So, a good way to evaluate the activity is through the PPG or ECG signals from which it is possible to extract the HR. Another way to measure the SNS activity is to measure the EDA, because this signal is solely stimulated by the SNS.

The emotional physiological responses can be assessed with several different physiological signals: ECG [1,24]; PPG [1,7]; Respiration [37]; Skin Temperature [1,7]; Electromyogram (EMG) [40]; EEG [41]; EDA [7, 42]

The work developed where is going to be focused on the analysis of emotion in group settings

using physiological data. However, at the moment there are not many viable solutions for simultaneous data acquisition in several participants. One of the existing options is the device FMCI [43], which only acquires the EDA signal. This signal has been shown across several pieces of research to be correlated with the emotional state of the user (being especially correlated with the arousal dimension of emotions) [44, 45]. Furthermore, this signal is the most commonly applied in researches in the field of emotion recognition. Some studied have also been conducted using physiological data only from the EDA and achieving very promising results [6, 42, 46]. So, for these reasons the EDA will be the main physiological data source used in the context of this thesis.

As it is possible to see, the ANS is responsible for a wide variety of functions, from a body regulator, through homeostasis, to an activator, by allocating body resources to better respond to stimulus, amount many order functions which were not mentioned here. So, due to this multi-functionality of the ANS, it is hard to establish a direct correlation between a subject's emotional state and their current physiological signals since a change in these signals can be a result of an emotional state but it can also be a result of one of the ANS non-emotional functionalities [7].

2.4 Electrodermal Activity

EDA (also referred to as Galvanic Skin Response (GSR)), measures the electrical conductivity of the skin. This measurement is correlated with the activity of the sweat glands in the skin. Although sweating is usually a result of thermoregulation, the sweat glands are also controlled by the SNS, hence being responsive to psychophysiological stimuli. For example, an increase in the activity of the SNS leads to an increase in the activity of the sweat glands, which in turn increases the skin conductance.

During periods of arousal, the SNS is active leading to an increase in the sweat glands activity, which can be noticed mainly on palmar and plantar sites, in axillary and genital regions, as well as on the forehead. This increase in the sweat gland activity culminates with an increase of the EDA amplitude, thus, one can say the EDA signal is correlated to the arousal level experienced [47]. The electrodes should be placed in regions of the skin with a high density of sweat glands, e.g. palm/finger (Figure 2.4) or feet [4].

There are two different methods of measuring EDA: without applying an external current (endosomatic method) or by applying a current (exosomatic method). In the endosomatic method, an electrode is placed on an active site, such as the palm, and a second electrode is placed on an inactive site, such as the forearm. Through this method, the Electrodermal Responses (EDRs) can be easily identified although the waveform is more complex than in the other methods, i.e. the waves can be mono-, bi- or triphasic, while in the exosomatic method these waves are always monophasic [4]. In the exosomatic methods, a Direct Current (DC) or Alternating Current (AC) is applied to the skin. The measurement of

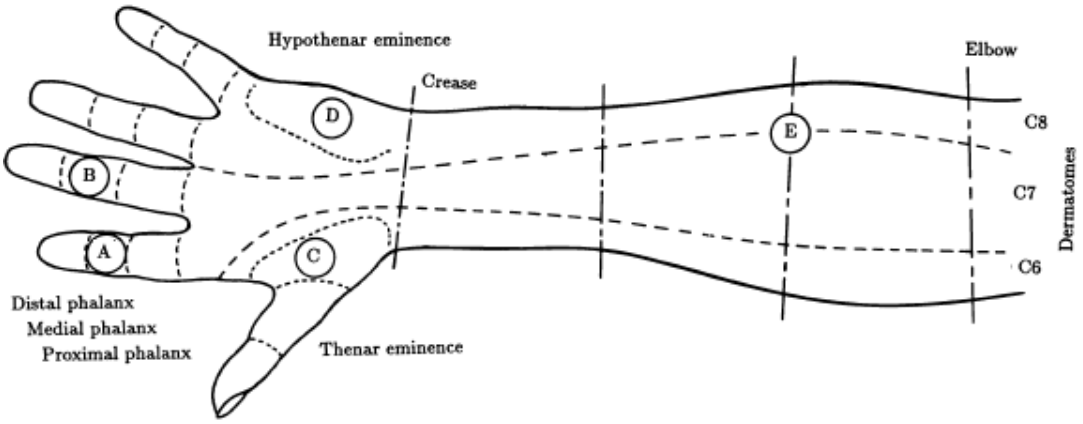


Figure 2.4: Preferred palmar electrode placement (A-D) and recommended placement for the inactive electrode for endosomatic approach [4].

EDA using direct current with Silver–Silver Chloride (Ag/AgCl) electrodes and an electrolyte of sodium or potassium chloride is the most commonly used method.

This method consists in applying a small constant voltage on two electrodes placed on the palm's skin of the same hand (to avoid ECG artefacts). Given that the applied voltage is a known value, the skin resistance can be determined based on the measured current using the Ohm's law (a more extensive review on the measurement of the EDA can be found in [47]). Usually, the EDA is measured in micro Siemens (μS) [48]. Using the exosomatic method with direct current may lead to the polarization of the electrodes, resulting in a behaviour similar to a rechargeable battery by the electrodes, with a voltage opposing the applied one, and introducing a bias in the recordings of the skin conductance [48]. To circumvent this problem associated with the use of direct current, the exosomatic method using alternated current can be applied. However, this method is not very common, since it requires more elaborate instrumentation and it has a complicated comprehension of the acquired data (due to the phase shift caused by the combined conductance and capacitance in the skin) [48].

The EDA can be decomposed into two main components: phasic and tonic. The latter one corresponds to a baseline signal with low bandwidth ($f < 3Hz$) and it expresses the Electrodermal Level (EDL) (also known as Skin Conductance Level (SCL)) [7]. This is a slowly changing signal, which reflects the overall baseline, not directly related to any stimulus [4]. The EDA signal also contains nonspecific skin conductance responses, which are phasic peaks in the EDR signal, but they occur in the absence

of external stimuli or artefacts (Non-Specific EDR).

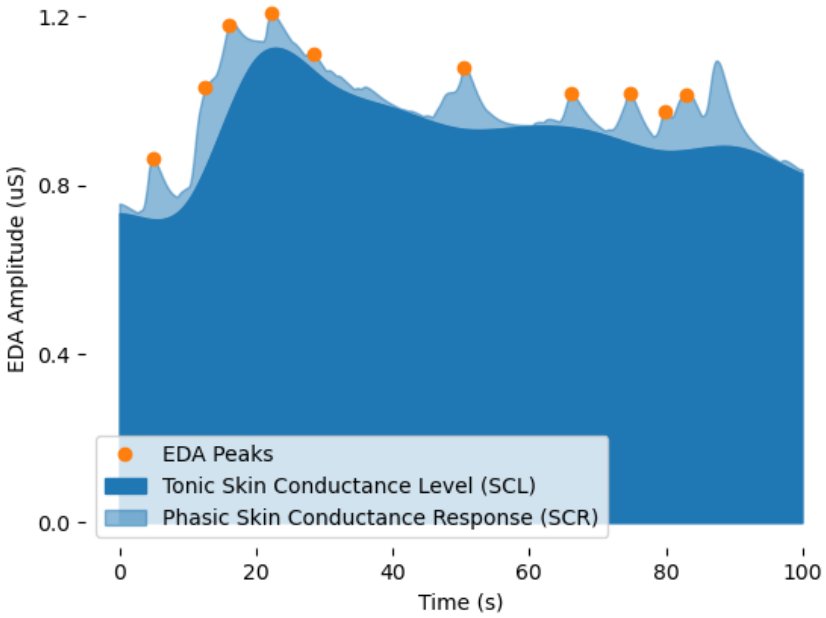


Figure 2.5: Example EDL and EDR signals.

The manifestations of the ANS on the EDA signal are expressed by the phasic component, in order words, the physiological changes caused by the SNS as a response to a stressful situation, for example, are contained in the EDR component. These responses are represented by short-lasting changes in the signal called EDR, or Skin Conductance Response (SCR) which can be elicited by distinct stimulus [48]. The peaks detected in the EDR are usually in response to a stimuli, thus they are called Event-Related EDR. On the other hand, peaks with a similar shape but occurring in the absence of external stimuli are called Non-Specific EDR. These responses have a relatively long latency period, between stimulus and signal onset, ranging from 1 second (s) up to 5s; it is important to note that this measurement is influenced by external factors such as the room temperature and recording site. The mean latency time in a room with a temperature of 30ºC is 1.5 s [4].

In this work, we focus on the exosomatic EDR (with direct current, which as previously described is the most commonly found in the literature). In Figure 2.6 it is possible to observe a prototypical EDR waveform with a short rise time and longer recovery. Usually, the rise time is shorter than its recovery time. The rise time has a range between 0.5 and 5 s, with a mean time of 2.184 s and standard deviation of 0.643 s, the distribution being slightly positively skewed and platykurtic [49]. On the other hand, the half recovery time has a mean value of 4.144 s with a standard deviation of 2.466 s, the distributions being slightly positively skewed and leptokurtic [49]. In cases with multiple high arousal stimulus being

presented in a short time period, several EDR events can become overlapped (Figure A.1).

The level of distortion depends on the proximity and amplitude of a preceding EDR. These levels are labeled from ”type 1” (represented in Figure 2.6) to ”type 3”. In Figure A.1, it is possible to observe ”type 2” and ”type 3” of overlapping and three methods of calculating the EDR amplitude. In the ”type 2” overlapping, the first EDR does not reach the half recovery time, thus method A or B can be used to measure the amplitude of the second EDR. In this scenario, method A estimates the amplitude of this response by measuring the vertical distance from its peak to the extrapolated recovery line of the first response. On the other hand, method B estimates the amplitude of this response by measuring the vertical distance from its peak to its onset. In the ”type 3” overlapping, there is no recovery after the first peak of the curve, but instead another ascent. In this scenario, method C considers a single EDR and measures a single amplitude. Considering these approaches, method B is the most widely applied and it has been shown to produce sufficiently accurate results in most cases [4].

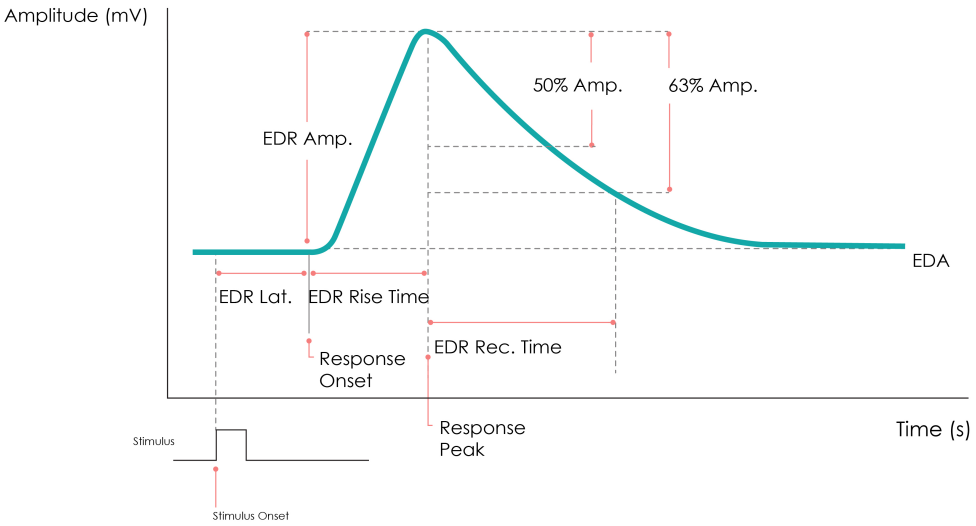


Figure 2.6: Prototypical EDR waveform with the respective characteristic to be extracted from it. Adapted from [4].

Beyond the decomposition of the EDA, another method to extract further information is to extract features. These metrics compactly characterise the signal and allow the comparison of the signal across different subjects or time instants enhancing the information that can be extracted from the signal [7]. The features can be specific to a certain signal or general time series characteristics. The extracted features can be grouped into the following classes: temporal, statistical and spectral. Some of the most commonly extracted features are presented in Table 2.1. As stated above, one can also extract specific metrics from the EDA. The list of commonly extracted features in the literature from the EDA signal are shown in Table 2.2, although some features provide more information regarding the emotional state than others, e.g. the mean value of the signal is correlated to the level of arousal. Feature extraction is critical to achieve good outcomes, thus finding features of the physiological signals that correlate with emotional

Table 2.1: Features commonly extracted from time-series signals grouped by their domain [7].

Domain	Features
Temporal	Maximum, minimum, centroid, median/mean absolute deviation/difference, zero crossing rate, linear regression, range, absolute integral
Statistical	Mean, median, Standard Deviation (STD), variance, interquartile range, root mean square, skewness, kurtosis, histogram
Spectral	Total energy, spectral centroid, spectral spread, spectral skewness, spectral kurtosis, spectral slope, spectral decrease, spectral roll-on/off, spectral variation

Table 2.2: Features commonly extracted from EDA signals grouped by their domain [7].

Domain	Features
Temporal	EDL degree of linearity, temporal features on EDR signal, number of EDR events, sum of SCR startle magnitudes and response durations, area under the SCR events, temporal features on SCR amplitudes, rise and 50%/60% recovery times
Statiscal	Statistical features applied to: SCR signal, amplitudes, rise and 50%/60% recovery times
Spectral	Spectral features applied to SCR signal, 10 spectral power bands in the 0.2-4Hz range

states of an individual is an arduous task.

2.5 Elicitation Material

Emotions are highly subjective and have great variability, hence obtaining data that corresponds to a particular emotional state can be very challenging. To overcome this challenge, researchers rely on a controlled environments (e.g. a lab), where emotions can be triggered with specific emotional stimuli following well designed protocols. These emotional stimuli are pre-validated, in an attempt to reliably induce the desired affective state, and different elicitation methods exist, namely: pictures, videos, Virtual Reality videos, games, sound, words, recall, acting. Nevertheless, the state-of-the-art has been progressing to less controlled real world scenarios [7].

Due to its low cost, and easy replicability, the use of pictures is one of the most common approaches, with several sets of standardized images for the elicitation of attention and a wide range of emotional experiences being proposed over the years. For example, the International Affective Picture System (IAPS) [10, 44, 50] contains images rated in terms of valence, arousal and dominance. The Geneva affective picture database (GAPED) [51] consists of 730 images, separated by positive, negative or neutral content, and rated in terms of valence and arousal. These are just a few sets of standardized images.

A second preference in the state-of-the-art in emotion elicitation, is to use videos. Emotion elicitation based on films or short-duration audiovisual video clips have shown to be the most reliable material for emotion elicitation [7]. Most studies use short movies (under 20 minutes) or small clips of films; this is the case in the validated video databases, such as: LIRIS-ACCEDE [52], HUMAINE [53], and MAHNOB-

HCI [54]. Only the database DEAP [55] has videos with more than 20 minutes, containing 40-minute music videos self annotated by 32 participants in terms of Valence-Arousal, Like-Dislike familiarity and Dominance.

However, in daily life, emotions are experienced based on random events or triggers, which can lead to a variety of complex emotions that depend on the subject. So, eliciting emotion with calibrated data does not accurately mimic the emotions commonly experienced in the real-world [56]. Furthermore, the majority of the emotions experienced in the real-world are not felt in constrained environments, such as the labs. Emotions are usually experienced in social contexts, where emotions are not only dependent on the participant and the elicitation stimulus, but also on the environment in which the individual is inserted. So, there is need to further study emotional dynamics in real-life condition using uncalibrated and previously unseen elicitation materials, such as a movie [17, 27].

3

Collective Emotion Assessment

Contents

This chapter presents the data acquisition protocol used to evaluate collective physiological responses while watching a long-duration uncalibrated audiovisual content in a real-world setting. Section 3.1 presents the reasoning which led to the evaluation of emotion in a group setting, namely the impact of being in a group settings. Section 3.2 comprises an overview of the state of the art in the scope emotional assessment in terms of the acquisition protocols used in previous works. Section 3.3 describes the data acquisiton devices, the annotation tool used and the data acquisition protocol.

3.1 Motivation

In real life scenarios, emotions are usually experienced in social contexts, where emotions are not only dependent on the participant and the elicitation stimulus, but also on the environment in which the individual is inserted and the implicit and explicit interactions that can occur between the group members [17, 27]. However, most studies focus on the individual setting emotion analysis, and have ignored the important dimension that is the group setting.

In a group environment, there are several subconscious processes that lead to changes in the emotions experienced by the group members. Primitive emotional contagion is the main process used to explain group emotions; in this process an individual mimics another person's emotion unconsciously and automatically [57]. For example, individuals in a crowd can develop similar emotions not due to individuals observing each other but because they unconsciously mimic each others' emotions [58]. Other subconscious processes are behavioral entertainment and synchrony of interaction, in which individuals adjust their emotions in order to be coordinated and synchronized with the rest of the group; this occurs through the observation of facial, postural, and behavioral expression [59].

Besides subconscious processes, there are also conscious processes within a group environment. In these conscious processes, individuals acquire cues regarding each other's emotions, and compare them with their own to judge whether their emotions are appropriate to the current situation [58]. Additionally, emotions can also be shared through social interaction, e.g. a conversation regarding a certain emotional state may lead to an individual relating to the circumstance, thus reliving the emotions expressed [58].

Group emotion have a great effect on group outcomes. Positive group emotions reduce conflict, increase cooperation and performance, although it may also lead to the spread of unrealistic euphoria. On the other hand, negative group emotions are associated with a decrease in creativity, reduce the span of attention and cooperation [58]. This part of the work aims to investigate the impact of emotional dynamics in a group setting, through the evaluation of collective physiological responses while watching a long-duration uncalibrated audiovisual content in a real-world setting, based on EDA signals. The data is used to create a database for collective emotion assessment containing annotated EDA signals in

terms of Valence, Arousal and level of confidence in the annotation.

3.2 State-of-the-Art

In the work of Haag et. al [60], physiological signals are collected in a controlled environment (laboratory) with the participants being alone in a room, while emotions are being elicited using a calibrated image dataset. The work of Domíngues-Jiménez et. al [2] follows a similar approach, but in this case using a calibrated video dataset.

In daily life, emotions are experienced based on random events or triggers, which can lead to a variety of complex emotions that depend on the subject. In contrast, eliciting emotions with calibrated data restricts the range of reactions from the participants. To overcome this constrain, one can use uncalibrated and previously unseen elicitation materials, such as a movie. In the work of Lee et. al [56], the authors use uncalibrated movies to elicit positive or negative emotions according to the movie genre. The movie was displayed in an individual setting, while acquiring ECG and EDA signals from the participant.

However, Humans are highly social beings that tend to live in collective structures. Hence, studies that elicit emotions in an individual setting are ignoring an important component for the study of affect. In [6], Wang et. al studies the influence of a commercial audio track on an audience of 15 subjects through the evaluation of EDA data. Participants were in a controlled room watching, in turn, three different sets of the same video with different audio tracks, while EDA data was being collected on the fingers. Before and after each video, participants individually filled two questionnaires to assess their engagement to the clip shown.

The work developed by Miranda-Correa et. al [27] studies the influence of individual versus group settings. In this work, participants performed the experiment either in an individual or a group setting (constituted of 4 subjects). The experiment consisted of watching a movie clip with 14 minutes (or longer) and filling a self-assessment questionnaire based on the SAM, before and after the elicitation clip. The clips were extracted from different movies and did not require any previous knowledge to be understood. In order to maximize group interactions, groups were formed with people that already knew each other beforehand, although the experimenters did not say if the participant could talk to each other for the interactions to be spontaneous. The experiment was conducted in a recording room, and during the visualisation of the clips physiological signals were collected, namely, EEG, ECG and EDA along with frontal HD video of the participants.

In the studies described above emotions are elicited in a controlled environment, however, the majority of the emotions experienced in the real-world are not felt in such a constrained environment, thus motivating the need to further study emotional dynamics in real-life conditions [61]. Ojha et. al [46]

studies the arousal state of subjects in an individual uncontrolled real-world scenario consisting of a walk through the streets of Zürich, using EDA data. Fleureau et. al [45] evaluates an audience arousal response in a real-world scenario (a film festival), while watching a movie in a cinema room. To this end, the EDA signal is recorded for every participant during the whole duration of the movie, although no assessment method is applied to obtain the participants opinion regarding the movie. This study uses an uncalibrated content, in a group setting and in a real-life context, effectively measuring the audience arousal level solely based on the EDA signal and the movie highlights according the group opinion as a whole.

3.3 Proposed Methodology

This section is divided into three subsections. The first subsection describes the characteristics of the data acquisition device used to collect the participant's EDA signal. The second subsection presents the emotional annotation tool used to obtain a retrospective emotional self-annotations. Finally, the third subsection describes the experiment setup and data acquisition protocol.

Wearable Device

The device chosen to acquire physiological data during the visualization of audiovisual elicitation contents was the Xinhua Net FMCI device. The device was selected due to its capability to wirelessly acquire EDA data from up to 20 devices simultaneously. The performance of this system was evaluated in [43], showing the feasibility of signal acquisition with no significant data loss, in collective settings. The FMCI device consists of a small wrist bracelet with two electrodes connected; two electrodes are attached to the palm or finger area, making it a highly mobile data acquisition device. The device collects EDA data through an embedded sensor designed to acquire EDA signals with a bandwidth between 0 and 5 Hz with a sampling frequency of 1 Hz, and uses a combination of low-power communication and and has a battery that enables it to work for over 50 hours. The sensor consists of an operational transconductance amplifier and a low-pass filter. The former is used to increase the amplitude of the weak potential differences generated from the biological electric signals. Additionally a second-order low-pass Butterworth filter (gain = 2, cut-off frequency = 5Hz) is applied [43].

