TÉCNICO LISBOA

Using Biometric Signals and Virtual Reality to Evaluate Movies and/or TV Shows

Daniel Filipe Garcia Gonçalves

Thesis to obtain the Master of Science Degree in

Computer Science and Engineering

Supervisor: Prof. Augusto Emanuel Abreu Esteves

Examination Committee

Chairperson: Prof. Rui Filipe Fernandes Prada
Supervisor: Prof. Augusto Emanuel Abreu Esteves
Member of the Committee: Prof. Daniel Simões Lopes

November 2022

This work was created using LaTeX typesetting language
in the Overleaf environment (www.overleaf.com).

Acknowledgments

Every time I look to this document I can't avoid thinking about every hour writing these letters, every bug-fixing nightmare turned into intense joy by that "eureka" moment that followed, every reaction of my friends and colleagues when I explained this concept. Even though I've had my ups and downs, my long nights trying to sleep while thinking about how to solve a specific issue, this roller coaster ended, without a doubt, at the end of the biggest climb. While writing this dissertation, every word seemed to make the work behind it more and more worth it. I was finally able to present something that started as a personal necessity and ended up as mature concept, shared with others and out in the open.

First I would like to thank my supervisor **Professor Augusto Esteves**, who handled my almost existential doubts with a high degree of finesse and was able to guide me during this long and new process. For every critical question, every direct and straight to the point answer, I can only express my deep appreciation.

To my parents I can only try to find words, because that is a hard enough task. Every bug-related scream, every existential crisis, even every meal kindly brought to me while I was spending long hours working on this project, always making sure that I was not disturbed by that attitude that meant the exact opposite than a disturbance of any kind. Twenty three years later I still do not have words, maybe I will find them in the next decades.

I would also like to thank something that is more than a single person: Theater. For almost ten years those old black planks have been my life raft in a sea of doubts and insecurity, and it was not different in the last year. The only thing that forced me to leave my PC and take a breath were the rehearsals, and for that, I can't say enough. Art has been part of my life since I can remember and many of this dissertation's insights were made while letting art flood my brain and compartmentalize everything else.

Lastly, I want to thank Instituto Superior Técnico, an institution that received me as one of its own, even with me spending my first three years of academic formation elsewhere. I never felt anything but unconditional support, well, maybe some headaches. What matters is, even when the biggest nightmares were caused by it, I knew that a love-hate relationship is always funnier.

Abstract

Every multimedia asset has one major goal: Invoke an emotional reaction on the viewer. Furthermore, they are even judged for how well they perform this task. Movies and TV Shows are some of the major types of multimedia content that are fully structured and developed to attend to the viewers' emotional needs or to enter their personal world by promoting some kind of psychological influence.

With the growth of Virtual Reality, several studies have been made about what the future of this area will be in regards to the movie industry and how it is expected to become the main approach to watching movies, even when compared to the cinema.

When trying to evaluate the impact of a movie next to the audience, the industry tends to use questionnaires or other types of subjective information gathering methods, resulting in ambiguous feedback and wrong content classification. In this thesis, a new approach is presented with the intent to to use biometric signals taken directly from the viewers, at the precise moment of watching some video content, and use it to measure the viewer's arousal during the whole watching experience, providing a new data point to be used for multimedia artifact classification. Virtual Reality is also a decisive factor, since it will guarantee greater presence, less distraction induced measuring errors and a glimpse into the future of the industry.

Keywords

Movie; TV Show; Arousal; Virtual Reality; Biometric Classification; Emotion Recognition.

Resumo

Todos os conteúdos multimédia têm um objetivo principal: provocar uma reação emocional no espectador. Além disso, estes são ainda julgados por quão bem realizam esta tarefa. Filmes e Séries são alguns dos principais tipos de conteúdo multimédia que são totalmente estruturados e desenvolvidos para atender às necessidades emocionais da audiência ou para entrar no seu mundo pessoal, promovendo algum tipo de influência psicológica.

Com o crescimento da Realidade Virtual, diversos artigos têm sido feitos sobre qual será o futuro desta área no que diz respeito à indústria cinematográfica e como se espera que esta se torne a principal abordagem para assistir filmes e séries, mesmo quando comparada ao cinema.

Ao tentar avaliar o impacto de um filme junto ao público, a indústria tende a usar questionários ou outros tipos de métodos de colheita de informação subjetivos, resultando em feedback ambíguo e numa classificação errada do conteúdo.

Neste documento, uma nova abordagem é apresentada com a intenção de utilizar sinais biométricos retirados diretamente dos espectadores, no momento preciso de visualização de algum conteúdo de vídeo, e usá-los para medir a ”Arousal” do espectador durante toda a experiência de visualização, proporcionando um novo ponto de dados a ser usado para classificação de artefatos multimédia. A Realidade Virtual é também um fator decisivo, uma vez que garante maior imersão, menos erros de medição induzidos por distrações e um vislumbre ao futuro da indústria.

Palavras Chave

Filme; Série Televisiva; Arousal; Realidade Virtual; Classificação Biométrica; Reconhecimento de Emoções.

Contents

List of Figures

List of Tables

Acronyms

HRV	Heart Rate Variability
TWE	Traditional Watching Experience
VRWE	Virtual Reality Watching Experience
HMD	Head Mounted Display
VR	Virtual Reality
ECG	Electrocardiogram
ANS	Autonomic Nervous System
PNS	Peripheral Nervous System
SNS	Sympathetic Nervous System
PSNS	Parasympathetic Nervous System
MHR	Mean Heart Rate
MRR	Mean RR interval
ReLU	Rectified Linear Unit
EI	Emotional Intensity
MAE	Mean Absolute Error
MSE	Mean Squared Error
MAEP	Mean Absolute Error Percentage
IST	Instituto Superior Tecnico

1

Introduction

Contents

1.1 Overview

With the growth of Movie/TV Show streaming, audiences have become the primary object of study when trying to evaluate the effectiveness or quality of such multimedia content. As Marine Boulanger, Médiamétrie's Director of Cinema, pointed out [3]:

"The cinema is a place for emotions. Our research shows that among some 42 million cinemagoers in France in 2018, 3 out of 4 viewed the cinema primarily as a catalyst for emotions."

And when we turn to the industry professionals' side, here represented by the legendary Stanley Kubrick, we get a similar response:

"A film is - or should be - more like music than fiction. It should be a progression of moods and feelings. The theme, the meaning behind the emotion, all that comes later."

Although the movie industry has suffered from a boom of new technologies and ways to develop itself, such as new watching devices, ways to distribute content, or even real-time personalized storytelling methods, the impact of the provided content next to the audience is still a tricky factor to measure.

If we look at the contemporary methods of formally requesting feedback from the audience regarding some Movies or Tv Shows, the main ones are Focus groups, Critics' reviews, Online Reviews, Anonymous Questionnaires, and user data retrieved by the streaming platforms, such as watch-time or simple qualitative evaluations. These methods have one thing in common: they all return results that require a cognitive process by the feedback giver. Results such as these are deeply influenced by many factors that may make them useless or even invalid.