Emotional Annotation

The volunteers' emotional annotation was performed retrospectively using the EmotiphAI annotation tool developed by Bota et al. [5]. The user interface for emotional self-reporting can be seen in Figure 3.1. At the top of the interface (Figure 3.1 - A) it is possible to select which participant is currently performing

their emotional self-report; to protect the privacy of the participants, these are anonymously identified by their device ID. The segments to be annotated by the user are displayed below (Figure 3.1 - B). These segments correspond to the time area where the subjects' EDA signal responded with higher intensity. Furthermore, since the annotations are being performed in specific time segments, each annotation has the corresponding time span. On the left there is a visible video player (Figure 3.1 - C) for the users to see the video clips which they are currently annotating. The goal is that, through the visualization of the clip, the user can recall the emotional state experienced during the first visualisation of the content. This emotional state should be annotated on the right (Figure 3.1 - D) using the digital SAM [3] Arousal and Valence self-reporting scale. Below each dimension there are two uncertainty buttons (Figure 3.1 - E), for the user to rate how certain they were of their Arousal and Valence reports, respectively. The uncertainty of the participant in the annotation of his emotional state may arise from a lack of motivation or engagement in the study, not fully understanding the Arousal and Valence concepts, or even the clip displayed encompassing complex emotional states. Lastly, at the bottom of the interface (Figure 3.1 - F) there is an optional box for the user to introduce further comments. After all the fields have been filled in, the next timestamp will be automatically selected for the user to annotate, changing the highlighted box at the top and loading a new clip in the video player.

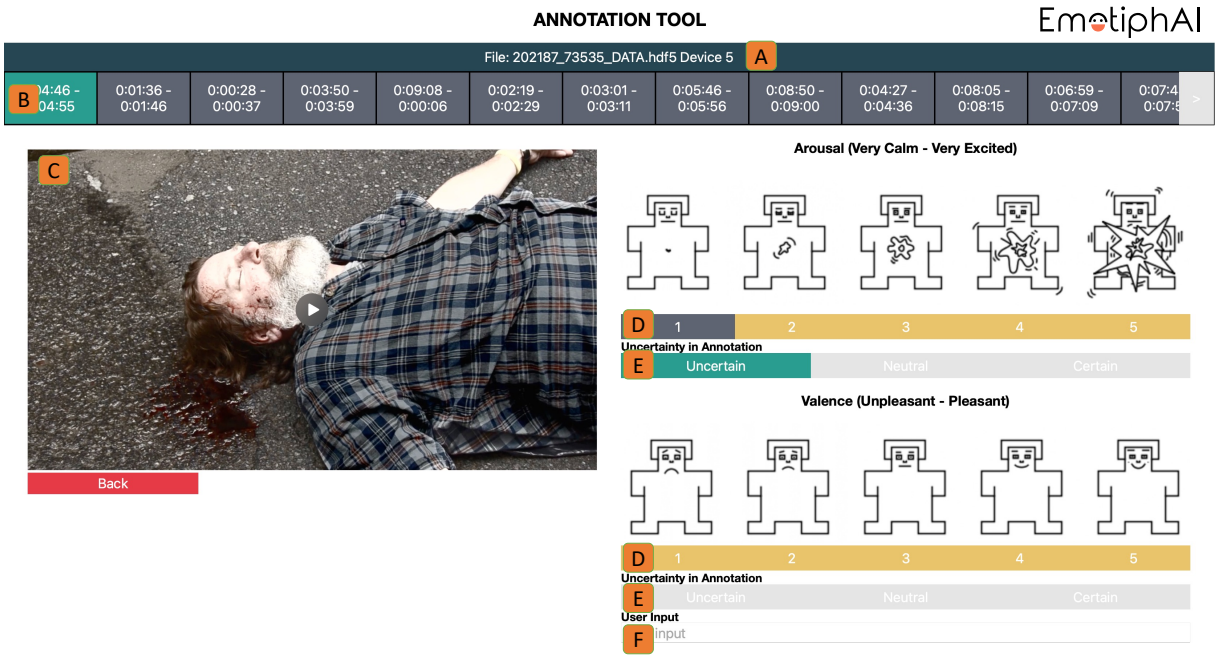


Figure 3.1: EmotiphAI self-reporting annotation interface [5].

Emotion Elicitation and Data Acquisition Protocol

The participants involved in these acquisitions were volunteers, older than 18 years old, without any known pathology. Participants were asked not to be under the effect of alcohol or medication before and during the experiment. Participants showing any physical or psychological impairment needed for the experiment, were not enrolled. The documents to be sign by the participants (e.g. informed consent document) as well as the experimental procedure were authorized by the ethics committee of the University under the process #1005890.

A selection of audiovisual elicitation content was performed based on the availability of the subjects subscribed platforms (e.g. *Netflix*, *HBO*, *HULU*), national TV programme (*RTP*) and the content genre (e.g. Comedy, Drama). The aim was to select novel long-duration content, with a duration longer than 40 minutes (μ=100.6 min, σ=32.6 min). The selected content consisted of recent uncalibrated movies that premiered in the last 3 years, thus approaching current topics. Furthermore, these contents covered 7 different genres to elicit a broad range of emotions.

Each experiment was realized with a group of between 9 and 4 participants (μ=5.1, σ=1.6). An assistant was present to ensure the protocol was properly followed; due to the COVID-19 pandemic, all experiments were performed following the sanitary guidelines from the national health authority (Direção Geral da Saúde - DGS). The trials were conducted in a familiar and comfortable environment for the participants to eliminate any bias, such as the stress of being in a new environment, and simulate as much as possible a real-world unconstrained scenario (Figure 3.2a).

Before the experiment, subjects were asked to sign an informed consent document, along with authorization to use their pseudo-anonymized data. Each participant was only enrolled in the experiment after agreeing to the terms in these documents. To ensure the data privacy of each participant, a pseudonym was assigned and the data collected was disassociated from the participant's private information (e.g. name). Furthermore, the documents signed by the participant were kept in a room with restricted access and the registered data maintained confidential, without the association of the participant identity, and password protected.

A *Raspberry Pi 4 Model B* was used as a set-top media center to display the elicitation content, run the EmotiphAI data acquisition and annotation software, and store the physiological data. This media center was connected to a LCD monitor (*SAMSUNG UE65TU7025* with 65" - 165 cm) to exhibit the elicitation content. Furthermore, this device also receives the data sent in real-time by the EDA acquisition devices, and stores it locally. For the data acquisition, each participant had one Xinhua Net FMCI device connected to their wrist or forearm, this device has two EDA electrodes, which were placed in the palm of the hand (according to Figure 3.2b) with Ag/AgCl electrodes. After the acquisition system was set up and the sensors have been placed, the devices were turned on and a data quality assessment was performed by the assistant, to see if every element was set up correctly. The Raspberry

(a) (b)

Figure 3.2: Experimental conditions of the acquisition process. (a) - Layout of the sitting area in relation to the LCD monitor. (b) - Device and electrode placement at the wrist and palm of the hand, respectively.

Pi embedded EmotiphAI software ensured the synchronization between the data acquisition and the video, by starting the two simultaneously.

After, the viewing of the movie, participants were asked to fill their self-assessment annotations regarding the content watched. The subjects' emotional annotation was performed using EmotiphAI's annotation tool using their mobile phones or a PC provided by the research assistant.

Both the EDA signals and the emotional annotations are stored in the same HDF5[1] file in a hierarchical format. For each user a HDF5 dataset is created, containing all the information acquired from this user i.e. the EDA signal and the user's annotations. The data from each source is saved in a different group, resulting in each dataset containing five groups (plus one optional group containing the user annotations comments): EDA signal; Arousal annotations; Arousal uncertainty; Valence annotations; Valence uncertainty.

[1] https://www.hdfgroup.org/solutions/hdf5/

4

Data Analysis

Contents

This chapter summarizes the emotional analysis performed on the collective data. Section 4.1 presents the reasoning which led to the evaluation of emotion in a group setting, namely the impact of being in a group settings. Section 3.2 comprises an overview of the state of the art in the scope of collective emotional assessment. Section 3.3 describes 2 different analysis: the first consists in evaluating the annotations given by the participants and the assessment of the similarities of synchronous annotations across different participants. The second analysis suggests a new approach to identify time regions where the participants and the audience reacted with higher intensity, based on EDA data and unsupervised machine learning techniques. Section 4.4 presents the results achieved with each analysis and Section 4.5 discusses such results in terms of the annotation performed by the participants and the evaluation of the content based on the EDA data

4.1 Motivation

Emotion recognition is a skill that is inherent to almost all human beings, so it is often taken for granted, although the same is not observed when replicated by a computer. In the filed of Affective computing, emotion recognition is still a great challenge. Previous work on emotion recognition focuses mainly on the analysis of emotion in an individual setting and in controlled environments, ignoring important dimensions. Hence, to evaluate emotions experienced by the subjects in a group setting, it is necessary to collect the data simultaneously from all participants.

Furthermore, the majority of studies applying machine learning techniques to perform emotion recognition use supervised algorithms. However, supervised learning techniques require each sample to be annotated with a ground-truth, in contrast to unsupervised learning (that does not require samples to be labelled). The collection of ground-truth data is a difficult process, especially the accurate labelling of emotional states, where attributing a label in a reasonable and scientifically-valid manner is even more complicated due to the subjectiveness in the appraisal and expression of emotion states. Most studies typically circumvent this issue by using short-duration (a few minutes) calibrated content to elicit specific emotions (see Section 2.5). A question is put upon whether a few minutes validated clip is able to elicit emotions similar to a real-life experience, when a longer content is appraised by the subject and a build up to each emotional state is experienced.

In the literature, the application of unsupervised learning algorithms to perform emotion recognition is considerably unexplored. Within the field of emotion recognition based on physiological signals, unsupervised learning is based mostly on EEG data, although it is still in a smaller scale when compared to the number of algorithms found for external expressions of emotions.

This part of the work seeks to analyze the emotional responses in a group setting when long-duration uncalibrated elicitation content is used. Being in a group environment can have an effect on the emotions

experienced at the time, as referred in Section 3.1, and the elements in a group are expected to react in a similar manner with regards to the elicitation content being watched, so the present work aims to analyze the similarities in simultaneous annotations across different participants, along with an analysis of the correspondent EDA signals dynamics. Furthermore, this work also suggests a new approach to identify time regions where the audience reacted with higher intensity on the EDA data based on unsupervised learning techniques.

4.2 State-of-the-Art

As seen in Chapter 2, the EDA signal is strongly related to the emotional states experienced by the subject, namely the Arousal dimension of emotions. Thus, studying this signal provides a bridge to understand human emotions. The authors in [62] performed a thorough review on innovations in the EDA signal processing, namely in signal decomposition tools. Out of all the decomposition tools analysed, the cvxEDA [63] stood out due to its low computational cost, good decomposition results and implementation in a Python environment. This algorithm describes the EDA as the sum of three components: the phasic component, the tonic component and an additive white Gaussian noise component (which incorporates the model predictions errors, measurement errors and artefacts). The algorithm is inspired by the physiological characteristics of the EDA and explains this signal based on Bayesian statistics, mathematical convex optimization, and sparsity. Furthermore, this method has also been widely used in many applications [44, 62], it is robust to noise, and overcomes the issue of overlapping EDR events (explained in Section 2.4), so it is considered to be a viable option for the analysis and decomposition of the EDA signal.

Feature Extraction

In the literature, the number of metrics used to analyse the EDA is highly diverse, varying from study to study [64, 65]. In [66] the author extracts 8 features, mainly temporal and statistical, from the EDA signal (peak to rise time sum, peak amplitude sum, half-recovery sum, peak energy sum, rise rate average, decay rate average, percentage decay, and number of peaks), coupled with data from other physiological signals, achieving an accuracy of 89.23% in detecting the stress level of drivers when using Layer Recurrent Neural Networks algorithms. Recently, Martinez et al. [40] proposed a new feature which is the surface area comprising the difference between the EDA signal and its linear regression. According to the authors, this feature has a low computational load while providing great physiological significance. The value of this feature will be smaller if fewer EDR events take place, thus the higher the value the more EDR events took place, highlighting significant ANS activation. Martinez et al. [40] achieved an accuracy between 64.9 and 99.1% in detecting the stress level of participants when using

different supervised machine learning algorithms.

Regarding the analysis of emotion based on EDA data, in the work developed by Li et al. [42], the authors establish a temporal correspondence between a weighted mean of the EDA signal and a continuous arousal annotation. This correspondence was achieved by simultaneously acquiring EDA data and a continuous arousal annotation from 13 participants, throughout the duration of the elicitation content (with close to 2 hours duration). These results validate the correspondence between the EDA signal and the arousal dimension of emotions.

Fleureau et al. [45] evaluated the audience reaction during a regular cinema show solely based on EDA data collected simultaneously across all participant in the audience. For this goal, first the Individual Affective Profile was calculated by pre-processing the EDA signal, truncating its derivative to positive values to highlight relevant phasic changes and normalizing the resulting signal. However, this measurement only reflects the individual reaction, leading to an evaluation of the audience reaction based on the Mean Affective Profile (MAP) computed by averaging the Individual Affective Profiles of every audience member. The resulting MAP proved to be an effective method of evaluating the audience arousal reaction to an elicitation content.

In terms of the application of unsupervised learning techniques in emotion recognition, the major use of such methods are applied using external demonstration of emotion such as speech and facial expressions. These are the cases of the works developed by Eskimez et al. [67] and Huelle et al. [68], which use speech and facial queues, respectively. Regarding the use of other physiological signals, Lakhan et al. [41] applies clustering algorithms (a type of unsupervised learning technique) to EEG signals. The EEG data was acquired while the participants watched different movie trailers, and the clustering algorithms were applied to features extracted from this signal, with the goal of grouping the movie trailers which elicit similar emotional reactions. The algorithms applied were two standard clustering methods, namely K-means and Gaussian Mixture Model, achieving accuracy scores between 0.63 and 0.70 when predicting whether the elicited signals were associated with a high or low level of valence and arousal. In a similar approach, Zhang et al. [69] use the K-means algorithm to group similar individual reactions to different sound stimuli based on EMG and HR data. The results obtained were able to successfully group the emotional states into groups of high and low valence when compared to the emotional annotations given by the participants, confirming K-means as a reliable approach to analyse emotional reactions to different stimuli.

For this part of the work, we analysed the data according to two different methodologies. The first approach consisted in using the participants' self-reported annotations; this has three objectives: 1) Evaluate the annotations throughout the content, in terms of number and values, to identify potential "hotspots" of the content; 2) Evaluate the similarities in temporally related annotations in terms of EDA events and annotation values; and 3) Establish correspondence between these synchronous annota-

tions and the elicitation which triggered them. The second approach seeks to expand the state-of-the-art, by using EDA data to determine the time regions where the audience reacted with higher intensity, based on clustering algorithms applied to features extracted from this signal, and grouping the time region where the audience reacted in a similar manner.

Clustering Algorithms

Clustering is a task that consists in grouping a set of objects so that each group contains objects that are more similar to each other than to those in other groups. In the literature there are several different clustering algorithms, which can differ significantly in their definition of what constitutes a cluster and how to efficiently divide them, thus they often produce different outputs for the same data [70]. This work focuses on the application of hierarchical clustering, namely agglomerative linkage hierarchical algorithms and K-means algorithms.

Hierarchical algorithms can be divided into two different types, agglomerative and divisive [71]. Agglomerative clustering consists of a bottom-up approach, in which a cluster is created for each individual sample and, in iterated steps, clusters separated by the shortest distance are combined. Divisive clustering is the opposite; all observations start in the same group and, with each step, the groups are split into subsets of clusters [72]. Agglomerative algorithms can be further divided into two groups. The first group consists in the linkage methods, in which no additional information is used besides the input data. On the other hand, in the second group of algorithms, it is necessary to specify the center point of each cluster - centroid; samples are grouped in clusters according to the proximity to the closest centroid. [72]. Regarding the stopping criteria, these algorithms stop merging clusters when a predefined number of clusters is reached. The number of clusters can be determined using different methods, some of which are: manually and automatically determined using (e.g.) the life-time criteria [70].

In the current work, four different hierarchical linkage methods are applied: single linkage, complete linkage, average linkage and ward linkage. In the single linkage method the distance between two clusters is determined by the minimum distance between all observations of the two sets [72]. An alternative approach is the complete linkage, in which the distance between two clusters is determined by the maximum distance between all observations of the two sets [72]. Average linkage uses the mean distance between each clusters based on the mean distances of each observation of the two sets [70]. Lastly, the ward linkage minimizes the variance of the clusters being merged [70, 72].

The K-means algorithm clusters data into groups with equal variance by minimizing the inertia of each cluster or the within-cluster sum-of-squares. In this algorithm the number of clusters, X, has to be specified since it initiates by determining X centroids, one per cluster (these centroids can be updates throughout the clustering process) [73].

4.3 Proposed Methodology

This section is subdivided into 3 distinct subsections. The first subsection describes the pre-processing steps taken regarding the EDA signal, namely, the filters applied, signal decomposition and fiducial points detection. The second subsection presents the methodology used in evaluating the annotations given by the participants. Lastly, the third subsection suggests a new approach to identify time regions where the participants and the audience reacted with higher intensity.

Signal Pre-processing

Data processing was performed on a Python 3 environment, with the support of the BioSPPy(version 2) toolbox [19], a publicly available set of signal processing tools to analyse biosignals. The first step in the pre-processing of the EDA was outlier removal and manual selection of which signals/participants to use. A function was developed to detected and remove the outliers of an EDA signal. This function is based on two criteria: the amplitude difference with the mean of the signal and the derivative of the signal. If a signal point has a large amplitude difference in relation to the mean of the signal and a large derivative, it is considered to be an outlier and it is removed. To maintain the signal characteristics after the outliers are removed, the signal is interpolated using the original sampling frequency with a cubic spline interpolation (to replace the eliminated data points)[1]. The function also provides a confidence level, which indicates the quality of the signal, i.e., if the signal has too many outliers the confidence level is lower indicating low quality. The exclusion criteria for the manual selection was based on the overall quality of the signals, that is, saturated signals, interruptions amidst the acquisition, and signals with a constant amplitude were removed. Figure 4.1 shows two examples of excluded signals, where it is possible to see that the data contains a great number of outliers, possibly from disconnections mid-acquisition.

The EDA signal was interpolated to 10 Hz using a cubic spline interpolation. Afterwards, the signal was filtered with a 4th order low pass Butterworth filter with a 1 Hz cutoff frequency. Following the filter, the signal was smoothed using a 10 point moving average following the approach described in [74]. After these procedures, the signals were normalized per subject so that its range is between zero and one [2].

As seen in Section 2.4, the EDA signal can be decomposed into 2 different components: the phasic component EDR, and tonic component EDL. So, to decompose the EDA into these components the *cvxEDA* algorithm was applied. In Figure 4.2a it is possible to see the decomposition of the EDA signal into its components. From this image it is possible to see that the EDA signal corresponds to the sum of its two components, EDR and EDL. Moreover, as discussed in Section 2.4, the EDL component is

[1] This was achieved using the scikit-learn toolbox which is publicly available
[2] *minmax _scale* function of the scikit-learn tool

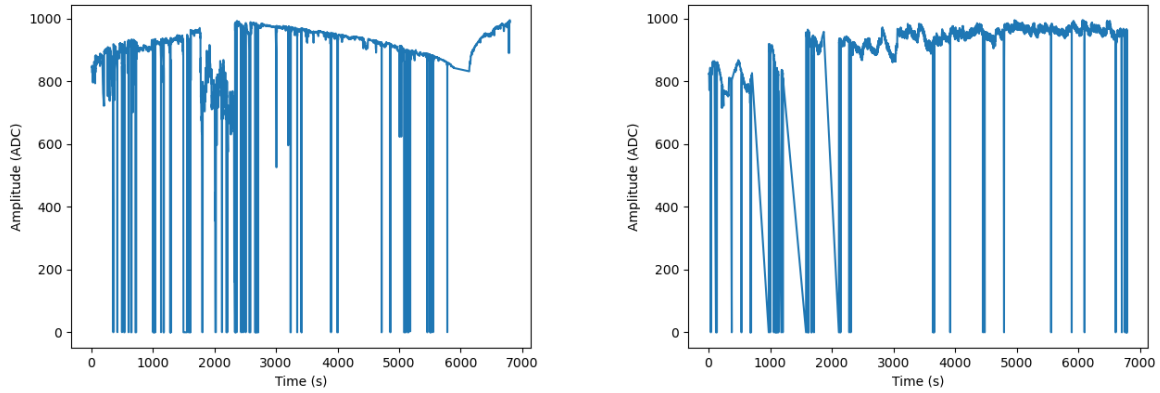
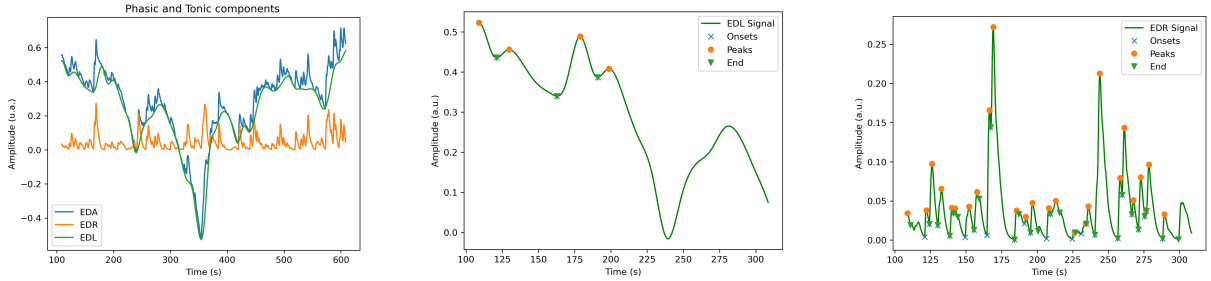


Figure 4.1: Example of two EDA signals excluded from the analysis.