Precision is one of those factors that falls short of the current mediums of feedback retrieval and content classification in the Movie Industry. Still, others exist, such as Accuracy, Ambiguity, and even external factors that may have influenced the viewing experience and directly affected the results. The following figure **??** from a paper named "Cross-evaluation of metrics to estimate the significance of creative works" [1] is an example of the problems that can be found when comparing this kind of evaluation method. Even though some different means are presented, it is possible to conclude that these problems are often related to personal mental factors. For example, "Proxy for popularity" and "Rater biases" directly result from conscious procedures. We can then conclude that **there is a lack of unconscious/objective metrics able to be used in this context to more accurately classify/evaluate such multimedia contents.**

By increasing the accuracy of the contemporary feedback retrieval methods, the industry would not only provide a better experience to the viewer, but it would also contribute to better predicting the audience response to an artifact and, consequently, lead to better executive decisions, fine-tuned creative paths and well-supported marketing strategies.

Class	Method	Property	Strengths	Weaknesses
Expert opinions	Preservation board (e.g., NFR)	Significance	Consistent selection process	Binary value
			Careful deliberation	Long time delay
	Critic reviews (e.g., Roger Ebert)	Quality	Subjective	Poor data availability
			Many independent samples	Limited value range
	Awards (e.g., Oscars)	Quality	Distinctive	Affected by promotion
			Information for older items	Restricted to small subset of films
Wisdom of the crowd	Average rating (e.g., IMDb user rating)	Quality/impact	Quantitative	Rater biases
				Unknown averaging procedure
	Total vote count (e.g., IMDb user votes)	Impact	Simple	Proxy for popularity
			Quantitative	
Automated/objective measures	Economic measures (e.g., box office gross)	Impact	Quantitative	Proxy for popularity
				Data availability
	Electronic measures (e.g., Wikipedia edits)	Impact	Quantitative	Proxy for popularity
				Complex interpretation
	Citation measures (e.g., PageRank)	Influence	Quantitative	Complex interpretation

Figure 1.1: Approaches for estimating the significance of films [1]

1.2 Motivation

The main objective of this approach in movie/TV show classification/reaction data gathering is to **create a new classifier** based on **returned unconscious data**, impermeable to personal psychological interpretations, social pressure, and external interference. This project was initiated by developing a hardware-based solution to capture a set of Heart Rate Variability (HRV) metrics, followed by a gathering testing phase on an experimental setup comprised of a Traditional Watching Experience (TWE) and a Virtual Reality Watching Experience (VRWE). The relationship between biometrics and arousal metrics, such as Attention, has already been tested and approved [4, 5]. Using this newly generated data, a sequential regression model was developed to classify a specific movie clip from one to seven based on the participants' reaction to it, based on a database that contains an arousal value previously determined for the such clip. Finally, a small usability test was conducted to understand the validity and necessity of a mock-up streaming application that uses a set of arousal-related classifiers.

To better justify the comparison of results between the traditional methods of feedback gathering referenced before and the proposed one, a thorough theoretical analysis will be performed to fully understand the current status of the mentioned industry, currently developed/used methods, and their performance and validity.

1.3 Hypothesis

This document proposes a new approach to the arousal biometric assessment area by incorporating metric-gathering methodologies in a Virtual Reality (VR) environment. Through this, we will be capable of increasing emotional responses with higher values of immersion/presence and, consequently, reducing external interference and distractions.

The proposed approach is expected to deal with the following hypotheses:

- **H1)** This new classifier will reach an average of, at least, 60% regarding it being useful or not to the targeted public;

- **H2)** The data acquired will return valid results in real-time and isolated from any conscious mental processes, resulting in an Mean Absolute Error Percentage (MAEP) of less than 10%;

- **H3)** This approach to movie/TV show classification will be perceived by, at least, 60% of the target audience as more scalable and trustworthy than a traditional and widely used feedback questionnaire.

To explore these hypotheses, two different test sessions were promoted, one for the data-gathering part of the process and another for the utility evaluation of other uses of such data.

2

Related Work

Contents

2.1 Background

2.1.1 The Movie Industry - State of Play

Before the COVID-19 years, the movie industry was in what is known as its historic period, between 2015 and 2020. During this period, the movie industry has reached about $234.9 billion in value. It has achieved a compound annual growth rate of 2.4%, presenting a predicted market value of $318.5 billion and $410.6 billion by 2025 and 2030, respectively [6]. Several technological and even demographic factors justify this.

The COVID-19 pandemic has represented a major hit in the traditional branch of the movie industry, with movie theaters having lost, in the US only, around $10 billion at the box office [7]. With such toll being caused by sanitary closings resulting in several months without operation, a part of the movie industry quickly took advantage of this ”new normal”: Streaming.

As reported by José Gabriel Navarro [2]: Streaming was the preferred medium for watching movies of around 14% of the public in 2018, while watching on a movie theater represented doubled that percentage. With COVID, streaming has surpassed theaters by around 2.5 times, as seen in figure 2.1.

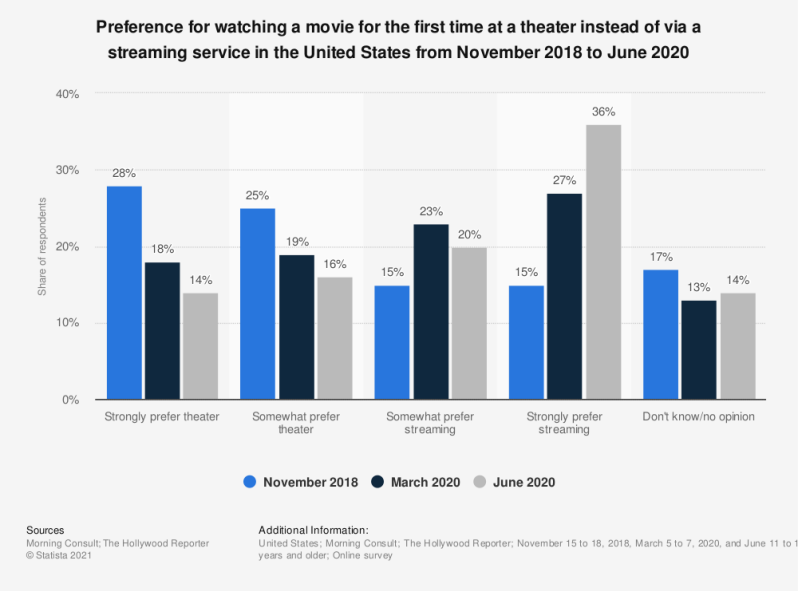


Figure 2.1: Preference for watching a movie for the first time at a theater instead of via a streaming service in the United States from November 2018 to June 2020 [2]

This significant influence jump may be considered provisional since the current health situation is also temporary. Still, experts say that this jump, although caused by COVID and the ”stay at home” directives given to the public across the whole planet, will be a permanent one because factors such as convenience, comfort, and cost have been taken with new care and importance and started being critical decision making prompts. Gene Del Vecchio [8] said that:

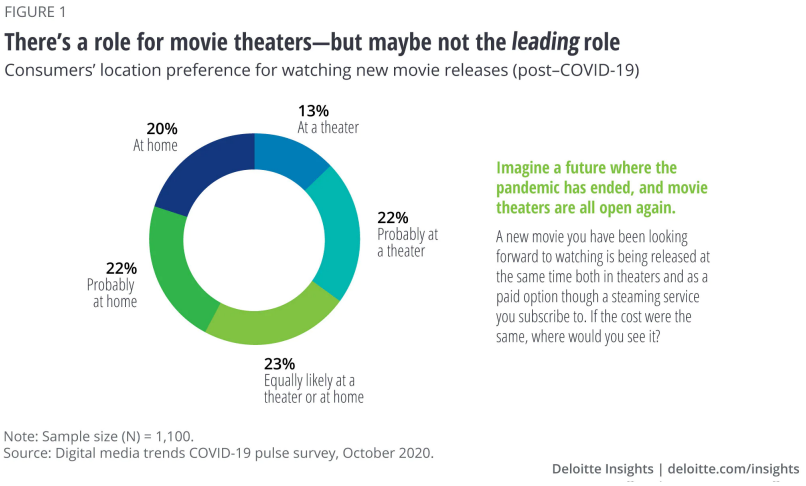


Figure 2.2: Consumers' location preference for watching new movie releases

"COVID-19 showed many consumers that it is far more convenient and cost-efficient to stream films and TV shows."

Looking at an article made for Deloitte Insights [9], the public's preference for watching new movie releases post-COVID registered seems to enforce this fact, as seen in the figure 2.2.

Experts have predicted that the video streaming market should reach $223.98 billion by 2028, with a CAGR of 21.0%. [10]

With the growth of streaming, movies and TV shows are suddenly closer than ever to the public, providing a much more personalized experience by giving control to the viewer, not of the content itself, but of How and When it is watched. The How part is exciting, considering that this new proximity to the viewers seems to have brought a lot of new content-watching methods to the table.

2.1.2 Model of Emotions

Since the proposed solution dramatically focuses on the impact generated by a multimedia artifact on a viewer, it's essential to evaluate how such an effect may be classified. Psychology and Effective Science have been trying to find a descriptive yet visual way to define human emotion. Although there seem to be a lot of different proposals, there are two main areas in which all of them seem to be included:

- **Discrete Emotions Theory -** Such theories say that emotions can be defined as simple, direct, and discrete concepts, fundamentally different from each other. One of the more widely distributed approaches is from Tomkins [11, 12], which presents nine different emotions: interest, enjoyment, surprise, distress, fear, anger, shame, dissmell, and disgust.

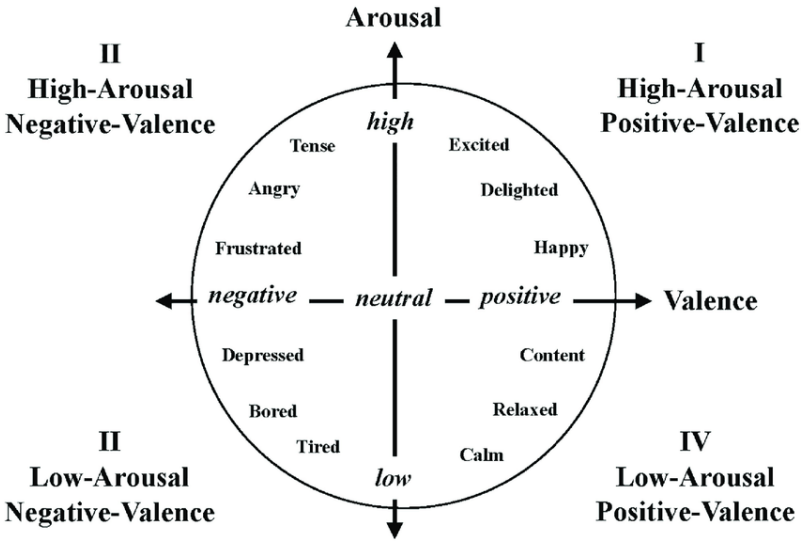


Figure 2.3: Emotions in the valence and arousal dimensions

- **Dimensional Theory of Emotion -** This kind of theory refuses the objectivity of discrete emotion classification. Instead, they propose that a dimensional model should be used to better account for slight variations and interdependence between emotional states. Wilhelm, considered the father of modern psychology, is one of the followers of such type of classification solutions, bringing to the public the idea of three main dimensions to be considered when identifying human emotions: ”pleasurable and unpleasurable,” ”arousing and subduing” and ”strain and relaxation. ” [13].

Among all the already published Dimensional Theories of Emotion, one seems to be the most referenced and applied in the most diverse occasions: The Circumflex Model. This model was developed by Russel [14] and used two different axes: Arousal, which represents the vertical axis and is commonly associated with the intensity of an emotion, and Valence, which means the horizontal axis and is known for portraying the positive or negative measure of sentiment, as seen in the figure 2.3.

2.1.3 Virtual Reality and its relation to Movies/TV Shows and streaming

As referenced in the Conviva Q4 2020 report [15], it’s possible to see that TVs only take globally about 17% of the viewing time, and the other 83% is distributed throughout other technological mediums. This shows that a conclusion can be taken: Streaming opens doors for new and innovative technologies that can improve the viewing capabilities of some movies or TV shows and even dialogue with such content to provide a more immersive and accurate to the initial purposes’ watching experience.

The use of an Head Mounted Display (HMD) to enter an entirely virtual environment has grown. According to eMarketer [16], and as seen in the figure 2.4, only in the US, which is also the country with the higher percentage of the penetration rate of subscription video-on-demand services (48%) [17],

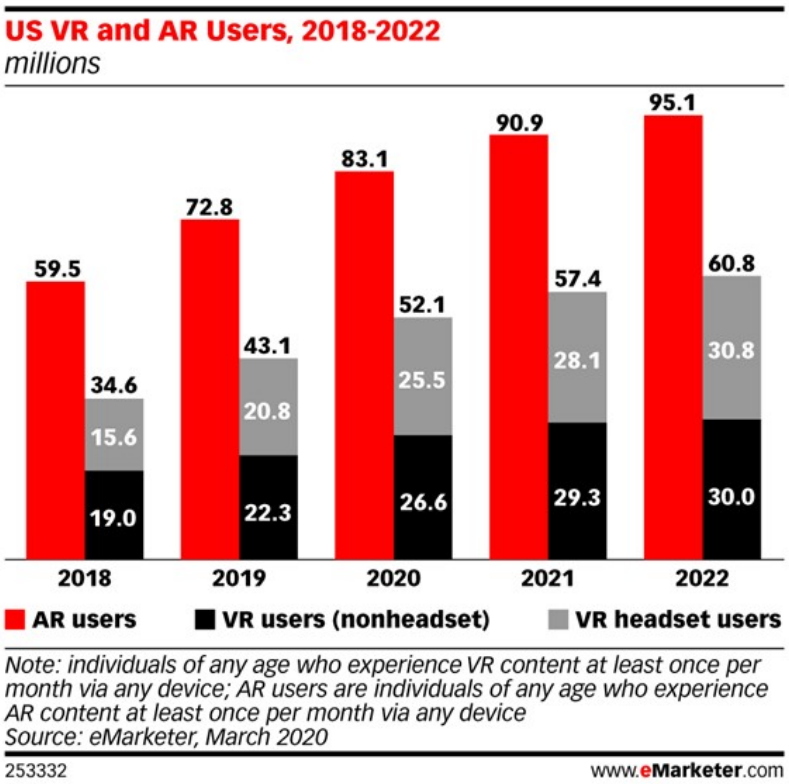


Figure 2.4: US VR and AR Users, 2018–2022 (millions)

there are 58.2 million VR users, which represents an increase of around 37% since 2019.