(a) **(b)** **(c)**

Figure 4.2: (a) - Example of the EDA data decomposition into EDR and EDL components using the *cvxEDA* algorithm. (b & c) - Example of the detected fiducial points for the EDL, (b), and for the EDR, (c).

characterized for being a slowly changing signal, while the EDR is distinguished for representing short-lasting changes which are usually a response to a stimuli. These characteristics of the EDA components are observed in Figure 4.2a, where the EDL, in green, represents a slower varying signal with the baseline of the EDA, while the EDR contains the rapid changes of the EDA associated responses to different stimulus.

Finally, the identification of the fiducial points was achieved based on the method proposed in [75], which has the advantage of not requiring any type of threshold. The algorithm returns the onset, peak and end of each event (the end point corresponds to the 63% recovery time, or in cases this point is not reached before the next event it is the same as the next event onset point). The detected fiducial points for the EDL, are represented in Figure 4.2b, and the EDR, are represented in Figure 4.2c. It can be seen that the EDR signal has a greater number and smaller events than the EDL.

Analysis of the Synchrony between Annotations

A first analysis was performed of the self-reporting performed by the volunteers upon watching a movie. Namely, the potential synchrony between time regions in which each participant performed an annotation, throughout the duration of the movie. The total number of annotations across all the participants throughout the duration of the movie was obtained. The total number of annotations was achieved by summing the number of annotations performed by different participants in 1-second windows. Figure 4.3 gives an example on how these metrics were achieved. The red and blue lines represent the number of annotations throughout the duration of the movie, during the first 2 seconds only one annotation from Participant 1 is available. During the third second, annotations from Participants 1 and 2 are available, thus accounting for a total of 2 annotations for this time instant. Lastly, for the remaining of the duration (seconds four and five) only the annotation from Participant 2 was available. The same process was followed throughout the duration of the movie to determine the number of participants annotating in each time frame.

Afterwards, the annotations were evaluated in qualitative terms, thus a histogram was plotted with the density of each annotation value for the Valence and Arousal dimensions. Each histogram had 5 columns, one per each annotation value from 1 to 5, with the total number of annotations performed with those values; e.g. if there were 5 annotations with a level 1 Arousal, column 1 in the Arousal histogram would have an amplitude of 5. The annotations were also evaluated throughout the movie, this was achieved by plotting the mean value of each dimension (Valence and Arousal) in each 1 second time window.

A further step was to analyse simultaneous annotations, which consist in 2 or more annotations which overlap time wise. Simultaneous annotations are illustrated in Figure 4.3; the two annotations displayed overlap during one second. Although the annotations from Participant 1 and 2 represented in light blue and green, respectively, do not start and end at the same time, there is a period in which they overlap, so, in this analysis they are considered simultaneous. Afterwards, the EDA data was concatenated for the entire duration of the time window considered synchronous. In the example demonstrated in Figure 4.3 EDA data from Participants 1 and 2 would be joined from the start of the annotation of Participant 1 until the end of the annotation of Participant 2. The EDA signal from the remaining participants (those who did not annotate in that time region) was also concatenated into a different group, for the entire duration of the time window considered synchronous. Furthermore, the mean EDA signal was calculated, along with the STD for simultaneous annotations periods. Likewise, the procedure was implemented for the participants with no annotations during the timestamp. The movie clip for each simultaneous time period was extracted in order to establish correspondence between the simultaneous annotations and the possible elicitation stimulus. With the goal of comparing different simultaneous annotations, the annotations values, along with the number of EDR events and the Pearson Correlation

Coefficient (PCC) between the EDA signals of each participant involved in the simultaneous annotations were acquired.

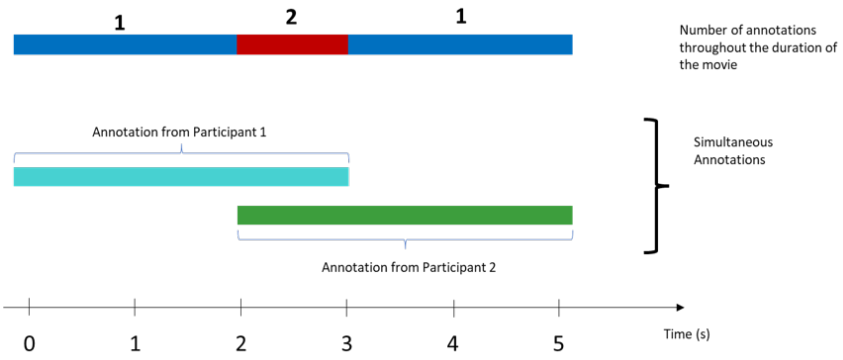


Figure 4.3: Example of the number of annotation in each time instant and illustration of a simultaneous annotation across two participants

Lastly, the movie clips during the time periods where no annotation was performed by any participant were extracted, to evaluate the hypothesis if these regions correspond to uneventful parts of the movie.

Collective Intelligence Analysis

This analysis focuses on determining the time regions where the audience reacted with higher intensity. To this end, the first step is to extract representative features from the EDA signal. A wide range of features are observed in the literature [7], from which 13 were selected based on the information provided by each feature for emotion recognition. The features extracted were: Number of EDR onsets; Number of EDL onsets; Area under EDR events; Sum of the startle magnitudes; Sum of EDR event amplitudes; Sum of EDL event amplitudes; Sum of the response duration; Sum of the onset-peak times; Mean EDR; STD EDR; Dynamic range; Mean of the EDA derivative; Surface area between the EDA and its linear regression. These features were selected based on a trade-off between the number of features and the information provided by each individual feature.

These features were extracted from each participant's EDA signal and a group EDA, thus 2 different analysis were perform, a group analysis to see where the audience as a whole reacted, and an individual analysis to see where each participant reacted. The group EDA signal was calculated by determining the mean EDA across all participants using a moving window with a length of 3 seconds and an overlap of 2 seconds. Note that the participant's EDA signals used to calculate the group EDA signal were previously normalized across all participants. The resulting group EDA signal has a 1 second resolution, which represented the mean EDA signal of the group. For feature extraction the signals were divided into windows, in the literature, it was observed that windows should be between 10 to 300s [40]. In the current work, a window size of 20s with an overlap of 5s was used. Given that the average movie

scene duration has between 1 and 3 minutes, a window smaller than this would be too granular and may not contain sufficient information for emotion recognition, while a longer window could encompass very different reactions, compromising the results. For each window, the aforementioned 13 features were extracted. The feature vectors were given as input for clustering algorithms. The clustering algorithms were applied to group the periods of the movie in which the participant reacted similarly, or in the case of the group EDA features, the periods of the movie where the audience reacted similarly. The number of clusters in the hierarchical clustering was determined using the life-time criteria, while for the K-means several number of clusters were tested as input to the algorithm. In particular, the number of clusters was increased from 2 to 8 until there was a some distinction in the movie scenes in each clusters. From the resulting clusters, the corresponding movie clips were extracted to analyse the scenes which triggered such reactions. This analysis consisted in counting the number of clips in each cluster, their total duration along with a visualization of the clips to see if there were any similarities between them, and check if there was an emotional context behind such scenes. Assuming the premise that each emotional scene triggers an emotional response and that neutral scenes do not trigger any emotional response. This would mean that, ideally, every scene in clusters that only contain strong or emotional clips, triggered an emotional reaction, and all the emotional reactions elicited by the clips in that cluster were similar, thus being in the same cluster. Note that this method is correlated with the intensity of the users' emotional reaction (Arousal), and less correlated to how positive or negative an emotion is (Valence). The current work relies on the EDA, which is strongly related to the Arousal dimension.

In a third approach, the MAP was determined using the methods described in the work of Fleureau et al. [45]. The MAP is a validated methodology in the literature that reflects the arousal variations of a global audience during a movie. This measurement was used as a ground truth to evaluate the performance of the clustering methods. To derive this value, the first step, after preprocessing the EDA data, is to truncate its derivative to positive values. Afterwards, the mean of the truncated derivative is calculated in each 20 second window (with a 5 second overlap), and the signal is normalized (area under the curve equal to one). The resulting signal for each individual is titled "Individual Affective Profile". Finally, the MAP is calculated by averaging the individual affective profiles of every participant. For further details in this process, we refer the reader to [45].

4.4 Results

This section is subdivided into 3 distinct subsections. The first subsection describes the dataset used. The second subsection presents the results achieved in the first analysis, i.e. evaluation of the the annotations given by the participants. Lastly, the third subsection displays the the results achieved in the second analysis, i.e. a approach to identify time regions where the participants and the audience

reacted with higher intensity.

Data characteristics

The current work focuses on the analysis of the data collected using as elicitation content the movie "Spider-Man: Far From Home" directed by Jon Watts and written by Chris McKenna, Erik Sommers and Stan Lee. Due to the pandemic situation in which this work was developed it was hard to gather volunteers for the experiment leading to several delays in the acquisition process, thus only one movie was analysed. The movie at hand had a total duration of 1 hour and 55 minutes (6900 seconds), and it is classified on IMDB as an Action, Adventure and Sci-Fi film[3]. The movie tells a story about Spider-Man, Peter, during his school vacation in Europe, when a new villain appears disguised as a superhero to fight the elemental beasts that emerged in some of the European capitals. Data was acquired from 7 volunteers, from whom 2 where male; the average age of the participants was 20 years old, with a STD of 0.7.

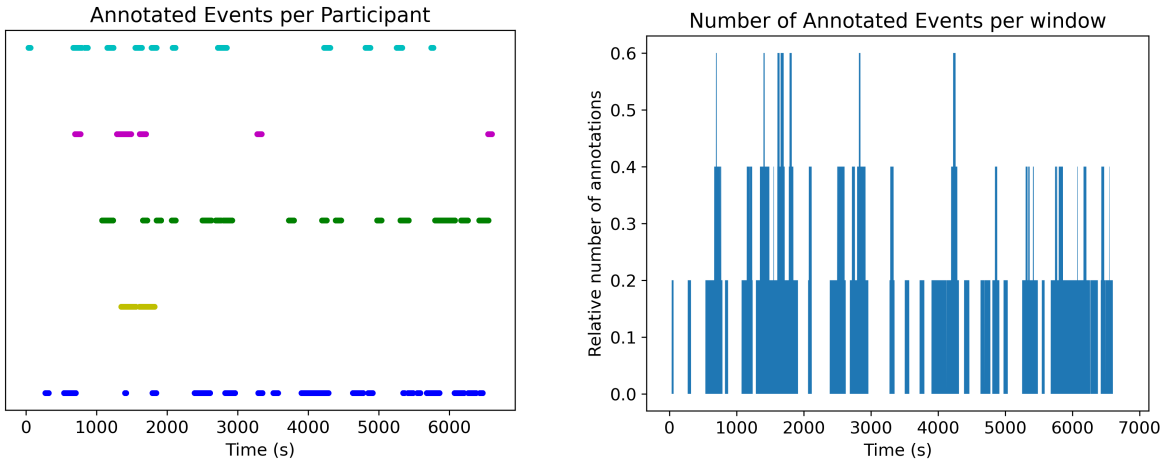
Participants 0 and 5 were excluded from the analysis, due to poor EDA signal quality, as per the criteria defined in Section 4.3. Lastly, on account of the low number of participants, a specific colour was attributed to each one, thus all plots with the same colour correspond to the same participant, from now on. This colour code can be seen in Figure 4.4

Figure 4.4: Participant's colour code

Analysis of the Synchrony between Annotations

In Figure 4.5a it is possible to observe the annotation performed by each participant throughout the duration of the movie. As expected, not all participants performed the same number of annotations, with Participant 2 only having annotated in a small period, while the annotations from Participant 1 are more spread throughout the movie. On the other hand, Figure 4.5b displays the total number of annotations performed by all participants throughout the movie. As it can be seen, the number of annotations fluctuates a lot with time, existing many and long periods with no annotations, and few and short periods having up to 60% of the audience annotating. The total number of annotations was 61, corresponding to a total of 5375 annotated seconds, and 2895 seconds (42.0%) of film without any

[3]https://www.imdb.com/title/tt6320628/



(a) Time regions in which each participant annotated.

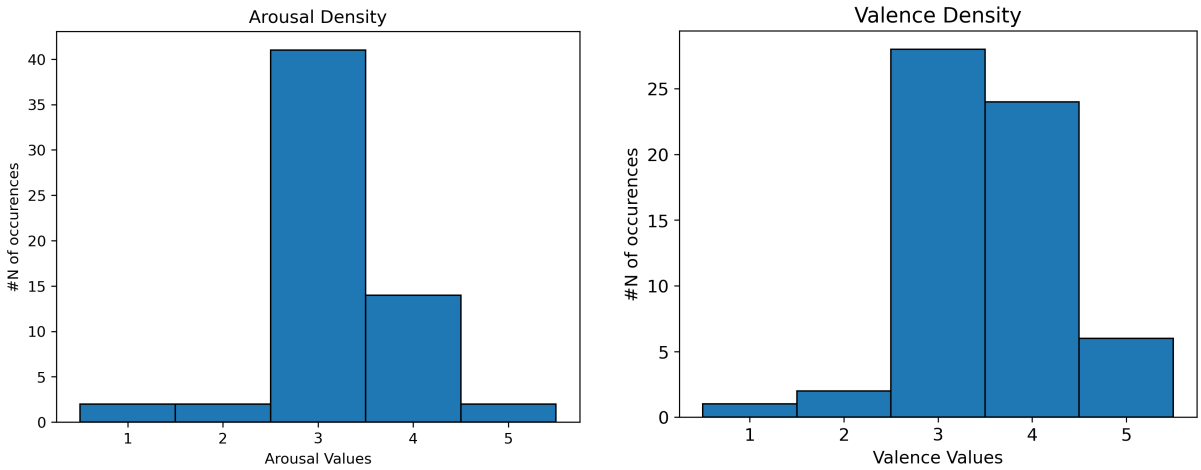
(b) Total number of annotations throughout the movie duration.

Figure 4.5: Temporal distribution of the annotations.

annotation. Furthermore, out of the 61 annotations, 39.3% (or 24 annotations) corresponded to a neutral state of Valence and Arousal, with both dimensions having a value of 3.

In Figure 4.6 it is possible to see a qualitative representation of the annotations performed by the audience. Figures 4.6a and 4.6b display a density histogram with the total number of annotations per value, for the Arousal and Valence dimensions, respectively. From these figures, it is possible to observe that most events were annotated with a 3 (especially in the arousal dimension), on a scale from 1 to 5, with a slight tendency for higher values. These annotations mostly represent neutral emotional states, with slight variation throughout the whole duration of the movie.

Figure 4.7 is an example of one of the simultaneous annotations; this particular one occurred across Participants 1, 2 and 4. This simultaneous annotation was chosen to be represented because it occurs during the longest peak in Figure 4.5b (this peak is the second maximum in 1000-2000 second time range), so it should correspond to a higher intensity emotional part of the movie, where the audience was more in tune. The annotation pair Valence-Arousal given by each participant during this period were: (3,3); (3,1) and (3,4), respectively. With regards to the scene itself, it displays a fight scene between Spider-Man and a water monster in the city of Venice, when a new superhero, Mysterio, appears for the first time. In Figure 4.7a is possible to see the individual EDA signals of the 3 participants who annotated in this time region. Although these annotations were considered to be simultaneous, they may not start and end at the same time, so the vertical lines present the beginning and end of each annotation, i.e. the first discontinuous blue vertical line represents the beginning of the event annotated by the Participant 1, and the second discontinuous blue vertical line represents the end of this event. Figure 4.7b displays the individual EDA signals of the participant who did not annotate. Lastly, Figure 4.7c displays the mean EDA in blue and the mean EDA $\pm$ STD in shaded color for the participants who annotated, and



(a) (b)

Figure 4.6: Density histograms with the total number of annotations per value for the Arousal (a) and Valence (b) dimensions.

Figure 4.7d represents the same but for the remaining participants.

To establish a comparison between several simultaneous annotations, Table 4.1 displays 8 representative simultaneous annotations, out of the 20 total simultaneous annotations, along with the annotation values of each participant involved, the number of EDR events during that time period and the PCC between the EDA signal of each participants. In addition to this, the table also contains a brief description of the movie scenes associated with each simultaneous annotation represented.

Table 4.1: Comparison of the annotations and the EDA signals of the participants involved in each simultaneous annotations, along with a description of the correspondent movie scene

Participants	Annotations (V,A)			# EDR events			PCC			Scene description
1,4,6	1 2,2	4 4,4	6 4,3	1 16	4 16	6 15	1-4 0,0424	1-6 0,6203	4-6 0,4707	Spider man is stressed with complicated questions and funny scene with may
1,2,4	1 3,3	2 3,1	4 3,4	1 51	2 94	4 18	1-2 0,2027	1-4 0,4841	2-4 0,3826	Intense Fight Scene
1,2,6	1 3,3	2 3,3	6 4,2	1 49	2 15	6 29	1-2 0,2679	1-6 0,0529	2-6 0,1913	Funny scenes and jokes / talk between Spider man and Fury
1,3	1 3,3	3 4,5		1 42	3 22		1-3 0,154			Peter almost mistakenly kills a friend
1,3,6	1 3,3	3 5,4	6 4,3	1 57	3 34	6 40	1-3 0,6982	1-6 0,6492	3-6 0,8703	Peter is scared for his friends safety
1,3,6	1 3,3	3 3,3	6 4,3	1 34	3 23	6 21	1-3 0,5624	1-6 0,71	3-6 0,2504	MJ finds out about Peter being Spider-Man
1,4	1 3,3	4 3,3		1 14	4 4		1-4 0,6265			Fight scene with lava monster
2,4,6	2 3,1	4 3,4	6 4,3	2 56	4 26	6 35	2-4 0,3578	2-6 0,1482	4-6 0,468	Intense Fight Scene with water monster

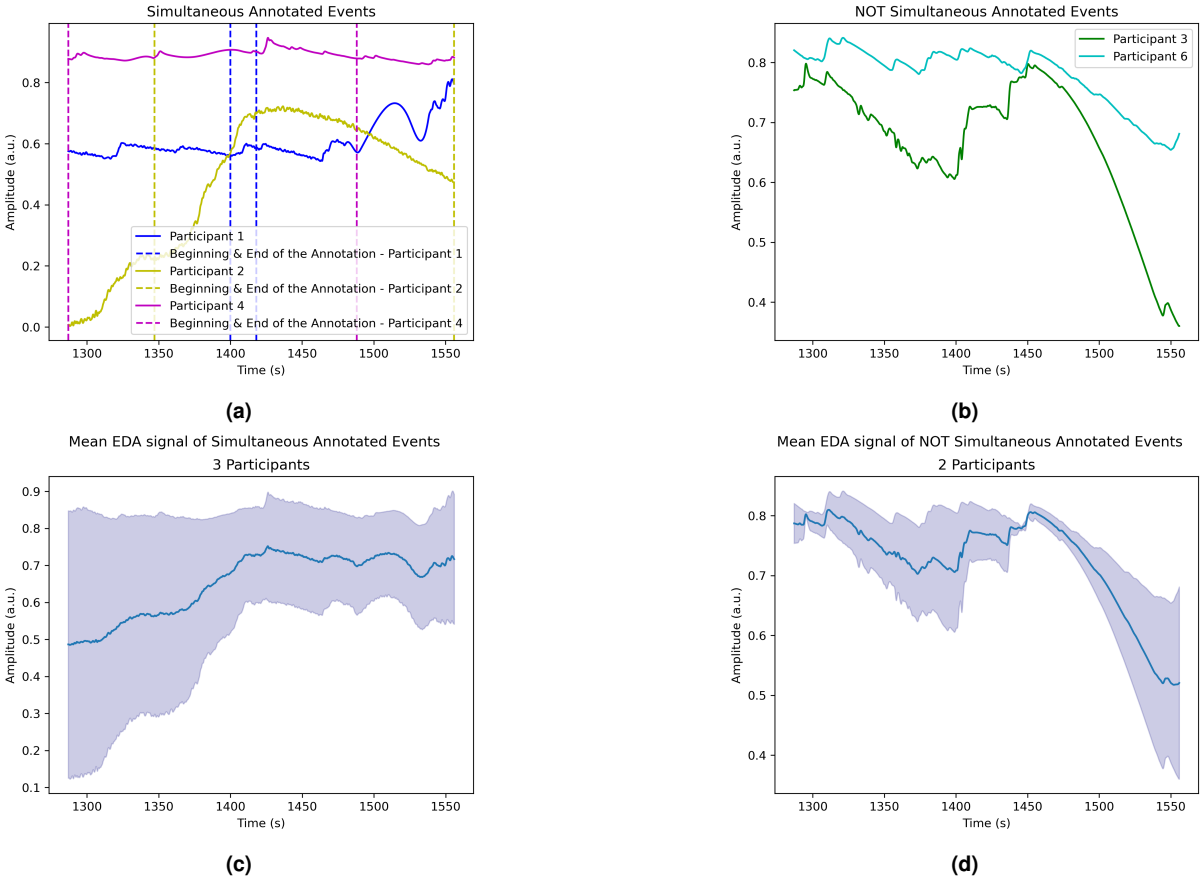


Figure 4.7: Representation of simultaneous annotations across Participants 1, 2 and 4. (a) - EDA signals of each participant that annotated in this time period, where the vertical lines represent the beginning and end of the annotations of each participant; (b) - EDA signals of each participant that did not annotate in this time period; (c & d) - Mean EDA signal in blue and the mean EDA $\pm$ STD in shaded color, for those who annotated (c) and those who did not annotate (d).