VR's capabilities represent a broad spectrum of new approaches to be taken and developed for several different areas, and the movie industry is no exception. Oculus, one of the most known VR-related companies owned by Facebook, has even created a dedicated department to cinematic storytelling through VR: Oculus Story Studio [18].

When we talk about the use of VR in the movie industry, we are automatically talking about the future, but maybe not a very distant one. Examples of the use of Virtual Reality to improve this industry and create new applications, consequently expanding it, are plenty and widely available with a single search online. Watching movies with an HMD and a specialized and personalized video player like SKYBOX [19], or simply entering the world of a 360º movie through YouTube are just some examples of already widely used connections between VR and Movies/TV Shows. Connections between this technology and streaming are also present.

2.1.4 VR as a movie/TV show generated emotions' amplifier

One of the main goals of a movie is to provoke a roller-coaster of emotional states next to the viewers. These might be joy, anger, fear, etc. Still, all of them have one thing in common: they are directly related to what the viewer is watching and hearing, and they can be distinguished into two main categories: the ones generated by the illusion of being in a fictional world (the death of a villain may cause joy, for example), and the ones directly related to the viewer's knowledge that such fictional world is being delivered to the audience by a concrete artifact (good acting by one of the actors/actresses may bring some enjoyment to the viewer, for example) [20].To better understand the relationship between these emotional responses and using an HMD to watch a movie or TV show, let's talk about the central concept in Virtual Reality: immersion. Slater and Wilbur [21] have pointed out that the correlation of four factors can define immersion: inclusion, extension, surround effect, and vividness. The first one is the degree to which outside influences are filtered out from experience, the second one is related to the number of sensory features that are affected, the third one measures how those sensory stimuli are distributed around the viewer, and the fourth one is related to the resolution of the display.

It would be unfair to talk about immersion without talking about presence; these two terms are often hard to distinguish, as they both build the base for a successful VR experience. As stated by Slater [22], Presence is also related to the illusion of being transported into the new virtual world. Still, in contrast with immersion, it is a perceptual concept, not a technological/technical one. Slater gives a pretty simple example of the difference between a perceptual and a cognitive response:

"where the perceptual system, for example, identifies a threat (the precipice) and the brain-body system automatically and rapidly reacts (this is the safe thing to do), while the cognitive system relatively slowly catches up and concludes 'But I know that this isn't real. But by then, it is too late; the reactions have already occurred." [22].

We can conclude that presence is directly related to the viewer's reaction, while immersion is a much more VR experience-related metric.

Suppose we try to find a relation between the emotions provoked by a movie and these two VR-related metrics. In that case, we can quickly conclude that a more considerable immersion can lead to a more significant presence. A more prominent presence directly leads to more robust and genuine emotions, at least those from the fictional world. With regards to the second type of emotions pointed out by Tans [20], we can also relate them to the degree of immersion and presence of a VR experience, just because our evaluation of an actor's performance, for example, can be influenced by how we felt at that moment when his character killed the villain. This proportional correlation between immersion and the emotional reactions performed by the viewers was already tested in a report by Vish et al. [23].

2.1.5 Movie/TV Show classification

We've already seen that a movie or TV Show is intended to generate a sequence of emotional states next to the viewer, that immersion and presence increase the power of such emotions, and that the ones caused by perception are the truest due to their unconscious source so, when we think about how the audience can classify such multimedia artifacts, we should find out that such classification would be made based on perceptual responses, but that's not what happens. All of them seem to be a result of conscious thinking.

As pointed out in the previous section of this document, the main mediums by which a viewer is asked to classify some movie or TV show are the following:

- **Focus Groups**: Focus Groups can be established in several steps of the production stage, but the consensus in the industry seems to be that the Post-Production test screenings are the most valuable and common. In regards to the results portrait by this method of feedback gathering, the possibly damaging factors start even before the watching experience: The person may be having a bad day, and this will make them have more tendency to lower arousal values/negative emotions since this screening is a previously scheduled event and do not care about the individually will that the viewers show at that exact moment; the viewer may be provided with few to no information about what they are watching, what can result in stress, anxiety and this may affect their comments after the movie; an NDA (non-disclosure agreement) can also be asked to the viewer, resulting in significant amounts of stress and legal pressure and force them to not answer truthfully to the questions made to them at the end; the whole formal and exclusive environment, both social and physical, can also influence the viewer's mood and, consequently, the feedback given by the viewer. Another significant factor is related to the choosing of a representative audience for a test screening. Companies tend to choose a diverse group of people to achieve better results, which is a good practice. Still, the number of participants will never be significant since a few hundred chosen viewers do not represent the billions that exist. Another big problem is related to the content shown because, most of the time, it is not the final version, which can seriously impact the feedback given. David F. Sandberg, director of movies like "Shazam!" (2019) and "Lights Out" (2016), has developed a simple animated video that points out a few problems with this kind of feedback-gathering method [24].

- **Critic's/Media Outlet's Reviews**: Websites like metacritic.com have been an accessible database of reviews for those who want to check what critics, aka professionals of the movie industry, think about some movie or TV show before deciding to watch it or not. Movie companies even host private screenings before the official movie release to use such professionals to create social media buzz and sell more tickets and streaming subscriptions. But these reviews can also lead

the audience to incorrectly classify a movie and spread such ideas without even watching it. This would be a little less bad if the critic's opinions did not diverge from the audience, but it does not seem to be the case. Film Data Researcher Stephen Follows pointed out in an article named "Are film critics losing sync with audiences?" [25] that critics' opinions have gradually diverged from the public's evaluations of the same artifact. This article also shows another concerning fact: The genres that appease audiences are very different from those that the critics prefer, and differences such as this seem to aggravate through the years. Apart from this, we keep having the audience representation issue, as critics are in far less quantity than the actual possible viewers of some multimedia artifacts. Regarding Media Outlet's reviews, not only do they still have the same problems stated before, but some questions about independence and impartiality also rise, since several movie companies also own media outlets and/or participate in the movie review industry.

- **Online Reviews**: If critics' reviews are already controversial in correctly classifying a movie, online reviews are the wild west of movie/TV Show Reviews: There are no rules. Viewers that, apart from enjoying or not some movie, provide a correctly argumented review have their thoughts mixed up with Fans that would do everything to exacerbate the movie's characteristics and haters that constantly try to put down some movie, TV Show, franchise, director, actor, etc. Take, for example, an article written by J.Bailey for Flavorwire [26], where the author also lifts the veil of critics-audience relation, which can prove very damaging for multimedia content classification. Although, there is some silver lining in this line of criticism: everyone can drop a review, so the audience representation, at least in number, is far more accurate than the one from other methods of feedback gathering.

- **Streaming Recommendation Systems**: Another more hidden classification method of multimedia artifacts is related to the streaming recommendation systems. As stated in the literature [27], there are two main types of Movie/TV Show recommendation systems: Content-Based, in which similarities between multimedia artifacts (director, actors, genre, etc.) classify them as suitable for a single viewer or not, and Collaborative Filtering, in which the habits and preferences of every viewer influences the choice of a single artifact for a single viewer. Both systems use simple objective parameters to classify content, such as watching time, genre, etc., but these parameters can also generate wrong results. As an example: A viewer starts to watch an episode of some TV Show, but during the experience, receives a call and has to turn off the TV. Suddenly, the algorithm stores this as an incomplete watching period, which leads to classifying that TV Show as probably not suited for them. This simple occasion will also contribute to not recommending that content to other viewers. With this example, we can understand that external factors resulting from low immersion values can influence the classification process in a streaming environment.

Other subjective metrics are used for ranking and classifying multimedia artifacts, such as the traditional ”like button,” which can also introduce questions like personal understanding, for example.

- **Viewer Questionnaires**: We end this discussion with the most used method for gathering viewer feedback. This questionnaire can vary in application, type of questionnaire, and even in being mandatory or not. Starting with the first factor, a questionnaire can be delivered in almost every possible situation: After a focus group test screening, as an add-on to an online review, or randomly after a user of some streaming platform watches an episode of a TV show. The two main problems with such questionnaires seem to be, once again, the number of answers, since you can't possibly have every viewer of some movie or TV show answer a questionnaire, and we also have a significant problem: Personal interpretation, and not just by the viewers, but by the author as well. When someone is writing a questionnaire, passing thoughts or objective metrics to words and structuring those words into accessible phrases can result in imprecise questions. The viewer can then misunderstand these questions, which results in incorrect responses. An example: A question such as ”What did you enjoy about the movie” can result in ”I enjoyed the main actor, the sets, and the costumes.” Still, it can also result in ”The death of the villain was great, and it perfectly ended the character arc of the main character,” both valid yet significantly different, what can lead to distinct classifications. Another problem can be quantitative measures asked to the viewers, for example, ”On a scale from 0 to 5, how much did you enjoy this episode.”. A simple scale of numbers is not enough to account for every possible understanding of the word ”enjoy.” Another psychological factor is the mood of the viewer when answering the questionnaire; if they had a bad day and didn't particularly enjoy the movie, they can be persuaded to answer most of the questions with less favorable results than they usually would.

As stated before, every one of these feedback-gathering methods seems to rely heavily on conscious data, which can bring problems of different natures, like impacting precision, accuracy, ambiguity, and isolation from external interference, all factors that need to be considered for a sound classification system.

2.2 Related Studies

To better understand the concept proposed and analyzed in this document, three studies have been chosen since they represent different approaches for taking biometric signals in VR and relating them to the user's arousal. The first one, by Hirt, Eckard, and Kunz, is titled ”Stress generation and non-intrusive measurement in virtual environments using eye tracking. ” [28].

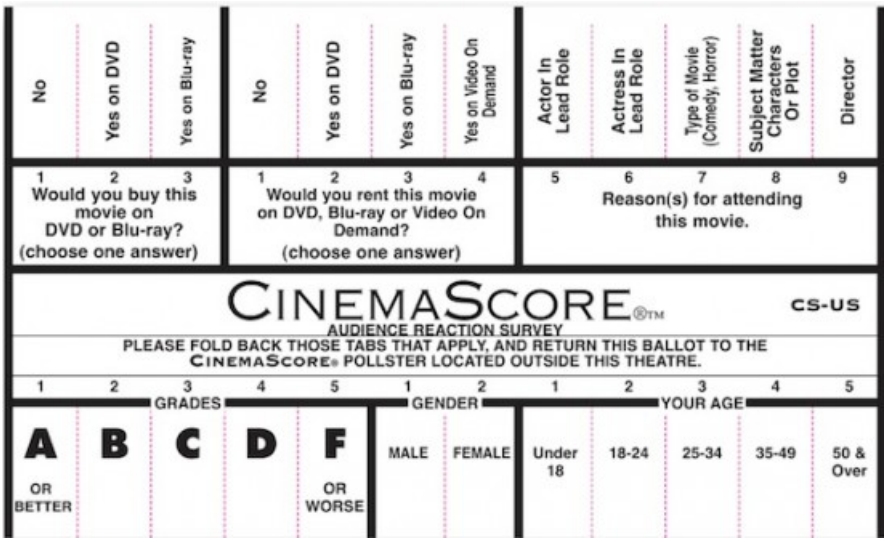


Figure 2.5: Example of an audience reaction form used for test screenings

The second one is ”Between-subject correlation of heart rate variability predicts movie preferences” [29] by So, Li, and Lau, and the third one is titled ”Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers” [30], from Schaefer, Nils, Sanchez, and Philippot.

2.2.1 ”Stress generation and non-intrusive measurement in virtual environments using eye tracking”

In this study, some essential and valuable conclusions were taken and evaluated. The authors’ main goal was to measure stress through a VR task and analyze biological reactions to some stressors induced by the application during the experience. When analyzing related work to their experiment, the authors referenced another report [31] that summarized stress measurement techniques as brain, heart and eye activities, skin conductivity, and cortisol level in saliva as being the most important ones when measuring pressure, and also pointed out that by having different error vulnerability, temporal needs and possible combinations with other metrics, makes them more or less suitable for VR experiences. After reviewing different setups used to take biometric signals and relating them to VR, the authors came to these conclusions:

- It is better to use combinations of biometric signals instead of single metric measurement;
- Heart Rate has been one of the most used metrics for this kind of arousal analysis;
- Regarding eye activity monitoring, it seems to be a non-intrusive biometric since many VR headsets already have this kind of sensor, guaranteeing that a high degree of immersion is maintained.

- Between Blink Duration (BD), Blinking Rate (BR), and Pupil Diameter (PD), the last one seems to be the most proved to be useful for stress measuring.

When discussing the results of this experiment presented by the authors, it is said that both an increase in average PD and an increase in Hear Rate do mean an increase in the stress level of the participant.

This report may prove helpful when trying to find accurate and non-intrusive biometric signals to be used at the time of the experience proposed by this document.

2.2.2 ”Between-subject correlation of heart rate variability predicts movie preferences.”

In this report, the primary conclusion taken by the authors is that Heart Rate Variability (HRV), taken in the moment of watching videos, does show synchronous characteristics between viewers that watched the same content. In contrast to the previous study, this one is not only related to the Movie Industry but also does not focus on a single arousal metric, such as stress; instead, it focuses on arousal as a wider one. This study uses different edits of the same content and then compares a cut with an expected higher arousal value to another one with a random distribution of clips. This comparison is then fortified with the results of a user survey to check if participants did feel more aroused by the first cut.