Collective Intelligence Analysis

In Figure 4.8 it is possible to see the MAP, where each point represents the mean arousal of the audience at that instant. The mean MAP throughout the whole duration was 9.64×10^{-5} with a STD of 1.33×10^{-4}. This image also contains a representation of the most relevant scenes of the movie, the shaded light blue areas represent the periods in which such scenes occur identified from (a) to (k). The description of the scenes associated with each area is: (a) - Spider-Man in an awkward situation, funny scene; (b) - Spider-Man gets stressed out with an interview, emotional scene; (c) - Peter and his friends are on a plane telling several jokes, funny scene; (d) - Spider-Man and Mysterio fight a water monster, action scene; (e) - Aunt May has a boyfriend, funny scene; (f) - Spider-Man and Mysterio talk about the past, emotional scene; (g) - Spider-Man and Mysterio fight a lava monster, action scene; (h) - Revelation that Mysterio is a villain, emotional scene; (i) - MJ finds out that Peter is Spider-Man, emotional scene; (j) - Mysterio tricks Spider-Man and attacks him, action scene and (k) - Final fight scene between Spider-Man and Mysterio, action/emotional scene. These scenes and their description were obtained based on the visualization of the movie by one annotator. The most relevant scenes of the movie are consistently located around the peaks of the MAP.

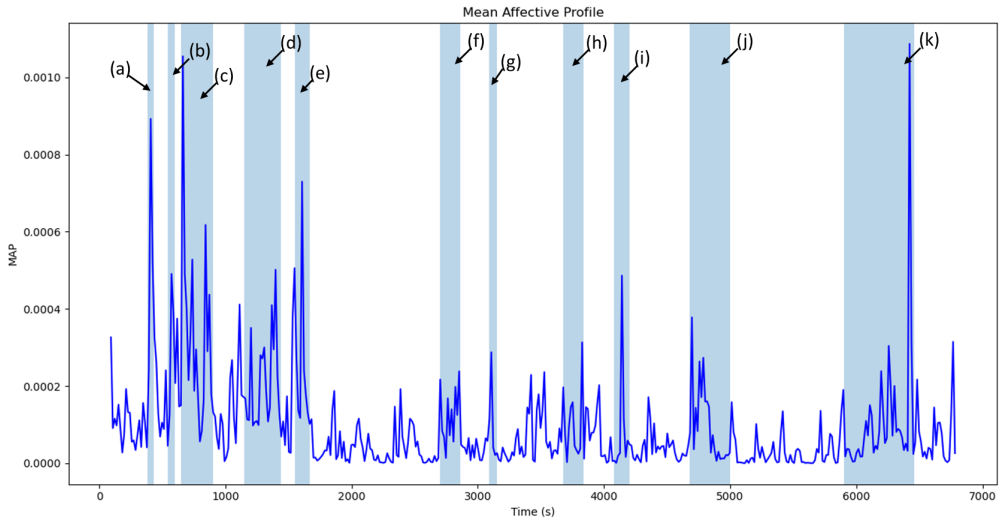


Figure 4.8: MAP calculated throughout the duration of the movie, along with a description and location of the most relevant scenes of the movie [6].

Figure 4.9 displays the group EDA signal of the audience, achieved by averaging the EDA across all participants. It can be seen that the signal shows prominent variations alternating between a period of high and low amplitude throughout the movie. Table 4.2 describes the results achieved with each clustering algorithm in terms of the number of clusters, number of clips, total duration of the clips in each group, mean MAP and STD MAP. These clusters were achieved based on features extracted from the

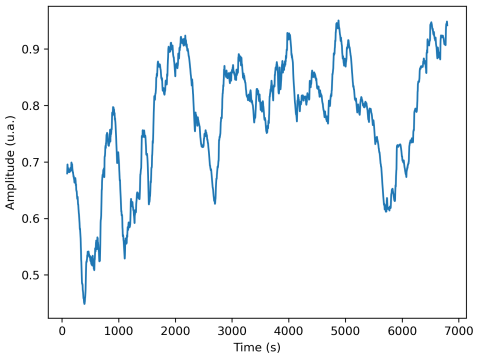


Figure 4.9: Group EDA signal.

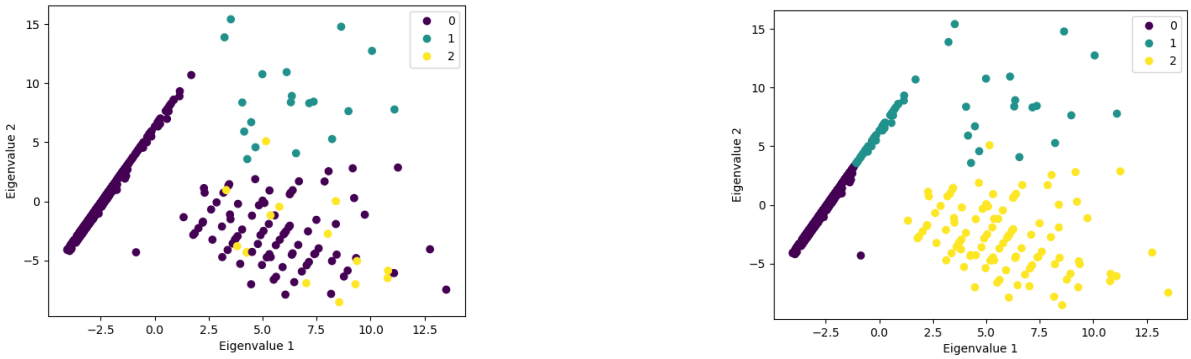
Table 4.2: Table with the characteristics of the group video clips achieved using different clustering algorithms with the group EDA signal.

Clustering Algorithm	Cluster	Counts	Length (s)	Mean MAP (E-04)	STD MAP (E-04)
Hierarchical Average Linkage	0	22	6327	0,90	1,88
	1	7	327	4,84	2,27
	2	15	315	1,73	1,27
Hierarchical Single Linkage	0	16	6576	1,57	1,26
	1	15	315	1,87	2,00
Hierarchical Complete Linkage	0	9	6414	0,95	1,86
	1	8	393	4,40	2,41
Hierarchical Ward Linkage	0	84	4614	1,25	1,16
	1	22	1182	2,38	2,09
	2	79	2019	1,03	1,75
K-Means	0	14	324	1,41	1,93
	1	74	2364	1,96	2,09
	2	15	315	0,62	0,76
	3	36	801	1,12	0,75
	4	26	891	1,21	1,19
	5	54	2559	3,37	2,05
	6	33	753	1,40	1,51
	7	5	240	2,63	2,58

group EDA signal, thus the resulting clusters should reflect the periods where the audience had a similar reaction. The characteristics of the clusters achieved using the features extracted from each individual vary a lot from participant to participant, thus it can be difficult to extract conclusive results from such analysis.

To evaluate the resulting clusters from the group EDA, Figure 4.10a displays a plot of the clusters in a two dimensional space achieved using a Principal Component Analysis (PCA) on the original 13 features extracted from the EDA data. This image displays the results achieved with the hierarchical clustering with average linkage and the hierarchical clustering with ward linkage, since these were considered to be the best and worst results obtained, respectively.

Lastly, in Figure 4.11a it is possible to see the MAP, in blue, and the time distribution of clusters 1 and 2 in the teal and yellow vertical lines, respectively (the areas which do not have any teal or yellow



(a) (b)

Figure 4.10: Plot of the clusters obtained with different clustering methods using the group EDA signal, in a feature corresponding to the 2 eigenvectors obtained by a PCA of the 13 dimension feature space. (a) - Hierarchical clustering with average linkage and (b) - Hierarchical clustering with ward linkage.

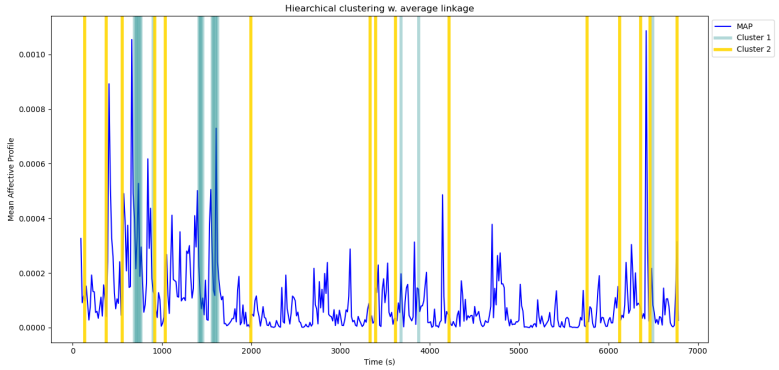
vertical lines correspond to the time distribution of cluster 0, which is the longest one). In each subfigure a different clustering algorithm was applied; the algorithms used were the hierarchical clustering with average linkage Figure 4.11a and the hierarchical clustering with ward linkage Figure 4.11b, which correspond to the best and worst results obtained, respectively. The cluster data represents the time frames that were classified into each cluster, thus a single blue point represents a 20 second time window of the movie, which was grouped according to the audience reaction into that designated cluster.

4.5 Discussion

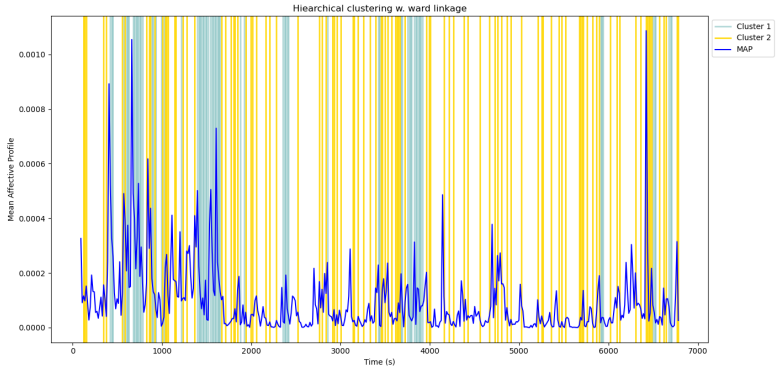
This section is subdivided into 2 distinct subsections, one per each analysis performed. The first subsection evaluates the annotations given by the participants and the second subsection evaluates the performance of the new approach suggested.

Analysis of the Synchrony between Annotations

This analysis focuses on the annotations performed by each participant regarding the Spider-Man movie, however, each person is different, thus the level of engagement of each participant and liking of the movie and/or the annotation task was different. This can be seen with Figure 4.5a in the olive color; the number and total duration of the annotations performed by Participant 2 was significantly smaller than the remaining participants, with Participant 4 (in magenta) also having a considerably smaller number and duration of annotation than Participants 1, 3 and 6. The reduced number of annotations by some participants can be a result of a lack of engagement in the annotation task, this phase can be quite monotonous since it requires a replay of several clips and their emotional annotation and, since the participants of this study were volunteers they could leave without annotating all the suggested clips.



(a)



(b)

Figure 4.11: MAP, in blue, and the time distribution of the clusters obtained with different clustering methods using the group EDA signal, namely, (a) - Hierarchical clustering with average linkage and (b) - Hierarchical clustering with ward linkage.

Regarding the number of annotations performed throughout the whole duration of the movie (Figure 4.5b), there were parts of the movie without any annotation, which can be expected since the participants were only performing annotations of the most relevant parts of the content. In terms of the period in which there were annotations, the number of annotators for a given timestamp varies considerably, since participants react differently to the same elicitation. As such, a clip that triggers an emotional reaction in one participant may not trigger an emotional reaction in another participant. Even when clips trigger an emotional reaction across several participants there could be a delay between emotional responses. Furthermore, it is to be expected that the periods in which the number of annotators is higher correspond to parts of the film with higher emotional content, since they triggered an emotional reaction in a greater percentage of the audience.

In terms of the values of the annotations, 39.3% corresponded to neutral states of Valence and Arousal and, as it is possible to see in Figure 4.6a and Figure 4.6b, the main annotation value in these dimensions was a 3, which is a neutral emotional state. These values were not expected since the movie consisted of a high-pass superhero movie, with several fight scenes, along with some emotional and comical parts. One would expect that the annotations of the relevant parts of the film were mainly positive. Predominantly in the Arousal dimension, 65.6% of the annotations values were 3, although, since this dimension measures the intensity of emotions, in an action movie such as the Spider-Man, with very intense fight scenes, plot twists and comical scenes it was expected that the emotions elicited would be more intense, as the film suggests.

These annotations suggest that a simple, unmeaningful conversation between two characters would elicit an emotion as intense as a fight scene where Spider-Man almost dies, or a comical scene when Peter gets caught by a friend in an awkward situation. The predominant Valence annotation value was also a 3, though in this case, it is clearly seen that these annotations tend to higher values, which describe a positive emotion as expected in these movies[4]. Nevertheless, in both dimensions, the number of annotations with values of 1 and 2 (extreme values) were very low (as expected), since these describe negative emotions, such as angry or scared, and inactive emotions, such as boredom or sadness, which are emotions that these kinds of movies do not aim to elicit in their audiences.

The simultaneous annotation, performed by Participants 1, 2 and 4, displayed in Figure 4.7, is a representative example of the simultaneous annotations, since the signal trends that can be seen in this example are also observed in most of simultaneous acquisitions data. From Figure 4.7a it is possible to observe that the EDA signals of different volunteers share very few similarities, however, the mean EDA signal displayed in Figure 4.7c shows a tendency to increase in amplitude. Similarly, in Figure 4.7b it is possible to see that the EDA signals do not share a lot of similarities, although the mean EDA signal (Figure 4.7b) displays a downwards tendency in amplitude. As mentioned above, these EDA trends

[4] In contrast to a horror movie where it should be expected that the emotion elicited would be mainly negative

are observed across the majority of the simultaneous annotations. In terms of the elicitation clip, this period corresponded to an intense fight scene between Spider-Man and a water monster, when the new character, Mysterio, first appears to fight side by side with Spider-Man. However, the Arousal annotations given by these 3 participants were all neutral, thus not corresponding to the intensity possibly projected by the director of the scene at hand. The Valence annotations ranged from 1 to 4, thus describing a broad range of elicited emotions from very negative to positive for the same video clip.

Regarding the global evaluation of the simultaneous annotations, in Table 4.1 it is possible to see that the PCCs (ranged between -1 and 1) are all close to 0, suggesting that the EDA signals in simultaneous annotations have small correlation between them. In fact, the mean PCC between EDA signals of simultaneous annotations is 0.45 with a STD of 0.24, which implies a low average correlation in these signals. Furthermore, the number EDR events detected in these periods can be quite similar in some cases. As observed in the first line of the Table 4.1 with Participants 1, 4 and 6, showing between 15 and 16 events, although the results are not consistent across all simultaneous annotations, with some cases having a very different number of events being detected (e.g. in the third line of Table 4.1 with Participants 1, 2 and 6). Therefore, it is possible to conclude that the number of EDR events does not present a reliable correlation between the EDA data responses in simultaneous annotations

In conclusion, the EDA signals in simultaneous annotations display a tendency to increase in amplitude over the period of the annotations, while for the remaining participants, which did not annotated in the same period, the same was not observed, displaying a decreasing trend in the EDA signal. Furthermore, the signals during simultaneous annotations display few similarities, as observed by the PCC obtained across the participant who annotated in the same segment. Regarding the individuals' self-reports, the obtained values do not follow the expected by a review performed by an expert annotator, thus revealing a lack of comprehension of the annotations scales by the participants, a lack of engagement towards the content and/or the annotation task, leading them to perform the annotations carelessly, with minimal attention. The expert annotator consisted in a person with a background in emotional analysis and a vast knowledge of the Valence and Arousal scales. These difficulties in assessment methods have already been described in [45], since these methods can be strongly biased by the level of attention of the participant, by subjective factors in the perception of the scenes or their annotation. Furthermore, they can also be considered to be intrusive, thus leading the participants to be reticent in performing a genuine self-assessment. A possibility to overcome the annotation tools limitations is to perform an emotional analysis solely based on the acquired physiological signals, in this case the EDA, and the movie content.

Collective Intelligence Analysis

Based on the methodology developed by Fleureau et al. [45] it was possible to determine the MAP of the audience for the movie at hand. Indeed, high correspondence is observed between the identified scenes by the annotator and the peaks of the MAP. Hence, verifying the correlation between higher arousal states given by the EDA data and stronger emotional scenes of the movie, i.e. intense scenes with strong emotions elicit high arousal emotional states in the audience.

Regarding the application of different clustering methods with features extracted from the group EDA signal, Table 4.2 presents a characterization of the resulting clusters. Since these results were obtained by extracting features from the group EDA signal, they represent the overall audience emotional response to the movie. Based on the analysis of the movie clips in each cluster, clusters 1 and 2 of the hierarchical average linkage, cluster 1 of the hierarchical single linkage and cluster 1 of the hierarchical complete linkage are all composed exclusively of intense movie clips that portray fight scenes, comical clips and emotional parts of the film, so these clusters successfully group the parts of the movie where the audience had a more intense emotional reaction (with higher arousal). It was observed that these clusters have the lowest duration (Length in Table 4.2) and contain the most relevant scenes, very similar to each other in the elicitation emotional content. For the remaining clusters (with greater length), although they may also contain scenes that can be labelled as emotional, they mainly contain very long-lasting clips with ”dead zones”, i.e. filling parts of the movie where the history is developing without eliciting any relevant emotion.

A further approach to verify the clusters containing ”dead zones” (besides the Length) is to identify the cluster with the initial instant of the movie, i.e. the first few seconds of the movie that displays the production companies. These instants are considered ”dead zones” and should be included in the clusters with low emotional-intensity response. Regarding the clusters obtained with the hierarchical ward linkage method, there was no clear distinction in the clips contained in each cluster, i.e. all clusters seemed to contain both intense emotional clips as well as ”dead zones” (even though cluster 1 appeared to contain less ”dead zones” and more emotional clips). Concerning the clusters obtained with the K-means algorithm, it is clearly seen that this method is the one that produced the greater amount of clusters, although it is also the most difficult method to evaluate. Despite the fact that the length of each cluster is smaller, and as said before for the hierarchical clustering method, smaller clusters meant that each cluster contained only relevant clips of the movie, in this case, some of the smaller clusters contain both relevant scenes as well as ”dead zones”. Even though two clusters seemed to stand out, clusters 3 and 5, they contained more intense emotional scenes than ”dead zones”, which suggests that the audience had a similar high arousal emotional response in these clusters.

In Figure 4.10 it is possible to observe the categorization achieved with 2 distinct hierarchical clustering algorithms, with the average linkage (Figure 4.10a) and with the ward linkage (Figure 4.10b). Based

on the data clustering with the ward linkage, it is possible to see, as expected, that each cluster is almost completely separated from the reaming ones; this happens since this algorithm aims to cluster data in order to minimal the variance within each group. On the other hand, the average linkage although it may suggest that the clusters are not correctly separated, especially clusters 0 and 2, this method determines the distance between an observation and a cluster based on the average distance between that observation and each element of the cluster, which can result in the clusters not appearing to be precisely separated in a reduced dimension case, such as this one.

Lastly, a comparison of the results achieved with the clustering methods with the MAP (considered as a validated method for emotional response profiling) is made in Figure 4.11. Since the clusters and the MAP have the same time resolution, the mean and STD of MAP was determined for each cluster. So, if the clusters containing exclusively emotional movie clips are located in the peaks of the MAP, this means that these groups contain the part of the movie where the audience had a stronger emotional reaction. This would validate the correlation between intense scenes and strong emotional reactions, and the successful clustering of the more intense emotional reactions of the audience. In Table 4.2 it is possible to see that the clusters obtained with the hierarchical clustering with ward linkage have the lowest mean MAP. Furthermore, in Figure 4.11 it is possible to see that all clusters of this method are spread out through the whole duration of the film, not being located exclusively near the MAP peaks, as such these clusters did not group the parts of the movie where the audience had a stronger reaction. On the other hand, the smaller clusters obtained with the hierarchical clustering with average linkage have higher mean MAP, meaning that the clusters obtained through this method correspond to the areas where the audience had a more intense reaction, as expected from the analysis of the clips contained in these clusters.

Moreover, in Figure 4.11 it is possible to see that these clusters are almost all located near the MAP peaks, thus confirming that this method achieved good results by grouping the parts of the movie where the audience had a more intense emotional reaction. Comparing the results of the hierarchical clustering with average, single and complete linkage, it is possible to see that the smaller clusters of these methods all achieved relatively good mean MAP, when compared to the average MAP throughout the whole duration of the movie. However, the average linkage method provides one additional cluster, being the one that obtains the highest amount of relevant clips; besides this, each clip in clusters 1 and 2 is located in the areas with higher audience arousal. Furthermore, clips in cluster 1 have the highest mean MAP of all clusters, thus tending to be located in the areas with higher audience arousal than the ones in cluster 2, suggesting that this method also provides a differentiation within the emotional parts of the films. With respect to the k-means method, the clips in each cluster seem to be randomly spread throughout the movie, although the two clusters that stood out tend to have more clips in the high arousal areas.