Although this report can sustain the desire to use HRV for this document's proposed solution, it is essential to note that no VR approach was used or even considered, which introduces other factors, such as immersion, able to influence the biometrics taken during the experiment.

2.2.3 ”Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers.”

This study acts as a verified set of information that will be used in the experimental phase of the proposed solution and proves the relation between watching a movie clip and rating it with an arousal-based metric. In this study, 364 participants watched 70 movie clips and, in the end, reported their perceived emotional status via a set of metrics such as Self-reported emotional arousal. If we take a look at the procedure of the developed experiment, we can take some important considerations when reviewing the initial indications presented to the participants:

- Participants were encouraged to report what they had felt and not what they believed people should feel in reaction to the movies;

- Participants were encouraged to report how they felt at the specific time they were watching the video excerpt and not their general mood of the day;

This data-gathering process is prone to interpretation issues, time influences, external distractions, and other error points external to the data-gathering process itself. These precious items clearly show that to more accurately assess the participants' emotional state while watching the video; the participants had to be advised against some ”bad practices” that could influence those results. The main reason that advice was necessary is that the emotion-related metrics evaluated, like arousal, are self-reported, which means that the participant must, willingly and consciously, report their data. The previously mentioned error sources are meant to be eliminated with the setup proposed in this document. After all, a large number of participants were used, which generated a big chunk of data with an error margin that ended up being considerably small, so, the results will be used as the fundamental basis for the machine learning algorithm. This happens because the data itself is not invalid, even though it was measured through imperfect means of data gathering.

3

System Design

Contents

3.1 Overview

The setup presented in this section intends to classify a particular movie clip with a continuous value from one to seven, related to the arousal felt during the process, based on the analysis of bio-signals extracted from a set of individuals while watching the such clip. To gather Electrocardiogram (ECG)-related data to be used when calculating some HRV-related metrics, twenty health participants were chosen to interact with the experimental setup. Each participant watched two sets of clips, one in a traditional way and another in a VR environment. This data was then used to train a Machine Learning algorithm, more precisely, a sequential model, utilizing the arousal values presented by an already developed study [30], since the used movie clips also came from the same survey. This model reported accuracy of 84,85% total, with 83,03% for Traditional-Only setup and 98,41% for VR-Only setup. In another phase of testing, and to prove that arousal-related classifiers would be useful for the users of movie/TV show streaming apps, a prototype of such an app was developed and exposed to a usability test with eight participants and not the same ones from the first experimental test.

3.2 Biosignal Recording

3.2.1 Lead I Sensor Placement

When constructing a full ECG spectrum of a specific individual, one of the main concerns is the amount and position of the measuring sensors. Before a decision was made about those two factors, theoretical research was done regarding the human heart operation. The component of the Peripheral Nervous System (PNS) that directly influences the function of the internal organs and, among them, the heart, which represents the primary data source of the proposed system, is the Autonomic Nervous System (ANS). The ANS is also proven to act unconsciously majorly [32]. The ANS is divided into two components: the Sympathetic Nervous System (SNS) and the Parasympathetic Nervous System (PSNS). These systems are opposites, the first related to immediate responses to a harmful event and the second to restore bodily functions to normal levels. The first is commonly associated with the ”fight or flight” system, and the second is with the ”rest and digest” system. When looking into heart rate behavior, the general conclusion is that the SNS is responsible for spiking it and the PSNS for slowing it down. Both the SNS and PSNS communicate directly with the sinoatrial node, also known as the heart’s natural pacemaker [33], which is responsible for electrically stimulating the heart and, by so, generating the heart rate. The timing of these impulses is the primary source of information for calculating each metric of HRV. The relation between emotion and an individual’s heart rate is also a proven factor [34], making this approach necessary for understanding the proposed setup. All HRV metrics can be calculated by directly analyzing an ECG taken from a specific individual in real-time. A correct lead must

be chosen and applied for the resulting ECG to be as accurate as possible. When trying to choose the best way to acquire this ECG data, one major factor was the need to minimize the intrusion level of the hardware setup. A lower intrusion level would lead to fewer worries and tension for the participant and, consequently, more accurate results. The minimum number of sensors needed to register an ECG is three, and only three leads can be used with such a small number of sensors: Lead-I, Lead-II, and Lead-III [35]. These three leads were proposed by Einthoven and form the famous Einthoven's triangle, which is represented in the figure3.1.

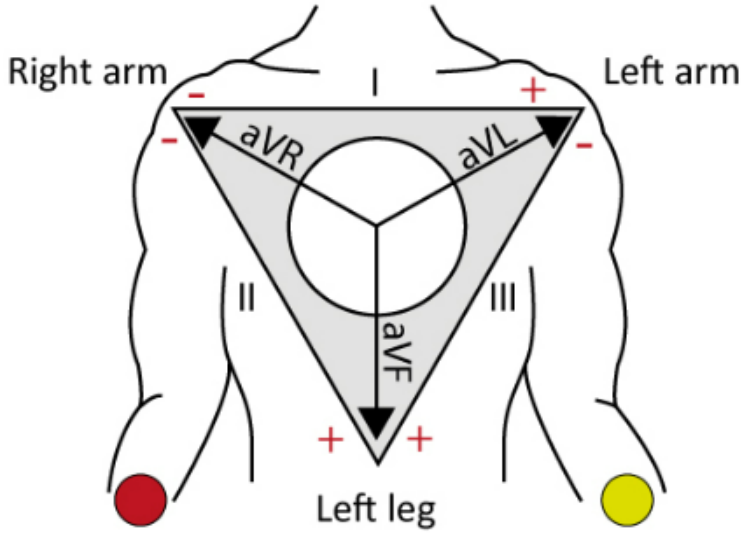


Figure 3.1: Einthoven's triangle Arms graph.

Still trying to minimize the participant's discomfort, the Lead-I was chosen since it fulfilled all the identified needs and required only negative and neutral sensors on the participant's left arm and a positive one on their right arm, as shown in the figure 3.2.

Figure 3.2: Lead-I sensor placement.

3.2.2 Plux setup

The experiment's first phase to test the concept presented in this document was the bio-signal recording process. To fulfill this phase, a hardware-based setup was developed to directly assess the behavior of the biometrics targeted by this approach. At first, it was attempted using a combination of a BITalino board and ECG plux sensors.

Important features like price and it being an open-source solution, when combined with the availability of such equipment at Instituto Superior Tecnico (IST) and the fact that, with Plux being a nationally originated company, technical support would be efficient and easy, made this equipment an obvious first choice.

Figure 3.3: Heartbeat - Plux hardware kit capable of measuring cardiac activity measurements (e.g., heart rate and heart rate variability) by evaluating electrocardiography (ECG) and Photoplethysmography (PPG) signals.

Paired with this hardware, Plux also provides software for recording and analyzing the biosignals acquired with the equipment they offer: OpenSignals. As also described on Plux's website [36]:

"OpenSignals is our easy-to-use and versatile software suite for real-time biosignals visualization, compatible with all PLUX devices.".