In conclusion, it is possible to determine the areas of the movie where the audience experienced an emotional reaction with higher intensity. When comparing this clustering methodology with the literature MAP, the best performing methodology was hierarchical clustering with average linkage, since it provides a higher number of clusters with more areas in which the audience had a more intense emotional reaction and it also differentiates the areas in which the audience reacted based on the intensity of such emotional reaction, i.e. it ranks the already stronger emotional reactions based on their intensity level into different clusters. Nevertheless, these results only provide insight to when and how much an audience reacts; they are mainly related to the Arousal dimension of emotion since the only physiological signal acquired was the EDA (see Section 2.4). To have an insight into the Valence level of the audience (how the audience reacted), other physiological signals related to the Valence, such as the PPG, should be analysed.

5

Real Time Collective Emotional Annotation tool

Contents

This chapter presents the development of the real-time collective emotional self-assessment tool. Section 5.1 presents the motive which led to the development of this tool and Section 5.2 comprises an overview of the state of the art in the scope of emotional self-assessment tools. Section 5.3 describes the approach followed during the construction of this tool, along with the steps taken during its testing. Lastly, Section 5.4 presents the results achieved in the evaluation of this tool and Section 5.5 discusses such results evaluating the usability of the tool developed.

5.1 Motivation

Emotion recognition has seen increasing popularity within the field of affective computing [11]. Most research in this area have a common denominator, which is the need for ground-truth data [7]; this is a core component in the development of emotion recognition algorithms, in particular, to train classifiers and validate their performance. This data is usually acquired with self-assessment methods, such as questionnaires filled by participants to report their own emotional state [24, 61]. The trend in the area is to explore data acquired in-the-wild, i.e., in environments where subjects observe previously unseen content [28, 46, 61], where standard annotation methods are difficult to apply.

Some assessment methods used are based on post-experiment analysis. So, there is a great latency period between experiencing an emotion and its annotation/description, since participants only describe the emotions experienced after the elicitation content is over, and provide only a global overview of the entire video clip. As such, there is an uncertainty as to which exact moment triggered their emotional response. This problem is exacerbated when using long duration elicitation contents, such as movies or documentaries [27, 52].

Furthermore, the current methods to annotate long duration content show some limitations, such as being built upon continuous emotion models, with predefined ranges on the emotional dimensions. When annotating data in-the-wild and in real-time, while observing previously unseen content, this can be a limitation since it is impossible to go back and correct given input levels misinterpreting the levels range, should the subject experience higher or lower extreme stimuli. Thus, if the participant were to annotate the maximum allowed value in one of the emotional dimensions, it would be impossible to correctly annotate a future emotion with higher intensity on that dimension. For example, using the Arousal-Valence model with values between 1 and 5 for each dimension, the participant annotates an emotion which he rates with a value of 5 in Arousal, although a further scene from the movie elicits a more intense emotion; in this case, it would impossible to correctly annotate the second emotion experienced. Another limitation of these methods is that most of them are desktop-based, and can only be used for emotional elicitation using video content, thus not being very versatile. Furthermore, performing an emotional annotation using a PC in a group environment would require users to have

their own PC or the annotation to be performed in turn which would take a significant amount of time. However, most people nowadays carry a smartphone with them, so using this device which users are very familiar with would be faster and easier [76].

With this in consideration, this work seeks to explore the use of smartphones as a self-assessment annotation tool. It builds upon the Valence-Arousal model, to enable subjects to report their rating with any sort of elicitation material, in specific while watching long duration audiovisual content in-the-wild (e.g. a movie or a documentary), providing minimal distraction from the content being watched.

5.2 State-of-the-Art

Although, in recent years some studies have been reported the use of emotional assessment in real time, these are still in smaller number than those not applying this method. Nevertheless, Cowie et. al [77] and Girard et. al [78] proposed a tool that allows participants to watch a video elicitation in half a screen and in the other half annotate one dimension of the emotion (Valence) experienced, using a slider scale. The work of Sharma et. al [79], Girard et. al [80], Yannakakis et. al [81] and Cowie et. al [33] improved on the previous developments, by creating an annotation tool with two dimensions, Valence and Arousal, where the annotation is performed with a joystick or mouse, while watching the elicitation video on the rest of the screen.

As the number of smartphone users increases, and with the pervasiveness of these devices, there is a growing interest in using smartphones for annotation. Thus, Zhang et. al [76], built upon the work of Sharma et. al, to create an annotation tool using a smartphone with a joystick to annotate emotional experience in a Valence-Arousal space, which is overlapped in the bottom right corner with the video elicitation shown in the rest of the screen. Another work using smartphones for emotional annotation was developed by Muaremi et. al [82], although in this case a discrete emotional model with the PANAS questionnaire was used, with the aim of discretely acquiring data throughout the day.

The works of Lopes et. al [34] and Melhart et. al [35], proposed two annotation tools similar to the ones developed by Cowie et. al [77] and Girard et. al [78], using a one dimensional continuous emotional model. Although, in these works, the Arousal state is measured using a scale without upper or lower limits. The lack of bounds enabled users to annotate more freely, without having to anticipate future experiences, thus being more intuitive to use. Furthermore, this data could be normalized post-hoc, restraining it between certain bounds and mapping it into Valence-Arousal scales which can be compared across studies, without affecting the annotator's experience. This also led to a increase in inter-rater agreement.

We extend the work found in the state-of-the-art by: 1) Proposing an annotation tool relying solely on the user's smartphone; 2) Testing two alternative designs for annotation; 3) Using a continuous emotion

model without upper or lower limits (unbounded); and 4) Allowing the annotation in different real-world environments, accommodating elicitation methods that may not necessarily use the smartphone for presentation (e.g. movie theaters, recitals, music concerts, live sports games, and related scenarios).

5.3 Proposed Methodology

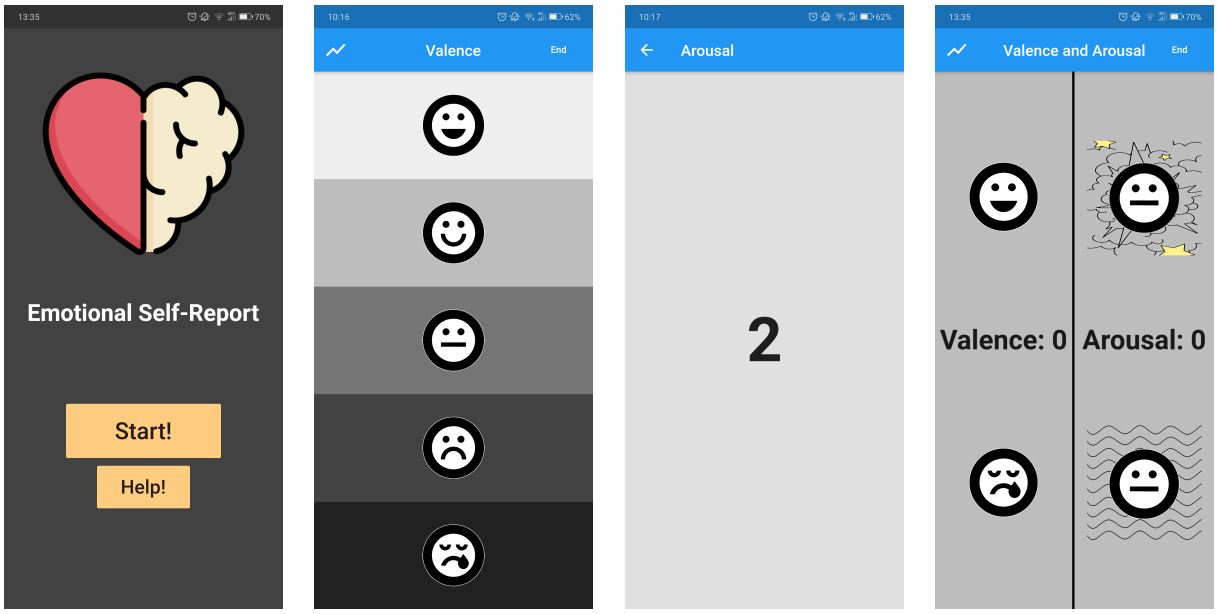
This section is divided into four sectors. The first sector describes the overall characteristics of the annotation tool developed, e.g. the two different version created, the emotion model used, etc. The second sector describes the annotation phases in each version of the application. The third sector presents the final pages of the application also describing how to store the acquired data. Finally, the last sector describes how both versions of the annotation tool were tested in terms of usability and mental workload.

Application Design

The developed application was designed to acquire the Valence and Arousal dimensions [15], which are annotated by the user, and the time stamp in absolute values, which is automatically registered when an annotation is given. By having the time stamp in absolute values it is possible to determine the instant of the elicitation content in which each state was reported by the subject, enabling this system to be used with a variety of elicitation materials and to always be possible to synchronize annotations and elicitations. To determine the best design, two different interfaces were developed using distinct approaches, namely a Two-step Sequential Annotation (TSSA) version and a One-step Matrix Annotation (OSMA). This allowed us to determine, based on an end-user perspective, which annotation method provides the least amount of distraction while efficiently collecting the data.

The first page of both versions of the app is the home page (Figure 5.1a). This page has two buttons, the ”Start!” and the ”Help!” button. The start button initiates the acquisition by acquiring a neutral Valence and Arousal pair together with the initial time instant, besides, it also directs the user to the annotation page. The Start button should be pressed simultaneously with the start of the elicitation content, so that the time instant saved corresponds to the start of the elicitation, although if the elicitation start instant is also recorded in absolute terms it is still possible to link both results.

The ”Help!” button, as the name suggests, leads to a help page that contains a detailed explanation on how to annotate one’s emotional state, together with the meaning of the terms Valence and Arousal, along with a usage guide. This allows the subject to learn how to use the app and to become familiarized with the app before using it. Since each version of the app has a different annotation method, each version has a different help page, specific to that version.



(a) Home Page. **(b)** Valence annotation TSSA. **(c)** Arousal annotation TSSA. **(d)** Annotation phase OSMA.

Figure 5.1: Main screens of both versions of the applications.

The applications developed were inspired by the SAM questionnaire [3]; as such, in both cases, pictograms were used in the rating buttons to represent the extreme emotional states of the Valence-Arousal axis. This approach was preferred over assigning words (e.g.) since cross-cultural issues can have an influence on the interpretation of the terms, while symbols such as emoticons enable the participant to relate to the emotions expressed by such images.

Annotation Phase

This sector is subdivided into two different parts, one for each version of the application developed. The first part describes the annotation phase characteristics in the TSSA version and the second part describes the annotation phase characteristics in the OSMA version.

Two-step Sequential Annotation

This version has two different pages, one for the Valence and one for the Arousal. After the start button is pressed, the user is directed to the Valence page (Figure 5.1b). This page has 5 different buttons, each one with a different pictogram from happy (top button) to unhappy (bottom button), and different color tones representing a 5-point scale for the Valence values. After selecting a valence value by pressing one of the buttons, the application proceeds to the Arousal page (Figure 5.1c). Here the user annotates his/her Arousal level by touching the screen; the higher the number of touches the higher the reported arousal. After one of the Valence buttons is pressed, there is no way of confirming the input value; on

the other hand, the arousal input is always displayed on the screen during the annotation of this value.

Since this app is aimed for emotional annotations in the wild with uncalibrated and unseen content, the participant has no way of knowing the intensity of future emotions. Furthermore, based on the work developed by Lopes et. al [34] and Melhart et. al [35], and considering their promising results using unbounded annotation of Arousal, the Arousal annotation in the TSSA version has no upper limit, enabling the user to adjust higher and lower intensity states taking into consideration their previous state level.

After the annotation of the Arousal value, the Valence and Arousal pair is registered along with the annotation input time instant. Five seconds after, the application automatically redirects the user to the valence page and waits for the next annotation. Hence, the annotation phase works in a cycle, where each annotation starts from a neutral state. There is no limit for the number of annotation neither for the time interval between annotations. This cycle is interrupted when the participant presses the ”End” button in the top right corner of the valence page; this ends the annotation phase and directs the user to the final page of the application. Ideally, this button is only pressed when the elicitation content being watched ends.

One-Step Matrix Annotation

In the OSMA version, the annotations are registered using only one page for both the Arousal and Valence values. Figure 5.1d shows the annotation page, where it is possible to see a layout with two columns, the one on the left for the Valence and the one on the right for the Arousal. On both columns, the user can annotate the level of the emotion experienced by pressing the upper or lower buttons identified with the pictograms. This would increase or decrease, respectively, the reported level that is displayed in the centre of the column.

The OSMA annotation method builds upon the work of Lopes et. al [34] and Melhart et. al [35], by using an unbounded annotation for both Valence and Arousal dimensions. So, since both annotations have no upper or lower limit the user can always introduce a more or less intense and more positive or more negative emotion than the ones previously experienced.

Throughout the visualization of a movie, subjects usually do not jump from a very positive emotional state to a very negative emotional state without passing through some neutral states in between. Therefore, the emotional annotation in this version is continuous, meaning that after an emotion is registered the app saves this emotional state as the baseline for the next emotion to be annotated. In order words, the user only has to adjust the previous Valence and Arousal state by increasing or decreasing each level by the desired amount, and this would result in the annotation of a new emotional state.

Using this method, participants only have to slightly adjust their emotional state instead of having to record a completely new state every time, enabling a more continuous annotation. This also helps

reducing the number of annotations required, since the user only has to annotate when a change in his emotional state is experienced. In this case, the Valence and Arousal pairs, along with the time stamp, are saved automatically after each annotation. The annotation phase is finished once the user presses the ”End” button in the top right corner of the page. Once again, this button should only be pressed when the elicitation content is over.

Post Study Questionnaire

The final page, which can be seen in Figure 5.2, has a small questionnaire with three questions where the user is asked to evaluate the content regarding its engagement towards it, familiarity and liking on a 5 point Likert scale. In this page, it is also possible to fill an optional field with the participant's number (useful to distinguish between each participant's annotation without needing any personal information, ensuring the privacy of everyone). This field was set to be optional since in certain acquisitions the users may perform the tasks individually, thus not requiring a participant number.

Lastly, to stored the data the user has to press the Send button at the bottom of the page. After this button is pressed, the participant can choose a variety of different options to save or send the data; these options can be seen in Figure 5.3. The preferable method is to send the data via email, in which case an email address is already predefined as the destination, although it is possible to change it by introducing an email address in the ”Email:” field and pressing the button ”Change email”. The data is exported via a single Comma-Separated Values file, which is a very versatile format that can be analysed using a variety of methods and environments.

Experimental Methods

To evaluate the performance of the two developed applications, an experimental evaluation was performed. Both versions of the app were made available in the *Google Play Store*[1,2]. Subjects were asked to download the app to their phones and watch a small audiovisual content with a duration of 5 minutes while using the app. After the download has been done, the participant read the help page of the app, to understand how to use it and to become familiarized with the application. If there were any questions from the subject regarding the use of the application or the procedure to be followed, these were clarified to ensure that they fully understood how to use the app and the experiment protocol.

Afterwards, the content display was initialized and the participant pressed the Start button to initiate the annotation phase. Throughout the visualization of the content, participants were asked to annotate their emotional state. When the video ended, subjects pressed the End button to terminate the annotation phase and fill the final page questionnaire (Figure 5.2). In the end, the annotations were sent via

[1] https://play.google.com/store/apps/details?id=com.emoteu.app
[2] https://play.google.com/store/apps/details?id=com.emoteu2.app

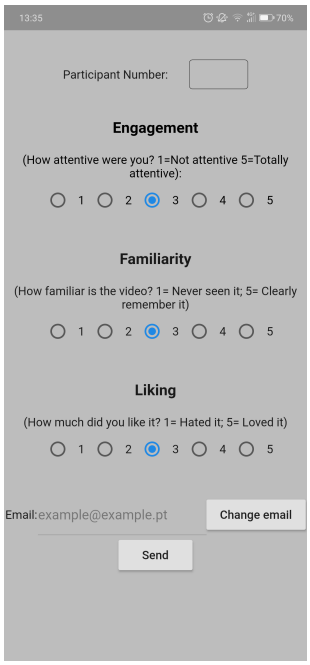


Figure 5.2: Engagement, familiarity and liking page.

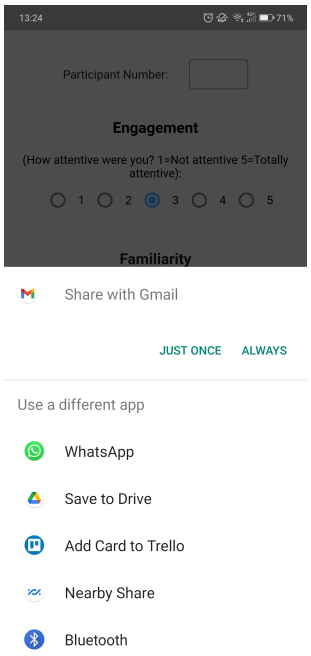


Figure 5.3: Annotation sharing options.

email, and participants were asked to fill an online questionnaire regarding their experience when using the application.

This questionnaire consisted of two groups of questions. The first one corresponded to the SUS questionnaire [83,84], which contains 10 questions; these questions are half worded positively, half negatively, and answers were given in a 5-point Likert scale with anchors for ”Strongly agree” and ”Strongly disagree” [83]. This method has been widely used and shown not to be biased against gender or certain types of user interfaces [84]. Additionally, two optional open questions were also added aimed at obtaining the users' opinion regarding the best and worst features of the app, along with improvement suggestions.

The first question of the SUS questionnaire, ”*I think that I would like to use this system frequently*”, was changed since this system is not aimed at being used on a regular basis. So, in order to better represent the software being assessed, this question was replaced with the question: ”*I would repeat the experience*”. This way it would not break the standard positive/negative statement balance of the SUS.

Lewis et al. [8] proposed a grading scale (represented in Table 5.1) constructed based on a large data sets with thousands of individual SUS questionnaires and hundreds of studies. This represents a good general guide for the interpretation of the SUS results, by grading them on a scale from A to F, based on the percentage of systems that obtained each score. In other words, out of all the systems contained in the data set evaluated by Lewis et al. [8], only the top 4% best systems achieved a SUS

Table 5.1: SUS grading scale and percentiles [8].

SUS Score Range	84.1 - 100	80.8 - 84.0	78.9 - 80.7	77.2 - 78.8	74.1 - 77.1	72.6 - 74.0	71.1 - 72.5	65.0 - 71.0	62.7 - 64.9	51.7 - 62.6	0.0 - 51.6
Grade	A+	A	A-	B+	B	B-	C+	C	C-	D	F
Percentile Range	96-100	90-95	85-89	80-84	70-79	65-69	60-64	41-59	35-40	15-34	0-14

Table 5.2: Deciles and Quartiles of the global NASA-TLX analysis [9].

Percentile	Min	10%	20%	25%	30%	40%	50%	60%	70%	75%	80%	90%	Max
Score	6.21	26.08	33.00	36.77	39.45	45.00	49.93	53.97	58.00	60.00	62.00	68.00	88.50

score between 84.1 and 100, being rated with an A+ grade.

The first step to obtain the SUS results is to convert the raw scores for each question to an adjusted score raging from 0 (poorest rating) to 4 (best rating). Given that the questions are half worded positively and half negatively, the scoring consists of subtracting 1 from the raw positively worded scores and subtracting the raw score of the negatively worded questions from 5. Afterwards, the sum of the adjusted scores is computed and then multiplied by 2.5 to obtain the standard SUS score, ranging form 0 to 100 [83, 84]. This calculation is summarized in Equation (5.1) (the $SUSn$ characters represent the raw score given in the n question of the SUS questionnaire):

$$SUS = 2.5(20 + \sum_{n=1 \mid odd\ n}^{9} (SUSn) - \sum_{n=2 \mid even\ n}^{10} (SUSn)) \tag{5.1}$$

A second questionnaire, the NASA-RTLX, was given to the volunteers to measure the participants' subjective mental workload. The mental workload is obtained based on six subscales: mental demand; physical demand; temporal demand; frustration; effort; and performance [85]. Each subscale is rated in a 7-point Likert scale with anchors for "Very Low" and "Very High". The results from each sub-scale are then converted to a relative scale from 0 to 100, and the average result per participant is determined. The NASA-RTLX is a simplification of the NASA Task Load Index since this questionnaire is composed of 2 separate parts and the NASA-RTLX only has one of these parts. In the second component of the NASA Task Load Index, participants choose the most significant sub-scale in each of the 15 combinatorial pairings of sub-scales [85]. However, there is a high correlation between the results from both questionnaires [9], so the NASA-RTLX was determined to be the best questionnaire to assess the participants' subjective mental workload due to its effectiveness and simplicity.

The mental workload of the system is obtained by performing the average of all the participants' scores [9]. In the work of Grier et al. [9], a rating scale was proposed which makes it possible to establish in which percentile of the NASA-TLX score the system developed is rated, concerning other systems evaluated with this method. This scale is represented in Table 5.2, where it is possible to see that if a given system has a mental workload score of 20, this system would be in the top 10% best systems evaluated in the work of Grier et al. [9].

The SUS evaluates the usability of the system, thus, obtaining a high SUS score means that the tool being evaluated has high usability and is easy to manipulate. The NASA-RTLX measures the mental workload during the use of the system and, consequently, a low mental workload score implies that the tool being evaluated does not require an elevated level of concentration to be used correctly. Given the analysis being performed in this work, the ideal results would be to achieve a high SUS score and a low NASA-RTLX score. Since this work aims to evaluate the use of the application itself, the Valence and Arousal ratings will not be analysed.

5.4 Results

A total of 16 participants, 9 female and 7 males, aged between 21 and 34 years old (μ=24.38, σ=3.35) participated in this experiment. This corresponded to 26 answers to the online questionnaire, 10 for the TSSA and 16 for the OSMA. The SUS scores obtained were 82.75 $\pm$ 5.29 for the TSSA and 82.66 $\pm$ 8.25 for the OSMA. On the other hand, the NASA-RTLX scores obtained were 44.52 $\pm$ 7.26 for the TSSA and 37.50 $\pm$ 6.77 for the OSMA. The results for each question in each scoring system can be seen in Table A.1 and Table A.2, respectively.

Regarding the answers to the open questions present in the questionnaire, these consisted mainly in small improvement suggestions, such as adding a small video tutorial in the help page or changing the application design to be more engaging for the user. On the other hand, some answers described the best qualities of the application for the user, these included comments regarding how simple and accessible both versions were. It should be noted that in participants who tested both versions of the app, the functionality of the OSMA version to save the current emotional state as the baseline for the next emotion to be annotated, was pointed out to be a highly appreciated feature, enabling a quicker and user friendly annotation with lower mental workload.

5.5 Discussion

Based on the the SUS results and on Table 5.1, both versions obtained an A grade, meaning that both applications are on the top 10% best systems according to the SUS scoring method [8]. On the other hand, based on the NASA-RTLX scores and on Table 5.2, the TSSA version is on the top 40% best systems evaluated through this method, and the OSMA is on the top 30% best systems. The SUS revealed that both version of the app have high levels of usability, thus being very easy to used. In terms of the NASA-RTLX, these revealed that both versions introduce some level of mental load, although it is low.

The obtained outcomes are very promising, showing that the developed applications are able to

report one's emotional experience without much distraction from the content being watched. The SUS scores from both versions had a minimal difference of 0.09 between them. However, the NASA-RTLX were quite different, with the OSMA obtaining a considerably lower score compared to the TSSA. So, both versions have a similar usabilities, although the OSMA has lower mental workload leading to a more intuitive use of this version, and constituting a lesser source of distraction from the content being observed. As such, the OSMA version was considered to be the preferable approach for real-time emotion annotation, better achieving the desired goals.

To conclude, the experimental results are in agreement with the literature, with rank unbounded annotation being the most promising approach for emotion annotation with higher reliability [34, 35]. Regarding the objectives set in the beginning of this work, these were considerate to be achieved with the development of the two smartphone applications for emotional annotation. The developed tools revealed a low mental workload and high usability, leading to the conclusion that they can be used in real world scenario, providing a reliable emotional annotation with minimal distraction. Lastly, since these tools were developed for smartphones, which have a high mobility and adaptability, it is possible to use them with a variety of elicitation materials.

6

Moving Forward

Contents

6.1 Motivation

The EDA is strongly correlated with the Arousal dimension of emotions, although it provides a limited amount of information regarding the Valence dimension. Thus, by only acquiring this signal an important aspect of emotion analysis is being disregarded. By combining the information obtained through the EDA with another physiological signal, correlated with the Valence dimension, it would fill in the blanks left by the EDA and enable a more comprehensive analysis towards emotion recognition. This way each signal would be strongly related to a different emotion dimension. The second signal to be acquired would need to fill some requirements, being the most relevant one that it needs to provide information regarding the Valence dimension of emotions. Another desired characteristic of this signal is that it should be non intrusive and easy to collect, preferably it should be possible to collect in the hand/wrist area as well. Based on these requirements, the proposed physiological signal to be used is the PPG [1].

To develop a biomedical device aimed at collecting any sort of physiological data, or to process physiological signals, an important factor is the minimum SF of the device. This characteristic has a great influence on the quality of the acquired data. Using a sampling rate below the optimal may result in aliasing, in which there is a great loss of data making it impossible to correctly reconstruct the signal at hand. On the other hand, using a sampling rate too high increases the computation load and storage necessary to process and save the data, increasing also the power consumption of the device, creating the need for a larger battery, hence leading to a larger device. Theoretically, the minimum SF can be determined by the Nyquist-Shannon theorem: according to this theorem the minimum SF should have a value be at least two times larger than the highest frequency of the signal ($w_s > 2w_m$, where w_s is the sampling frequency and the w_m is the maximum frequency of the signal) [86]. Although, the acquisition system characteristics are not always known or precisely specified, difficulting the processing tasks.

There are other factor which influence the decision to select a minimum SF. In the case of emotion analysis, the onset and peak detection on EDA signals may also influence this decision. Furthermore, the interpolation techniques used to reconstruct the signal into a higher SF also affect the minimum SF since some techniques allow the usage of lower SF. The type of sensors used and the characteristics of the filter should also be taken in consideration. Lastly, the experiment settings are also a crucial factor: individual versus collective acquisitions have very distinct requirements. In group acquisitions there are several devices in the same network interchanging data, thus considerably increasing the probability of data-loss due to collisions, and leading to the need of using a lower SF to reduce the data loss. In individual data acquisition, this is not such a considerable issue, thus allowing the use of higher SF compared to group settings.

As such, in order to acquire 2 physiological signals simultaneously in a group environment, a new device needed to be used, since the FMCI only allows the acquisition of the EDA signal alone, with

a sampling frequency of 1 Hz. The chosen device was the BiTalino R-IoT[1], which is able to acquired 2 physiological signals from several subjects simultaneously. The present work focuses on the benchmarking the BITalino R-IoT device against a reference device, to guarantee that it performs correctly and acquires high-quality data, without any loss of information.

6.2 State-of-the-Art

In the field of emotion recognition, several studies have been performed using only the EDA signal as a way of analysing the individual affective state. This is the case of the work developed by Wang et al. [6], in which the EDA signal is used to assess the volunteer response to an audio track of a commercial, and later to help in the design of new advertisements. However, the use of this signal alone has been seen as a limitation, since it mostly provides information regarding the Arousal dimension of emotions. Liapis et al. [26], focused on stress recognition in human-computer interactions, such as slow network speed, and although the findings of this study seemed promising, the usage of additional physiological signals was denoted as a requirement for improvement in future works. The work of Fleureau et al. [45] and Li et al. [42] both focused on measuring an audience reaction to an audio-visual content based on the EDA signal. Although the continuous model of emotion was utilized in both scenarios, only the Arousal dimension was evaluated since it is the one related to the EDA; this was seen as a limitation by the authors of both articles and could be solved with the collection of additional physiological signals.

Benchmarking is the process of determining the highest standards of excellence for a product and if necessary make the improvements needed to reach such standards. In other words, benchmarking raises the standard of products and identifies those who cannot keep up [87]. The comparison results between a new device and an established reference can be evaluated in terms of accuracy, reliability and feasibility. A new device can be labelled as sufficiently accurate if its measurements are shown to be comparable to those obtained by the reference in the same conditions. However, this evaluation is relative to the type of data and the goal of its use, for example, to develop a continuous monitoring system that measures the HR of a patient, collecting the PPG signal is sufficiently accurate in relation to the ECG signal. The reliability and feasibility evaluate the system in terms of reproducibility and applicability to a given specific settings, respectively [88].

Kleckner et al. [88] proposed six steps for benchmarking mobile devices in psychophysiological and physical activity research. These steps are: Step 1 - Identify signals of interest; Step 2 - Characterize intended use cases; Step 3 - Identify study-specific pragmatic needs; Step 4 - Select devices for evaluation; Step 5 - Establish assessment procedures; Step 6 - Perform qualitative and quantitative analyses. These steps can help guide the benchmarking process.

[1] https://bitalino.com/storage/uploads/media/manual-riot-v12.pdf

In the work developed by Batista et al. [89] the low-cost and easy-to-use toolkit BITalino (r)evolution was benchmarked against the established reference BioPac MP35 Student Lab Pro (BSL), using four signals: ECG, electromyography, EDA and electroencephalography. This work revealed that the signals acquired by the BITalino (r)evolution were similar to those acquired with the well-established devices which further confirmed the results found in [90].

The problem of finding the minimum SF has been tackled several times throughout the years, leading to several papers being published regarding this topic. Due to its clinical relevance, the majority of studies focus on ECG signals. In [91] the authors recommend a range between 250 and 500 Hz for ECG, with a slightly lower frequency of 100 Hz also being possible to use, although, only if coupled with quadratic interpolation to refine the R-peak. Ellis et al. [92] studies the effect of lower SF on an ECG signal originally recorded at 1000 Hz. In this work, the impact of a lower SF is evaluated across 24 widely used time- and frequency-domain measures of HRV on healthy subjects. Ziemssen et al. [93] analyses the impact of the sampling frequency in ECG signals originally recorded with 500 Hz SF and downsampled to 200 and 100 Hz. The 100 Hz was shown to be sufficient in healthy individuals, and to have minor influence in pathological patients. The overall recommend SF for ECG was 100 Hz with interpolation.

A relevant work is the one developed by Béres et al. [94], although this study was developed using PPG and not EDA, the methodology followed was very comprehensive. Data were acquired from healthy individuals at 1kHz and down-sampled by a factor of 2, 5, 10, 20, 50, 100, 200, 500. The decimated data were then interpolated back to 1kHz using cubic and quadratic splines. The results revealed that a SF of at least 50 Hz without interpolation and, 10 Hz and 20 Hz with interpolation is required to achieve an acceptable standard deviation and root mean square of successive RR-differences.

Regarding the EDA, the literature is very reduced. Boucsein [4] reports sampling rates between 10 and 40 Hz in his work. Moreover, the author recommends the use of 20 Hz SF, which can be decreased down to 1 Hz for EDL data. However, he notices that in cases where EDA decomposition may be necessary, the SF must be increased from 1Hz to values between 4 to 8 Hz.

6.3 Proposed Methodology

Benchmarking of the BITalino R-IoT

In previous researches the FMCI device has been benchmarked with the BITalino (r)evolution as a reference [43]. As such, the current work of benchmarking of the BITalino R-IoT, also uses the BITalino (r)evolution as a reference.

The BITalino (r)evolution has been tested several times in the past, revealing to be a reliable physiological signal acquisition tool, obtaining high quality data [89, 90]. Furthermore, this device is able to

acquire data from 6 different channels at the same time. However, with BITalino (r)evolution it is only possible to simultaneously acquired data from 3 devices due to its communication via Bluetooth and the high data throughput[2,3]. Since the BITalino (r)evolution has a limit of 3 devices acquiring simultaneously, it can not be used as an acquisition tool in group settings. Nonetheless, due to its scientifically validated multi-sensor acquisition capabilities, it can be used as a reference in this work.

The device being evaluated in the current work, the BITalino R-IoT, communicates via Wi-Fi which enables it to collect information from several devices simultaneously, eliminating the limitation set with the Bluetooth communication in the BITalino (r)evolution. Furthermore, the BITalino R-IoT acquires data with 200 Hz SF, which as been shown to be enough for most physiological signals, namely the ones herein foreseen (EDA and PPG) [94]. This device is composed of a rechargeable 3.7 V battery, it contains integrated accelerometer, gyroscope, magnetometer and Euler angles calculation with 3 degrees of freedom along with a temperature sensor [95][4]. Moreover, it is also possible to add two additional sensors with 12-bit resolution, used to collect EDA and PPG data in this case.

For the current work data was acquired from volunteers, older than 18 years old, without any known pathology. Participants were asked not to be under the effect of alcohol or medication before and during the experiment, and in case they show any limitation either physical or psychological, necessary for the realization of the experiment they would not be allowed to participate.

The data acquisition was carried out in an individual setting, using both devices simultaneously, with the electrodes placed on the subjects non-dominant hand according to Figure 6.1. Each acquisition was composed of 3 sequential tasks, each one with the duration of 20 minutes, comprising a total duration of 1 hour. The first task consisted in watching video chosen by the participant (usually correspond to an episode of a series). In the second task, the participant was playing some computer/mobile games. The last task consisted in following a meditation guide. The goal of these tasks was to elicit different states, i.e. the first task was to elicit a neutral state, followed by a stressful situation during the game, ending with a relaxation period, although for this part of the work the specificity of the events is not relevant.

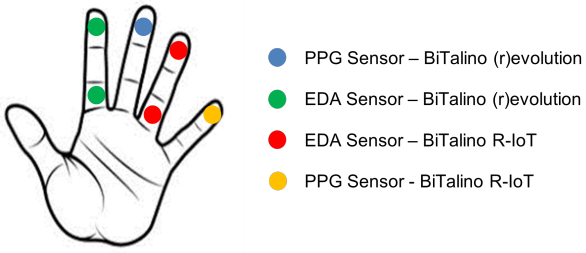


Figure 6.1: EDA and PPG sensor placement for benchmarking purposes.

To synchronise the data collected by the two devices, accelerometer data was acquired by the BITal-

[2]https://plux.info/software/43-opensignals-revolution-000000000.html
[3]https://www.bluetooth.com/specifications/bluetooth-core-specification/
[4]https://bitalino.com/storage/uploads/media/manual-riot-v12.pdf

ino (r)evolution using a sensor placed on the BITalino R-IoT shell. The synchronization was achieved by creating a prominent peak in the acceleration signals of both devices with a small stroke on the sensors. Given that the sensors remained still for the rest of the acquisition, the signals would be constant throughout the acquisition, except for this peak, thus marking the beginning of the acquisition in both devices. During the signal processing step, the acquired signals from each device were cropped on the acceleration peak location of their device.

The acquired signals from each device were stored in different HDF5[5] files in a hierarchical format. For each user a two signal dataset were created containing all the information acquired from this user (one for each device). Data processing was conducted on a Python 3 environment, with the support of BioSPPy (version 2) toolbox [19], a publicly available set of signal processing tools to analyse biosignals. Since the signals acquired with the BITalino R-IoT had a lower SF, these where interpolated to the same SF as the BITalino (r)evolution (1000Hz) using a cubic spline interpolation. Afterwards, the quality of the each signal was evaluated manually; this assessment was performed based on saturated signals, disconnections in the mid acquisition, and signals with a constant amplitude. The signals were smoothed using the 15$\times$SF point moving average following the approach described in [74].

Finally, the identification of the fiducial points was performed using the resources of the BioSPPy library, selecting amongst the several algorithms available based on approaches described in the literature. For PPG systolic peak detection, the algorithm proposed by Elgendi was used [96]. This algorithm relies on squaring, generating blocks of interest, and defining a threshold to detect systolic peaks. The parameters W1, W2 and Beta were set to be 0.18, 0.69 and 0.01, respectively. For the EDA, these points were detected using the same process described in Section 4.3.

The comparison between the signals acquired with both devices was achieved based on the detected peaks and onsets of the PPG and EDA signals, respectively. The first step in this comparison was to match the detected points for each pair of signals (PPG and EDA) acquired by the two devices, to avoid mismatching with false peaks from noise and motion artifacts.

To compare the data morphology, the PCC was calculated between the signals of the two devices. For the PPG these values were calculated for a time period which corresponded to 0.25 s before the detected peaks and 0.5 s after. For the EDA these values were calculated between the onset and end point of each event. Beyond these metrics, the number of detected peaks was also evaluated (for both EDA and PPG signals), along with an extraction of the HR from the PPG signals.

Influence of the Sampling Frequency in EDA signals

To perform the analysis of the minimum SF, the CoolWorking dataset provided by BrainAnswer [97] was used. The data set was acquired to study the impact in physiological signals and in the individuals'

[5] https://www.hdfgroup.org/solutions/hdf5/

Table 6.1: Overview of EDA sensors specifications.

	BITalino EDA
Range	0 – 25 µS (with VCC = 3.3V)
Bandwidth	0 – 2.8 Hz
Consumption	$\pm$ 0.1 mA
Type	Exosomatic

performance while performing psychometric tests during a 2-day sleepless event for students. For this study, 98 records from the dataset were used, in which the participants' ages ranged from 15 and 30 years old ($\mu = 20.4, \sigma = 3.3$), with 70.4% (69) identified as female and 29.6% (29) as male.

The data was acquired using a BITalino (r)evolution [98] with 10-bit resolution at 1KHz. The EDA was acquired with two Ag/AgCl electrodes: one was placed on the index finger and the other on the ring finger of the left hand [6], the specifications of the EDA sensor can be seen in Table 6.1 [7].

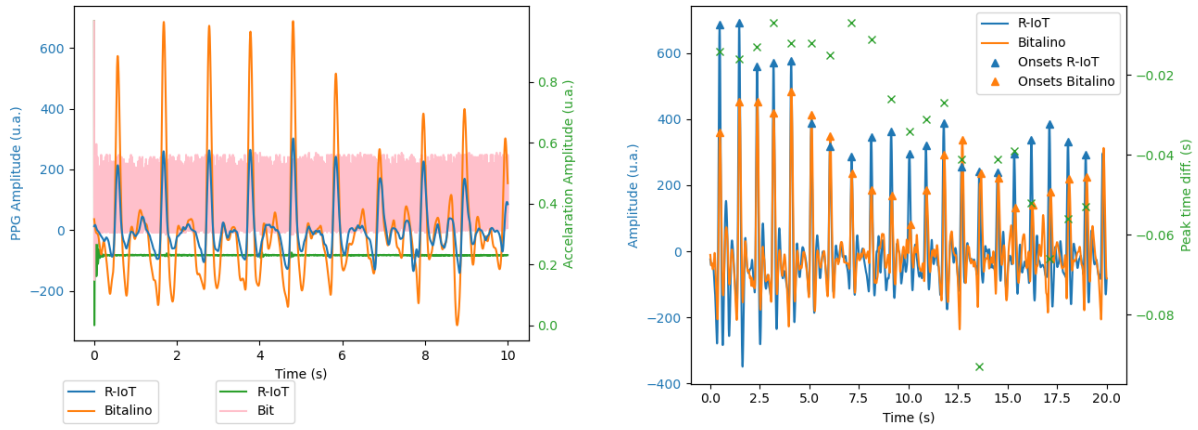
The selected files were initially analysed manually, which resulted in a removal of 8 files, leaving 90 to be further analysed. The selection criteria was based on morphology of the signals, i.e. saturated signals, disconnections mid acquisition, and signals with a constant amplitude were excluded.

To evaluate the impact of low SF on the quality of the signal, the data from each participant was downsampled to 500, 200, 100, 50, 20, 10 and 1 Hz. These signals were then interpolated back to 1000 Hz SF with 4 different interpolation methods to test their effect on onset time estimation. These methods were: no interpolation, linear spline, quadratic spline and cubic spline using the SciPy 1.7 Python library [99] implementation.

The onset detection was performed using the same methodology as the one described in Section 4.3. Afterwards, time and amplitude differences were then computed between the original 1 kHz signals and their downsampled versions.

To study how EDA signals are distorted with different sampling rates, the original 1kHz filtered signals and the interpolated downsampled signals were first segmented between each event onset and the end point. To perform a waveform distortion analysis, the PCC was used. This coefficient is computed between two signals, and it gives a normalized measure of linear correlation. By computing the PCC between a segment and its interpolated downsampled version, it is possible to verify the overall profile of correlation of the segments. The PCC was then calculated using the simultaneous event segments, these values were only calculated for the linear, quadratic and cubic spline, since the signals to be correlated are required to have the same number of points (in the case of the no interpolation method the signal has a fewer amount of points due to downsampling).

[6] https://plux.info/electrodes/60-non-gelled-reusable-agagcl-electrodes-870122114.html
[7] https://bitalino.com/storage/uploads/media/eda-sensor-datasheet-revb.pdf



(a) **(b)**

Figure 6.2: Representation of the synchronization process with the acceleration and PPG data (a) and determination of the PPG temporal difference between the two devices being evaluated (b).

6.4 Results

Benchmarking of the BITalino R-IoT

The current work is based on the data acquired from 3 volunteers, from whom 2 where male, the average age of the participants was 23 years old, with a STD of 1.4.

In Figure 6.2a it is possible to observe the two YY axis, the left one corresponds to the PPG data acquired from by the BITalino R-IoT and the BITalino (r)evolution, in blue and orange, respectively. The right YY axis corresponds acceleration data collected by the BITalino R-IoT and the BITalino (r)evolution, in green and pink, respectively. Figure 6.2b follows the same principles as the previous figure; it is also possible to observe two YY axis, with the left one corresponding to the PPG data acquired with both devices, using the same color code. The right YY axis represents the time differences in the detected peaks of each signal. In this figure it is also possible to see the detected peak presented with small triangles in the color of each device. These images only display the PPG signal since this is a higher frequency signal, having more peaks than the EDA and being more complex, so it is more accurate to evaluate the synchrony between the two devices with the PPG data.

In Table 6.2 it is possible to see a comparison between the PPG signals extracted with both devices in terms of the number of detected peaks, waveform similarities using PCC values, temporal difference in the detected peaks on each signal, and the HR extracted from each signal. In a similar manner, Table 6.3 displays a comparison between the EDA signals extracted with both devices in terms of the number of detected onsets, waveform similarities using PCC values, temporal differences in detected onsets on each signal and event duration.

Table 6.2: Comparison between the number of detected peaks, PCC, temporal difference and extracted HR for the PPG signal extracted with the BITalino (r)evolution and BITalino R-IoT.