This software solution also provides a catalog of add-ons, each with a specific purpose of returning results for a particular metric based on the captured biosignals. HRV is one of the available metrics to be evaluated and the one with the most interest for the proposed system. Because of this description of OpenSignals, it was decided to use such a tool to analyze the data provided by the Plux/BITalino software bundle.

After a period of integration testing with both the hardware and software solutions, and at the end of the steep learning curve expected for this process (since there was no previous contact with such equipment), the results were not correct. The ECG lines seemed erratic and almost digital, not analog,

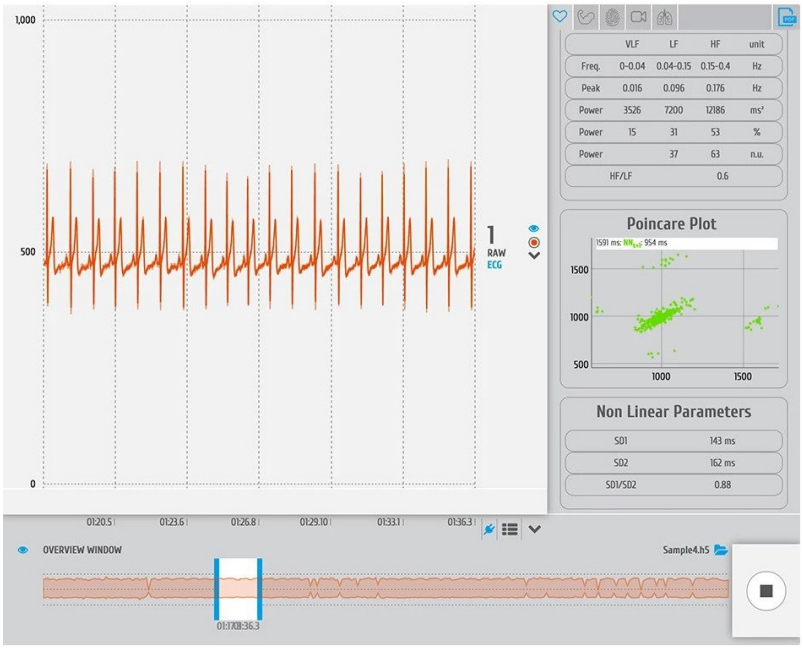
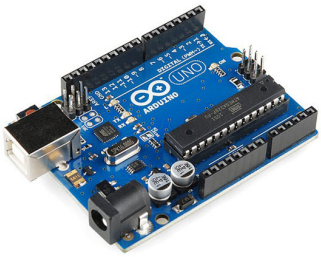


Figure 3.4: OpenSignals example of usage with the HRV plugin enable.

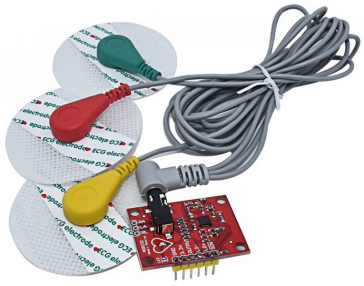
as expected. After the analysis of these results, the conclusion was that such equipment was prone to external radio interferences, directly influencing the extracted biosignals and rendering the results unusable. This conclusion assumes as the primary justification the need for the hardware to be close to a desktop connected to a power line, which created such interference problems. A second fact contributed to the final decision to abandon this system, the closed-down nature of the software and the impossibility of understanding exactly how every metric was being calculated. Even though the formulas used to calculate each HRV metric were present in a paper developed by the creator of this system. Since this data would then need to be used in another software environment (for the machine learning phase), it was decided to look for a better Open Source solution that could provide the necessary information and raw data to be processed apart from the hardware itself. This would also make the learning process more linear and contribute to a better understanding of the experimental setup.

3.2.3 Arduino setup

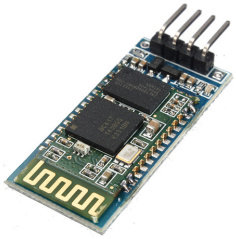
After rejecting a Plux-based hardware setup, the soon-to-be final solution was explored: an Arduino-based system. Being a widely distributed hardware, a plethora of information and documentation helped develop the final solution. The chosen board, for the sake of availability and set of functions, was the Arduino UNO. This board fulfilled all the needs of this project and was paired with two other pieces of hardware:



[Arduino UNO]



[AD822]



[HC-06]

Figure 3.5: Arduino UNO board hardware add-ons used.

- HC-06: This small module provided the capability of transmitting the acquired data wirelessly, through Bluetooth, to the laptop responsible for calculating the HRV metrics and respective data storing functions. This allowed for fewer physical connections between the subject of the experiment and the board itself, giving a higher degree of freedom to the participant and increasing the level of relaxation;

- AD8232: This ECG module was the most critical piece of hardware for this experiment. It provides three positive, negative, and neutral sensors, the exact sensors needed for a Lead-I ECG placement. This board also internally minimizes interferences, which helps get more accurate results. This device's operation was set to 1000Hz;

The combination of these three pieces of hardware resulted in a directly intractable system that wirelessly provided the ECG data needed for this experiment and combined the advantages of being cheap, well-documented, highly personalized, and widely available. The final setup is shown in the figure. 3.6

3.2.3.A Board's power problem

When this setup was initially tested, the results seemed either correct looking or strange, depending on the day. This recurring yet unpredictable problem made an early testing phase necessary to identify the problem and find the most indicated solution correctly. After changing each piece of hardware and

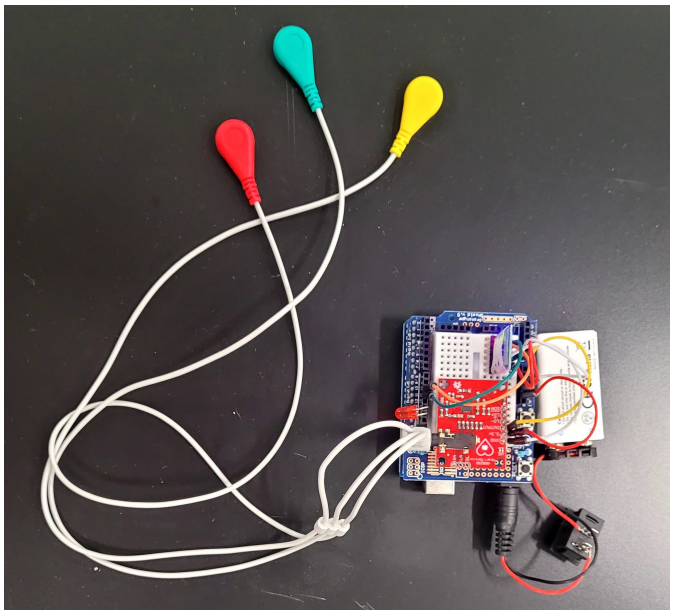


Figure 3.6: Final setup.

trying out different cables and software implementations, the problem was finally identified: The power source. Just like was previously happening with the Plux system, a direct line of power to the Arduino Uno, or even if it is powered from a computer connected to a mains line, generated a lot of unwanted noise that was not possible to be filtered out. To solve this problem, the Arduino board was powered via a USB-rechargeable 9V battery that, since it has a higher current supply than a standard 9V battery, was sufficient to power the whole setup for hours and without any mains-related interferences.