Participant	Nb Values			Waveform Similarity			Time difference (s)				HR (bpm)	
	#N peaks R-IoT	#N peaks (r)evolution	% matched peaks	Mean PCC	STD PCC	Min PCC	Mean	STD	Max	Min	R-IoT	(r)evolution
0	3633	3747	89,78	0,91	0,15	0,02	-0,02	0,02	0,10	-0,10	77,8	77,7
1	3287	3228	96,59	0,96	0,08	-0,02	-0,01	0,02	0,10	-0,10	62,29	62,29
2	3961	3979	96,33	0,97	0,06	-0,47	-0,04	0,02	0,09	-0,10	68,65	68,66
Overall	10881	10954	94,17	0,94	0,11	-0,47	-0,03	0,03	0,10	-0,10		

Table 6.3: Comparison between number of detected onsets, PCC, temporal difference and event duration for the EDA signal extracted with the BITalino (r)evolution and BITalino R-IoT.

Participant	Nb Values			Waveform Similarity			Time difference (s)			
	#N onsets R-IoT	#N onsets (r)evolution	% matched onsets	Mean PCC	STD PCC	Min PCC	Mean	STD	Max	Min
0	57	46	45,65	0,91	0,18	0,26	-0,17	0,62	1,04	-1,94
1	62	63	88,89	0,98	0,05	0,66	-0,17	0,47	1,88	-1,24
2	104	97	93,81	0,98	0,05	0,64	-0,13	0,46	0,87	-1,64
Overall	249	235	72,34	0,96	0,15	0,26	-0,15	0,51	1,88	-1,94

Influence of the Sampling Frequency in EDA signals

In Table 6.4, it is possible to see a comparison in terms of the number of detected onsets, time difference, amplitude error and PCC between the best interpolation method, cubic spline, and no interpolation (in the PCC comparison, the linear interpolation was used instead of the no interpolation). Furthermore, 6393 onsets were detected in the original signal.

Figure 6.3 presents the comparison between the different interpolation methods (no interpolation, linear spline, quadratic spline and cubic spline), for the same 10Hz DS, based on the box plots of the time difference (Figure 6.3a) and the box plot of the amplitude difference (Figure 6.3b).

Table 6.4: Comparison between time difference, amplitude difference and PCC distribution metrics for no interpolation and for interpolation by cubic splines in the EDA signal. The values for the different Downsampled Frequency (DS) are shown, as well as mean and STD for each distribution. To compare PCC values, linear spline interpolation values were used.

	Onsets		Time error (ms)				Amplitude error (nS)				PCC			
Interpolation	None	CubicSpl.	None		CubicSpl.		None		CubicSpl.		Linear		CubicSpl.	
DS (Hz)	Counts	Counts	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
1	4275	5844	-66.3	294.4	-72.0	186.5	11.6	26.2	1.0	33.7	0.94	0.10	0.97	0.07
10	6342	6392	-0.9	28.8	0.0	0.2	0.2	0.4	0.0	0.0	1.00	0.00	1.00	0.00
20	6377	6393	-0.1	14.4	0.0	0.1	0.0	0.1	0.0	0.0	1.00	0.00	1.00	0.00
50	6393	6393	0.0	5.8	0.0	0.0	0.0	0.0	0.0	0.0	1.00	0.00	1.00	0.00
100	6393	6393	0.0	2.9	0.0	0.0	0.0	0.0	0.0	0.0	1.00	0.00	1.00	0.00
200	6393	6393	0.0	1.4	0.0	0.0	0.0	0.0	0.0	0.0	1.00	0.00	1.00	0.00
500	6393	6393	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	1.00	0.00	1.00	0.00

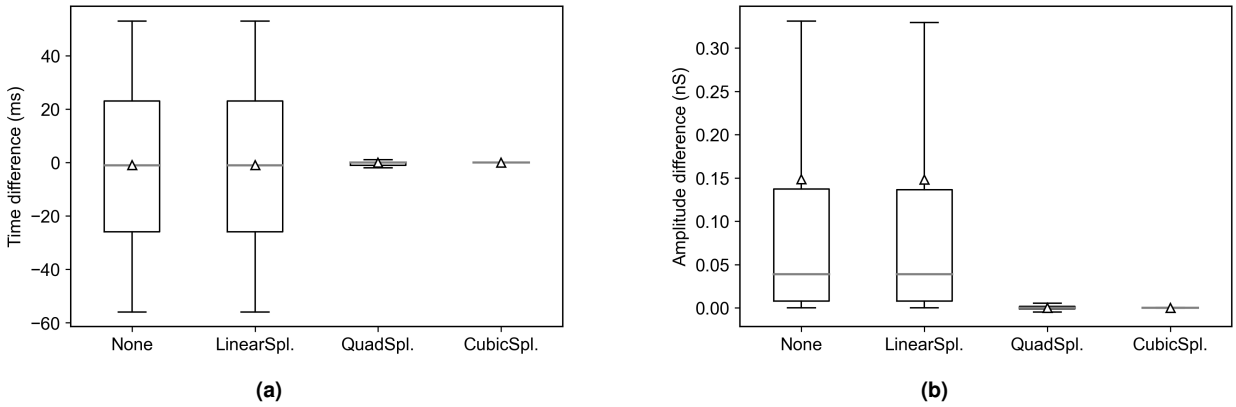


Figure 6.3: Box-plots of the EDA time (a) and amplitude (b) difference from the 1kHz signal with the different interpolation methods using a DS of 10Hz.

6.5 Discussion

Benchmarking of the BITalino R-IoT

Figure 6.2a illustrates the synchronization performed based on the acceleration data. In the first time instant it is possible to see an a very prominent peak in the acceleration data of both devices, while the rest of the time the acceleration signals have small oscillations between very defined ranges. As such, the initial peaks can be use to synchronize data from both devices marking the beginning of the acquisition phase. From Figure 6.2b it is possible to see an almost perfect synchronization between the two PPG signals, with the peak time differences being very close to 0.

Based on Table 6.2 it is possible to see that the PPG results achieved with both devices are very similar. In terms of the number of detected peaks these are very close to each other, with about 95% of the detected peaks being correctly matched between the two signals. Furthermore, the analysis on the waveform similarities around each detected peak show a mean PCC value very close to 1, indicating an almost perfect similarity in waveform of the PPG signals acquired with both devices. Even though, the minimum PCC is close to 0 in the first 2 participants and smaller than 0 in the last participant, which demonstrates almost no correlation in these areas, the STD is very small in all participants which indicates a very low deviation from the mean value. All time difference values across the table are considered to be very small, with the maximum and minimum time differences being 0.1 and -0.1 s, which expresses a negligible time deviation in peak detection in the signals from the 2 devices. Lastly, the differences in the HR extracted from each PPG signal are in the order of magnitude of 10^{-1}, which are also negligible considering the range of values for this metric.

Regarding the EDA signal comparison, Table 6.3 encompasses the results obtained. In this table it is possible to see that there are some differences in the number of detected onsets and % of matched onsets in participant 0, although in participants 1 and 2 the number of onsets are very close to each

other, with about 90% matching onsets. Overall, the number of onsets detected with each device is very similar, with the great majority of the detected onsets being matched between the two devices. In terms of the waveform similarity, the mean PCC are above 0.9 for all participants, with an overall value of 0.94, which indicates an almost perfect similarity in waveform of the acquired EDA signals. The overall minimum PCC is observed in participant 0, even though these value still demonstrates some similarities in waveform. The observed mean EDA time differences are all considerably close to 0, showing a good data correspondence across devices.

Lastly, it is possible to see that the mean time differences in both the PPG and EDA signals are all negative, suggesting that the BITalino R-IoT acquires data with a small delay when compared to the BITalino (r)evolution.

Influence of the Sampling Frequency in EDA signals

Based on Table 6.4, it is possible to see an improvement in the results from 1 to 10 Hz DS in both interpolation methods, furthermore, one can observe that the results obtained with the cubic spline are consistently better than with no interpolation methods. The number of onsets almost reaches the total number of the original signal. The mean and STD time differences reach values very close to 0 ms for the cubic spline; the no interpolation method still has a considerably high STD value. Regarding the amplitude differences, although they were already considerably small with the 1Hz DS (since these are represented in nS), with the 10 Hz DS these differences reach values of 0 for the cubic spline and values close to 0 for the no interpolation method. Regarding the cubic spline, the following frequencies of 20, 50 and 100Hz show minor improvements concerning the 10 Hz reaching, in the latter frequency, both time and amplitude errors of 0 across all criteria.

The remaining interpolation methods follow the same pattern as the previously mentioned. However, in the case of the linear spline, there is no frequency for which the differences are all 0, similarly to the no interpolation case; in the quadratic spline, this only occurs with the 500Hz downsampling frequency. Furthermore, comparing the same frequency across all interpolation methods (as it is possible to see in Figure 6.3 for the 10Hz DS). The best results are achieved using the cubic spline, closely followed by the quadratic spline, with a great difference in amplitude; the following methods are the linear spline and the no interpolation, being that the linear spline is still marginally better than the no interpolation.

In Table 6.4, it is possible to see the PCC for the linear interpolation case and the cubic spline. The experimental results show an improvement from from 1 to 10Hz downsampling frequencies, as well from the linear spline to the cubic spline. Although the results from the 1Hz frequency were already relatively good in both scenarios, since the mean is very close to 1 and SD close to 0; the results from the 10 Hz frequency have a mean value of exactly 1 with a SD of 0. The frequencies above 10Hz show perfect results across the table (with regards to the PCC). The quadratic interpolation methods follows the same

pattern as the one previously described.

A 10Hz signal interpolated with cubic splines presented accurate time and amplitude onset matching with the 1kHz signal. Minimal distortion values are observed for 10Hz sampling rate and above. Hence, the minimum SF recommended for a quality EDA acquisition is 10Hz; this value was selected based on the trade-off between having a low sampling rate and an accurate measurement of the desired fiducials. However, the 1Hz signal interpolated with cubic splines also achieved good results given the order of magnitude of the measurements extracted from this signal i.e. latency time, rise time, amplitude, which were presented in Section 2.4.

7

Conclusions

Throughout the current work, several problems related with real world group emotional analytics were addressed. Regarding the group emotion analysis (Chapter 4, Analysis of the Synchrony between Annotations), the annotations performed by the participants represent mainly neutral states, which was not expected given that the movie consisted of a high-pass superhero movie, containing several fight scenes, along with some emotional and comical parts. This revealed a lack of comprehension of the annotation's scales by the participants, a lack of engagement towards the content and/or the annotation task. Furthermore, the evaluation of the EDA signal in simultaneous annotations revealed a tendency to increase over the period of the annotations (which was not observed in the participants who did not annotate in the same period of time). Nevertheless, the signals during simultaneous annotations displayed few waveform similarities.

To overcome the annotation tool limitations, an emotional analysis solely based on the acquired physiological signals and movie content was performed (Chapter 4, Collective Intelligence Analysis). This analysis was conducted by extracting features from the mean EDA signal of the group and applying clustering algorithms to group the areas of the movie where the audience experienced a similar emotional reaction. The clustering results were then compared with the literature, namely the MAP. Based on this analysis it was possible to conclude that best performing methodology was hierarchical clustering with average linkage. This clustering methodology provides a higher number of areas in which the audience had a more intense emotional reaction, divided into two distinct clusters with 7 and 15 emotional time regions in each one. Furthermore, within the areas in which the audience had a more intense emotional reaction, this method also provides a differentiation in the intensity of the reaction with one cluster having a mean MAP of 4.84E-04 and the other cluster having a mean MAP of 1.73E-04.

In terms of real-time content annotation tool (Chapter 5), the results corroborate what was already seen in the literature, i.e. rank unbounded annotation is the most promising approach for annotation of previously uncalibrated and unseen content with higher reliability. Furthermore, a smartphone-based annotation tool was proposed, and displayed a high usability and a low mental workload, thus providing a reliable emotional annotation with minimal distraction. Although both versions achieved good results, the OSMA version was considered to be the preferred version.

To address existing limitations in the real-world collective data acquisition, the BITalino R-IoT was evaluated; this device revealed to be a great asset for the acquisition of physiological data in collective environments, being able to collect two physiological signals simultaneously across several member of an audience. With the use of this device it would be possible to collect EDA and PPG data, thus providing a window of information to the Valence dimension of emotions. Based on the evaluation of the minimum SF required for the acquisition of the EDA signal showed that the recommended SF is 10 Hz. This value was determined based on a trade-off between having a low sampling rate and an accurate measurement of the signal. This evaluation of the minimum SF and benchmarking of a new device able

of acquiring one additional physiological signal in a group setting, establishes a path to future works in the area of group emotion recognition.

Even though the recommend EDA SF was 10 Hz and the emotional analysis performed in this work as based on the acquisition of EDA data with 1 Hz, the results and conclusions obtained were still considered to be valid. First of all, the time and amplitude errors achieved with the EDA data downsampled to 1 Hz are relatively low when compared to the measurements extracted from this signal i.e. latency time, rise time, amplitude. Given the adverse circumstances (COVID-19 pandemic) in which this work was performed, the data acquisition tasks had to be performed remotely which coupled with the complexity of the problem (i.e. acquisition in group context requiring several participants in the same location simultaneously, especially during the pandemic situation; few acquisition device options and complicated experimental set-up) and submissions to the ethics committee made these the best acquisitions that could have been made. However, the current work gives a lead for futures work by evaluating the performance of a new acquisition device for future acquisition, along with an experimental protocol and a list of elicitation content to be used during such acquisitions.

Overall, the current work fulfilled the objectives drawn at the beginning, expanding the state-of-the-art by developing a new self-assessment tool and implementing machine learning methods to emotional assessment in a collective setting, namely of the audience EDA signal. Future work will focus on expanding the database using the protocol developed with the BITalino R-IoT acquiring EDA and PPG data; applying the developed emotional analysis method, namely the use of clustering algorithms on features extracted from physiological signals, on other movies and further validate the annotation tool developed.

Bibliography

[1] P. Schmidt, A. Reiss, R. Duerichen, and K. Van Laerhoven, "Wearable affect and stress recognition: A review," *CoRR*, vol. 19, no. 19, 2018.

[2] J. Domínguez-Jiménez, K. Campo-Landines, J. Martínez-Santos, E. Delahoz, and S. Contreras-Ortiz, "A machine learning model for emotion recognition from physiological signals," *Biomedical Signal Processing and Control*, vol. 55, p. 101646, 2020.

[3] M. Bradley and P. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[4] W. Boucsein, *Electrodermal activity*. Springer Science & Business Media, 2012.

[5] P. Bota, P. Cesar, H. Silva, and A. Fred, "Unveiling the potential of retrospective ground-truth collection for affective computing."

[6] C. Wang and P. Cesar, "Measuring audience responses of video advertisements using physiological sensors," in *Proc. of the Int'l Workshop on Immersive Media Experiences*, 2015, p. 37–40.

[7] P. Bota, C. Wang, A. Fred, and H. Silva, "A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals," *IEEE Access*, vol. 7, pp. 140 990–141 020, 2019.

[8] J. Lewis and J. Sauro, "Item benchmarks for the system usability scale," *J. Usability Studies*, vol. 13, pp. 158–167, 2018.

[9] R. Grier, "How high is high? a meta-analysis of NASA-TLX global workload scores," in *Proc.of the Human Factors and Ergonomics Society Annual Meeting*, vol. 59, 2015, pp. 1727–1731.

[10] S. Basu, A. Bag, M. Mahadevappa, J. Mukherjee, and R. Guha, "Affect detection in normal groups with the help of biological markers," in *Proc. of the Annual IEEE India Conf.*, 2015, pp. 1–6.

[11] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion recognition: a review," in *Proc. of the IEEE Int'l Colloq. on Signal Processing and its Applications*, 2011, pp. 410–415.

[12] P. Ekman, “An argument for basic emotions,” *Cogn. & Emot.*, vol. 6, no. 3-4, pp. 169–200, 1992.

[13] R. Plutchik, “Emotion: A psychoevolutionary synthesis,” *A psychoevolutionary synthesis*, 1980.

[14] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, “A review of emotion recognition using physiological signals,” *Sensors*, vol. 18, no. 7, p. 2074, 2018.

[15] J. Russell, “Affective space is bipolar.” *Journal of Personality and Social Psychology*, vol. 37, pp. 345–356, 1979.

[16] A. Mehrabian, “Framework for a comprehensive description and measurement of emotional states.” *Genetic, social, and general psychology monographs*, 1995.

[17] S. Barsade and D. Gibson, “Group affect: Its influence on individual and group outcomes,” *Current Directions in Psychological Science*, vol. 21, no. 2, pp. 119–123, 2012.

[18] G. Salvador, P. Bota, V. Vinayagamoorthy, H. Silva, and A. Fred, “Smartphone-based content annotation for ground truth collection in affective computing.” Association for Computing Machinery, 2021.

[19] C. Carreiras, A. Alves, A. Lourenço, F. Canento, A. Silva, H. and. Fred *et al.*, “BioSPPy: Biosignal processing in Python,” 2015–, [Online; accessed 14/10/2021]. [Online]. Available: https://github.com/PIA-Group/BioSPPy/

[20] D. Ellis and I. Tucker, *Social psychology of emotion*. Sage, 2015.

[21] W. Wundt, “Vorlesung über die menschen-und tierseele,” *Siebente und Achte Auflage*, 1922.

[22] A. Moors, P. Ellsworth, K. Scherer, and N. Frijda, “Appraisal theories of emotion: State of the art and future development,” *Emotion Review*, vol. 5, no. 2, pp. 119–124, 2013.

[23] S. Schmidt, C. Tinti, L. Levine, and S. Testa, “Appraisals, emotions and emotion regulation: An integrative approach,” *Motivation and emotion*, vol. 34, no. 1, pp. 63–72, 2010.

[24] Y. Hsu, J. Wang, W. Chiang, and C. Hung, “Automatic ECG-based emotion recognition in music listening,” *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 85–99, 2020.

[25] D. Ciuk, A. Troy, and M. Jones, “Measuring emotion: Self-reports vs. physiological indicators,” *Physiological Indicators*, 2015.

[26] A. Liapis, C. Katsanos, D. Sotiropoulos, M. Xenos, and N. Karousos, “Stress recognition in human-computer interaction using physiological and self-reported data: a study of gender differences,” in *Proc. of the Panhellenic Conference on Informatics*, 2015, pp. 323–328.

[27] J. Miranda, M. Khomami, N. Sebe, and I. Patras, “AMIGOS: A dataset for Mood, personality and affect research on Individuals and GrOupS,” *IEEE Transactions on Affective Computing*, 2017.

[28] A. Liapis, C. Katsanos, D. Sotiropoulos, M. Xenos, and N. Karousos, “Subjective assessment of stress in HCI: A study of the valence-arousal scale using skin conductance,” in *Proc. Int’l Conf. of Biannual Conference on Italian SIGCHI Chapter*, 2015, p. 174–177.

[29] J. Pollak, P. Adams, and G. Gay, “PAM: a photographic affect meter for frequent, in situ measurement of affect,” in *Proc. of the SIGCHI Conf. on Human factors in computing systems*, 2011, pp. 725–734.

[30] S. Shiffman, A. Stone, and M. Hufford, “Ecological Momentary Assessment (EMA),” *Annu. Rev. Clin. Psychol.*, vol. 4, pp. 1–32, 2008.

[31] E. Hayashi, J. Posada, V. Maike, and M. Baranauskas, “Exploring new formats of the self-assessment manikin in the design with children,” in *Proc.of the Brazilian Symposium on Human Factors in Computing Systems*, 2016, pp. 1–10.

[32] D. Watson, L. Clark, and A. Tellegen, “Development and validation of brief measures of positive and negative affect: the PANAS scales.” *Journal of personality and social psychology*, vol. 54, no. 6, p. 1063, 1988.

[33] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, “‘FEELTRACE’: An instrument for recording perceived emotion in real time,” in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[34] P. Lopes, G. Yannakakis, and A. Liapis, “Ranktrace: Relative and unbounded affect annotation,” in *Int’l Conf.on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 158–163.

[35] D. Melhart, A. Liapis, and G. Yannakakis, “PAGAN: Video affect annotation made easy,” in *Int’l Conf. on Affective Computing and Intelligent Interaction (ACII)*, 2019, pp. 130–136.

[36] M. Wirth and A. Gaffey, “Hormones and emotion.” *Handbook of cognition and emotion*, vol. 69, 2013.

[37] E. Jang, B. Park, M. Park, and S. Kim, “Analysis of physiological signals for recognition of boredom, pain, and surprise emotions,” *Journal of physiological anthropology*, vol. 34, no. 1, pp. 1–12, 2015.

[38] C. Stangor, R. Jhangiani, H. Tarry *et al.*, *Principles of social psychology*. Psychology Press, 2014.

[39] J. Waxenbaum, V. Reddy, and M. Varacallo, “Anatomy, autonomic nervous system,” *StatPearls*, 2019.

[40] R. Martinez, A. Salazar-Ramirez, A. Arruti, E. Irigoyen, J. Martin, and J. Muguerza, "A self-paced relaxation response detection system based on galvanic skin response analysis," *IEEE Access*, vol. 7, pp. 43 730–43 741, 2019.

[41] P. Lakhan, N. Banluesombatkul, V. Changniam *et al.*, "Consumer grade brain sensing for emotion recognition," *IEEE Sensors Journal*, vol. 19, no. 21, pp. 9896–9907, 2019.

[42] T. Li, Y. Baveye, C. Chamaret, E. Dellandréa, and L. Chen, "Continuous arousal self-assessments validation using real-time physiological responses," in *Proc.of the Int'l Workshop on Affect & Sentiment in Multimedia*, 2015, pp. 39–44.

[43] P. Bota, C. Wang, A. Fred, and H. Silva, "A wearable system for electrodermal activity data acquisition in collective experience assessment." in *Proc. of the Int'l Conf. on Enterprise Information Systems: ICEIS*, 2020, pp. 606–613.

[44] R. Laureanti, M. Bilucaglia, M. Zito *et al.*, "Emotion assessment using machine learning and low-cost wearable devices," in *Int'l Conf. of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 576–579.

[45] J. Fleureau, P. Guillotel, and I. Orlac, "Affective benchmarking of movies based on the physiological responses of a real audience," in *Humaine Association Conf. on Affective Computing and Intelligent Interaction*. IEEE, 2013, pp. 73–78.

[46] V. Ojha, D. Griego, S. Kuliga *et al.*, "Machine learning approaches to understand the influence of urban environments on human's physiological response," *Information Sciences*, vol. 474, pp. 154–169, 2019.

[47] A. Banganho, M. Santos, and H. Silva, "Design and evaluation of an electrodermal activity sensor (EDA) with adaptive gain," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 8639–8649, 2021.

[48] D. Fowles, M. Christie, R. Edelberg, W. Grings, D. Lykken, and P. Venables, "Publication recommendations for electrodermal measurements," *Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures*, vol. 18, no. 3, pp. 232–239, 1981.

[49] P. Venables, S. Gartshore, and P. O'Riordan, "The function of skin conductance response recovery and rise time," *Biological Psychology*, vol. 10, no. 1, pp. 1–6, 1980.

[50] P. Lang, M. Bradley, B. Cuthbert *et al.*, "International affective picture system (IAPS): Technical manual and affective ratings," *NIMH Center for the Study of Emotion and Attention*, vol. 1, pp. 39–58, 1997.

[51] E. Dan-Glauser and K. Scherer, “The geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance,” *Behavior research methods*, vol. 43, no. 2, pp. 468–477, 2011.

[52] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, “LIRIS-ACCEDE: A video database for affective content analysis,” *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43–55, 2015.

[53] E. Douglas-Cowie, R. Cowie, I. Sneddon *et al.*, “The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data,” in *Int’l Conf. on affective computing and intelligent interaction*, 2007, pp. 488–500.

[54] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, “A multimodal database for affect recognition and implicit tagging,” *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 42–55, 2011.

[55] S. Koelstra, C. Muhl, M. Soleymani *et al.*, “DEAP: A database for emotion analysis; using physiological signals,” *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[56] J. Lee and S. Yoo, “Design of user-customized negative emotion classifier based on feature selection using physiological signal sensors,” *Sensors*, vol. 18, no. 12, p. 4253, 2018.

[57] E. Hatfield, J. Cacioppo, and R. Rapson, “Emotional contagion,” *Current Directions in Psychological Science*, vol. 2, no. 3, pp. 96–100, 1993.

[58] S. Rhee, “Group emotions and group outcomes: The role of group-member interactions,” in *Affect and groups*, 2007.

[59] T. Sy, J. Choi, and S. Johnson, “Reciprocal interactions between group perceptions of leader charisma and group mood through mood contagion,” *The Leadership Quarterly*, vol. 24, no. 4, pp. 463–476, 2013.

[60] A. Haag, S. Goronzy, P. Schaich, and J. Williams, “Emotion recognition using bio-sensors: First steps towards an automatic system,” in *Tutorial and research workshop on affective dialogue systems*, 2004, pp. 36–48.

[61] M. Ragot, N. Martin, S. Em, N. Pallamin, and J. Diverrez, “Emotion recognition using physiological signals: laboratory vs. wearable sensors,” in *Int’l Conf. on Applied Human Factors and Ergonomics*, 2017, pp. 15–22.

[62] H. Posada-Quintero and K. Chon, “Innovations in electrodermal activity data collection and signal processing: A systematic review,” *Sensors*, vol. 20, no. 2, p. 479, 2020.

[63] A. Greco, G. Valenza, A. Lanata, E. Scilingo, and L. Citi, "cvxEDA: A convex optimization approach to electrodermal activity processing," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 797–804, 2015.

[64] H. Silva, A. Fred, S. Eusébio, M. Torrado, and S. Ouakinin, "Feature extraction for psychophysiological load assessment in unconstrained scenarios," in *Int'l Conf. of the IEEE Engineering in Medicine and Biology Society*, 2012, pp. 4784–4787.

[65] H. Silva, A. Fred, and A. Lourenço, "Electrodermal response propagation time as a potential psychophysiological marker," in *Int'l Conf. of the IEEE Engineering in Medicine and Biology Society*, 2012, pp. 6756–6759.

[66] R. Singh, S. Conjeti, and R. Banerjee, "A comparative evaluation of neural network classifiers for stress level analysis of automotive drivers using physiological signals," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 740–754, 2013.

[67] S. Eskimez, Z. Duan, and W. Heinzelman, "Unsupervised learning approach to feature analysis for automatic speech emotion recognition," in *IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5099–5103.

[68] J. Huelle, B. Sack, K. Broer, I. Komlewa, and S. Anders, "Unsupervised learning of facial emotion decoding skills," *Frontiers in human neuroscience*, vol. 8, p. 77, 2014.

[69] Z. Zhang and E. Tanaka, "Affective computing using clustering method for mapping human's emotion," in *IEEE Int'l Conf. on Advanced Intelligent Mechatronics (AIM)*, 2017, pp. 235–240.

[70] M. Ackerman and S. Ben-David, "A characterization of linkage-based hierarchical clustering," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 8182–8198, 2016.

[71] A. Fred and A. Jain, "Data clustering using evidence accumulation," in *Int'l Conf. on Pattern Recognition*, vol. 4, 2002, pp. 276–280 vol.4.

[72] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86–97, 2012.

[73] J. Yadav and M. Sharma, "A review of K-mean algorithm," *Int'l J. Eng. Trends Technol*, vol. 4, no. 7, pp. 2972–2976, 2013.

[74] S. Smith, *The scientist and engineer's guide to digital signal processing*. California Technical Publishing, 1997.

[75] K. Kim, S. Bang, and S. Kim, “Emotion recognition system using short-term monitoring of physiological signals,” *Medical and biological engineering and computing*, vol. 42, no. 3, pp. 419–427, 2004.

[76] T. Zhang, A. Ali, C. Wang, A. Hanjalic, and P. Cesar, “RCEA: Real-time, continuous emotion annotation for collecting precise mobile video ground truth labels,” in *Proc. of CHI Conf. on Human Factors in Computing Systems*, 2020, pp. 1–15.

[77] R. Cowie, M. Sawey, C. Doherty, J. Jaimovich, C. Fyans, and P. Stapleton, “Gtrace: General trace program compatible with emotionML,” in *Humaine Association Conf. on Affective Computing and Intelligent Interaction*, 2013, pp. 709–710.

[78] J. Girard, “CARMA: Software for continuous affect rating and media annotation,” *Journal of open research software*, vol. 2, no. 1, 2014.

[79] K. Sharma, C. Castellini, F. Stulp, and E. Broek, “Continuous, real-time emotion annotation: A novel joystick-based analysis framework,” *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 78–84, 2017.

[80] J. Girard and A. Wright, “DARMA: Software for dual axis rating and media annotation,” *Behavior research methods*, vol. 50, no. 3, pp. 902–909, 2018.

[81] G. Yannakakis and H. Martinez, “Grounding truth via ordinal annotation,” in *Int’l Conf. on affective computing and intelligent interaction (ACII)*, 2015, pp. 574–580.

[82] A. Muaremi, B. Arnrich, and G. Tröster, “Towards measuring stress with smartphones and wearable devices during workday and sleep,” *BioNanoScience*, vol. 3, no. 2, pp. 172–183, 2013.

[83] S. McLellan, A. Muddimer, and S. Peres, “The effect of experience on system usability scale ratings,” *Journal of Usability Studies*, vol. 7, no. 2, pp. 56–67, 2012.

[84] J. Lewis, “The system usability scale: past, present, and future,” *Int’l Journal of Human–Computer Interaction*, vol. 34, no. 7, pp. 577–590, 2018.

[85] S. Hart and L. Staveland, “Development of NASA-TLX (task load index): Results of empirical and theoretical research,” in *Human Mental Workload*, 1988, vol. 52, pp. 139–183.

[86] A. Oppenheim, A. Willsky, and S. Nawab, *Signals Systems*. Prentice-Hall, Inc., 1996.

[87] K. Bhutta and F. Huq, “Benchmarking-best practices: an integrated approach,” *Benchmarking: An International Journal*, 1999.

[88] I. Kleckner, M. Feldman, M. Goodwin, and K. Quigley, "Framework for selecting and benchmarking mobile devices in psychophysiological research," *Behavior Research Methods*, vol. 53, no. 2, pp. 518–535, 2021.

[89] D. Batista, H. Silva, A. Fred, C. Moreira, M. Reis, and H. Ferreira, "Benchmarking of the BITalino biomedical toolkit against an established gold standard," *Healthcare technology letters*, vol. 6, no. 2, pp. 32–36, 2019.

[90] D. Batista, H. Silva, and A. Fred, "Experimental characterization and analysis of the BITalino platforms against a reference device," in *Int'l Conf. of the IEEE Engineering in Medicine and Biology Society*, 2017, pp. 2418–2421.

[91] "Heart rate variability: standards of measurement, physiological interpretation and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

[92] R. Ellis, B. Zhu, J. Koenig, J. Thayer, and Y. Wang, "A careful look at ECG sampling frequency and R-peak interpolation on short-term measures of heart rate variability," *Physiological Measurement*, vol. 36, no. 9, pp. 1827–1852, 2015.

[93] T. Ziemssen, J. Gasch, and H. Ruediger, "Influence of ECG sampling frequency on spectral analysis of RR intervals and baroreflex sensitivity using the EUROBAVAR data set," *Journal of Clinical Monitoring and Computing*, vol. 22, no. 2, p. 159, 2008.

[94] S. Béres and L. Hejjel, "The minimal sampling frequency of the photoplethysmogram for accurate pulse rate variability parameters in healthy volunteers," *Biomedical Signal Processing and Control*, vol. 68, p. 102589, 2021.

[95] E. Ramos, H. Silva, B. Olstad, J. Cabri, and P. Lobato, "SwimBIT: A novel approach to stroke analysis during swim training based on attitude and heading reference system (AHRS)," *Sports (Basel, Switzerland)*, vol. 7, no. 11, 2019.

[96] M. Elgendi, I. Norton, M. Brearley, D. Abbott, and D. Schuurmans, "Systolic peak detection in acceleration photoplethysmograms measured from emergency responders in tropical conditions," *PLoS One*, vol. 8, no. 10, 2013.

[97] J. Valente, V. Kozlova, and T. Pereira, "BrainAnswer platform: Biosignals acquisition for monitoring of physical and cardiac conditions of older people," in *Promoting Healthy and Active Aging*. Routledge.

[98] H. Silva, A. Fred, and R. Martins, "Biosignals for everyone," *IEEE Pervasive Computing*, vol. 13, no. 4, pp. 64–71, 2014.

[99] P. Virtanen, R. Gommers, T. Oliphant *et al.*, “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python,” *Nature Methods*, vol. 17, pp. 261–272, 2020.

Appendix

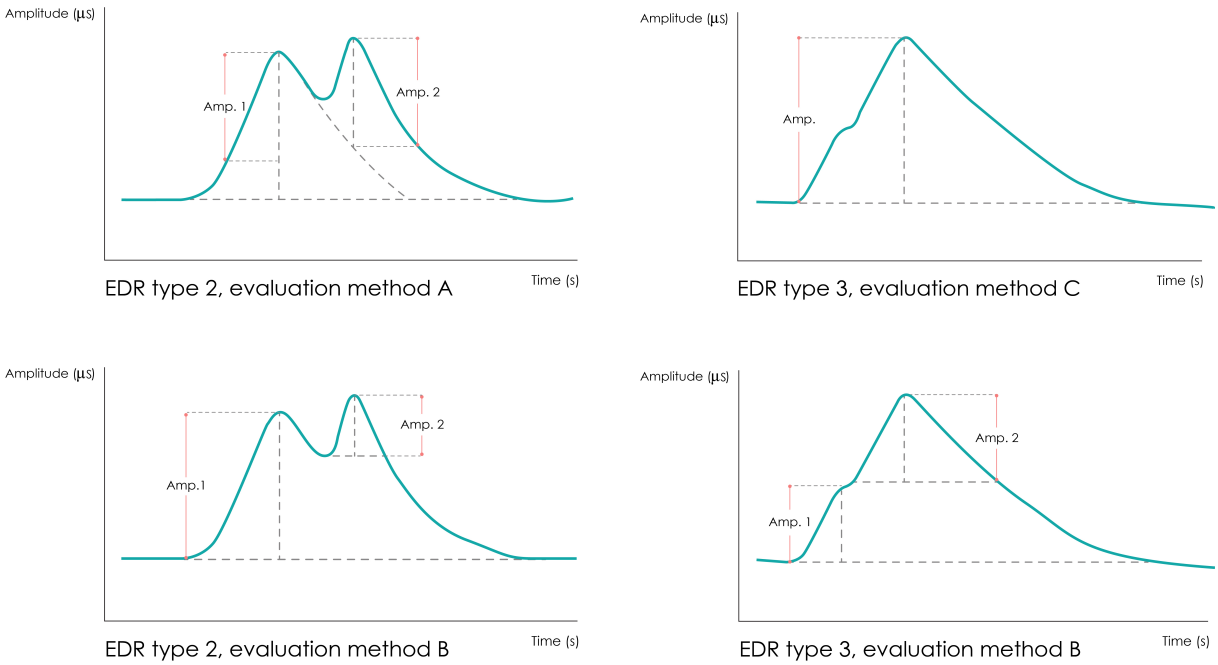


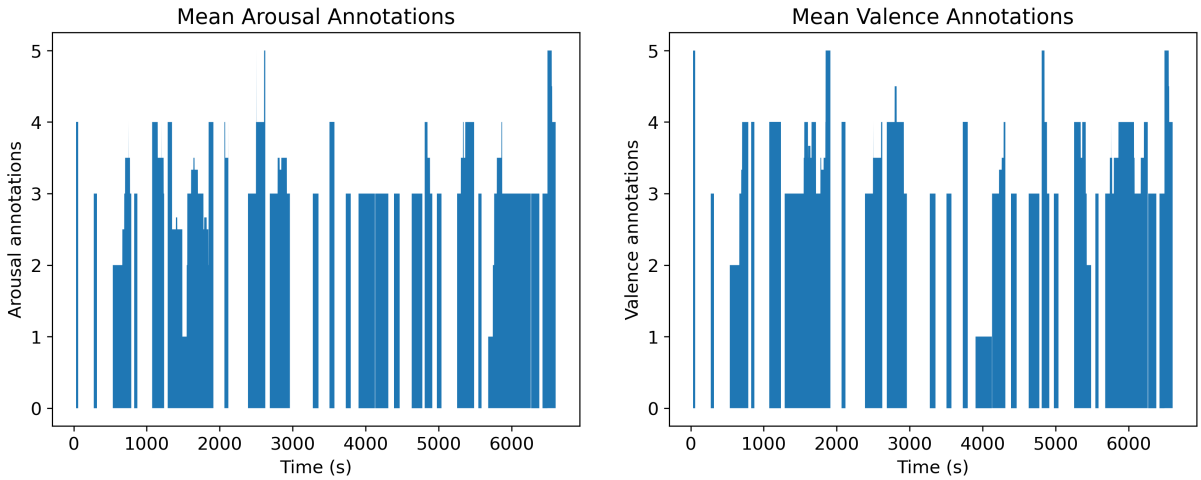
Figure A.1: Examples of overlapped exosomatic EDR and method to estimate amplitudes. Adapted from [4].

Table A.1: Average SUS score per question.

Question	TSSA $\mu \pm \sigma$	OSMA $\mu \pm \sigma$
I would repeat the experience	4.3±0.6	4.3±0.8
I found the system unnecessarily complex.	2.0±0.9	1.9±1.0
I thought the system was easy to use.	4.7±0.5	4.4±0.9
I think that I would need the support of a technical person to be able to use this system.	1.3±0.6	1.3±0.8
I found the various functions in this system were well integrated.	4.2±0.6	4.5±0.7
I thought there was too much inconsistency in this system.	1.6±0.7	1.9±1.1
I would imagine that most people would learn to use this system very quickly.	4.8±0.4	4.3±1.0
I found the system very cumbersome to use.	2.4±1.0	2.0±1.1
I felt very confident using the system.	4.0±0.8	4.3±0.8
I needed to learn a lot of things before I could get going with this system.	1.6±0.8	1.6±0.9
Final results.	82.75±5.29	82.66±8.25

Table A.2: Average NASA-RTLX score per question.

Question	TSSA $\mu \pm \sigma$	OSMA $\mu \pm \sigma$
How Mentally Demanding was using the app while watching the movie?	2.9±1.2	2.1±1.4
How Mentally Demanding was using the app while watching the movie?	3.6±1.6	2.6±1.6
How hurried or rushed was the pace of annotating your emotions using the app?	4±0.9	3.6±1.4
How successful were you in accomplishing what you were asked to do?	3.2±1.5	3.5±2.1
How hard did you have to work to accomplish your level of performance?	2.7±1.0	1.9±0.7
How insecure, discouraged, irritated, stressed, and annoyed were you using the app?	2.3±1.1	2.1±1.2
Final results.	44.52±7.26	37.50±6.77



(a) **(b)**

Figure A.2: Mean Arousal (A.2a) and Valence (A.2b) annotations throughout the duration of the movie

Table A.3: EDA time errors, amplitude errors and Pearson correlation coefficient values between downsampled (DS) and original 1kHz signal for different interpolation methods and sampling frequencies. The number of peak counts in each essay is presented as well as mean, STD, minimum (min) and maximum (max) values. The Pearson correlation coefficient values and kurtosis of the estimated distribution are also described.

Interpolation method: None									
DS (Hz)	Counts	Time error (ms)				Amplitude error (nS)			
		mean	STD	min	max	mean	STD	min	max
1	4275	-66.3	294.4	-975	942	11.6	26.2	-19	498
10	6342	-0.9	28.8	-56	53	0.2	0.4	0	8
20	6377	-0.1	14.4	-26	25	0.0	0.1	0	3
50	6393	0.0	5.8	-11	10	0.0	0.0	0	0
100	6393	0.0	2.9	-5	5	0.0	0.0	0	0
200	6393	0.0	1.4	-3	2	0.0	0.0	0	0
500	6393	0.0	0.7	-1	1	0.0	0.0	0	0

Interpolation method: Linear splines														
DS (Hz)	Counts	Time error (ms)				Amplitude error (nS)				Pearson Correlation Coefficients				
		mean	STD	min	max	mean	STD	min	max	mean	STD	min	max	kurtosis
1	5341	-60.8	295.6	-975	942	11.8	26.3	-19	498	0.94	0.10	-0.6	1.0	24
10	6375	-1.0	28.8	-56	53	0.2	0.4	0	8	1.00	0.00	1.0	1.0	979
20	6387	-0.1	14.4	-26	25	0.0	0.1	0	3	1.00	0.00	1.0	1.0	994
50	6392	0.0	5.8	-11	10	0.0	0.0	0	0	1.00	0.00	1.0	1.0	995
100	6393	0.0	2.9	-5	5	0.0	0.0	0	0	1.00	0.00	1.0	1.0	995
200	6393	0.0	1.4	-3	2	0.0	0.0	0	0	1.00	0.00	1.0	1.0	995
500	6393	0.0	0.7	-1	1	0.0	0.0	0	0	1.00	0.00	1.0	1.0	995

Interpolation method: Quadratic splines														
DS (Hz)	Counts	Time error (ms)				Amplitude error (nS)				Pearson Correlation Coefficients				
		mean	STD	min	max	mean	STD	min	max	mean	STD	min	max	kurtosis
1	5878	-80.1	198.5	-1000	963	1.8	32.7	-987	497	0.97	0.08	-0.7	1.0	45
10	6391	0.0	1.5	-35	27	0.0	0.0	-1	0	1.00	0.00	1.0	1.0	995
20	6392	0.0	0.5	-5	5	0.0	0.0	0	0	1.00	0.00	1.0	1.0	995
50	6393	0.0	0.2	-1	3	0.0	0.0	0	0	1.00	0.00	1.0	1.0	995
100	6393	0.0	0.1	-1	1	0.0	0.0	0	0	1.00	0.00	1.0	1.0	995
200	6393	0.0	0.0	-1	1	0.0	0.0	0	0	1.00	0.00	1.0	1.0	995
500	6393	0.0	0.0	0	0	0.0	0.0	0	0	1.00	0.00	1.0	1.0	995

Interpolation method: Cubic splines														
DS (Hz)	Counts	Time error (ms)				Amplitude error (nS)				Pearson Correlation Coefficients				
		mean	STD	min	max	mean	STD	min	max	mean	STD	min	max	kurtosis
1	5844	-72.0	186.5	-997	991	1.0	33.7	-1097	497	0.97	0.07	-0.6	1.0	49
10	6392	0.0	0.2	-2	1	0.0	0.0	0	0	1.00	0.00	1.0	1.0	995
20	6393	0.0	0.1	-1	1	0.0	0.0	0	0	1.00	0.00	1.0	1.0	995
50	6393	0.0	0.0	0	1	0.0	0.0	0	0	1.00	0.00	1.0	1.0	995
100	6393	0.0	0.0	0	0	0.0	0.0	0	0	1.00	0.00	1.0	1.0	995
200	6393	0.0	0.0	0	0	0.0	0.0	0	0	1.00	0.00	1.0	1.0	995
500	6393	0.0	0.0	0	0	0.0	0.0	0	0	1.00	0.00	1.0	1.0	995