3.2.4 VR setup

To implement the VR part of this experiment, it was using the HTC Vive Pro Eye in an attempt to, apart from the ECG recording process, also record the Pupil Diameter of the participant while watching a movie clip. This extra data was discarded since it represented an inconsistency between the traditional watching experience and the VR-watching experience, rendering the comparison between both experiences incoherent since one had an extra data point.

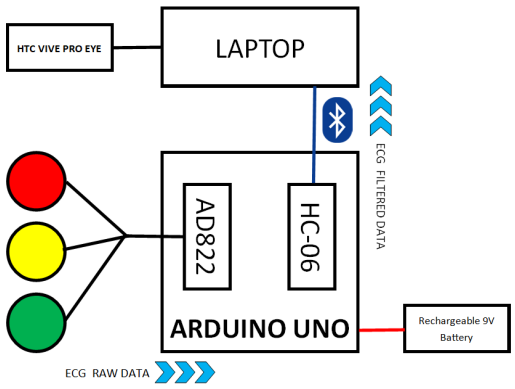


Figure 3.7: Diagram of the whole hardware setup.

3.3 Dataset

3.3.1 HRV metrics

When trying to understand which type of biosignal should be considered when the primary purpose was to measure and analyze the bodily reactions to emotional cues, the reached conclusion was looking into Hear Rate, more precisely, HRV. As indicated in [37], Heart rate variability (HRV) is the fluctuation in the time intervals between adjacent heartbeats. Another piece of information worthy of consideration is that HRV is directly connected to ANS processes, which makes the relation between emotion, ANS and HRV a proven link. HRV metrics can be divided into three sections: Time-domain, Frequency-domain, and Non-linear measures. Each of these measures was analyzed regarding their efficiency in a small capturing window and with relation to the hardware setup itself and its limitations. After such analysis, it was concluded that Frequency measures were influenced by the proximity to a Laptop or the battery that powered the Arduino board. This kind of interference would render such metrics unusable, so it was decided to discard Frequency-Domain measures. Regarding Non-Linear measures, and after conducting a theoretical investigation about their usage and limitations, the consensus [38] seems to be that these kinds of measures are not reliable for short capturing windows, which represents a significant setback for the proposed system since the content to be presented to each participant is only up to three minutes long. The last and not discarded type of HRV measure is the Time-domain. In total, there are about nine different Time-domain measures of HRV. All of these measures could be used for the purpose presented in this document, but due to the small capturing window available for such part of the process, considered even an ultra-short-term measuring window, some of those measures needed to be discarded in order only to get the most appropriate and literature supported results.

The table 3.1 provides information about each Time-domain measure and its smallest approved capturing window [37] [39].

Table 3.1: Time-domain HRV measures comparison. SDNN - Standard deviation of the NN intervals; SDRR - Standard Deviation of R-to-R; SDANN - Standard deviation average of the NN intervals; SDNNI - mean of the standard deviations of all the NN intervals for each 5 min segment of a 24-h HRV recording; NN50 - Number of pairs of successive NNs that differ by more than 50 ms; pNN50 - Proportion of NN50 divided by total number of NNs; RMSSD - Root mean square of successive differences between RR intervals

Measure	Unit	Smallest time window
SDNN	ms	60s
SDRR	ms	24h
SDANN	ms	24h
SDNNI	ms	24h
NN50	ms	60s
pNN50	%	60s
RMSSD	ms	60s
HR Max−HR Min	bpm	120s
HRV Triangular Index	ms	120s
Triangular Interpolation of the NN Interval Histogram	ms	90 s

After looking into these measures, the ones with the smallest capturing window (60 seconds) were chosen. Two others were added to this set: Mean Heart Rate (MHR) and Mean RR interval (MRR), with RR intervals being interbeat intervals.

3.3.2 FilmStim movie clips

An important factor that had to be considered when trying to construct a set of movie clips to be used in the data-gathering part of the process was that the duration of each clip had to allow for all of the chosen HRV metrics extraction. As pointed out in 2.2.3, the FilmStim database was used to establish a base relation between a specific movie clip and its corresponding arousal value. This database has 64 film clips, with each arousal level being a continuous value between 1 and 7. Reducing this amount of clips by selecting those with a 3-minute duration, at maximum, turned this set into a 19-clip pool. The main objective with this selection was to end up with three clips for the TWE and another three for the VRWE, each set of 3 clips with three different levels of arousal. The figure 3.8 demonstrates which movie clips were chosen and used.

	A	B	C	D	E	F	G
1	FILM SCENE	description	emotion	code	number	group	arousal
5	In the name of the father	Violent police interrogation leading to forged confessions.	Anger	310.00	30.000	3.000	3.84
6	There is something about Mary (1	Ben Stiller fights with a dog	Amusement	71.00	61.000	7.000	4.02
3	A fish called Wanda	One of the characters (John Cleese) is found naked by the owners of the house	Amusement	33.00	23.000	3.000	4.04
4	E.T.	E.T. is apparently dying	Sadness	19.00	9.000	1.000	4.16
6	Seven (3)	Policemen find the body of a man tied to a table.	Disgust	79.00	69.000	7.000	4.24
7	The silence of the lambs	Forensic examination of a dead body.	Disgust	32.00	22.000	3.000	4.39
1	Copycat	One of the characters gets caught by a murderer in a toilet	Fear	42.00	32.000	4.000	4.76
7	The Blair Witch Project	Final scene in which the characters are apparently killed.	Fear	65.00	55.000	6.000	4.95
8	The professional (2)	Stan (Gary Oldman) and his team kill Mathilda's (Nathalie Portman) family.	Anger	51.00	41.000	5.000	5.00
1	The Shining	The character played by Jack Nicholson pursues his wife with an axe.	Fear	38.00	28.000	3.000	5.11
2	City of angels	Maggie (Meg Ryan) dies in Seth's (Nicolas Cage) arms.	Sadness	46.00	36.000	4.000	5.15
3	Philadelphia	Andrew (Tom Hanks) and Joe (Denzel Washington) listen to an opera aria on the stereo. Ted describes to Jo	Sadness	72.00	62.000	7.000	5.24
9	Life is beautiful (3)	Mother and son are reunited	Tenderness	710.00	70.000	7.000	5.44
1	Life is beautiful (4)	In a concentration camp, a father "fakes" a translation of what an officer says in order to prevent his son to b	Tenderness	58.00	48.000	5.000	5.59
3	Seven (1)	By the end of the movie, Kevin Spacey tells Brad Pitt that he beheaded his pregnant wife.	Anger	75.00	65.000	7.000	5.69
4	Saving Private Ryan	Graphic war scene: fighting on the beaches	Disgust	37.00	27.000	3.000	5.73
5	A perfect World	Butch (Kevin Costner) is gunned down, at the end of the movie.	Sadness	53.00	43.000	5.000	5.78
6	Dead Man Walking	The main character is put to death by lethal injection	Sadness	31.00	21.000	3.000	5.87
8	Misery	Annie (Kathy Bates) breaks Paul's legs (James Caan).	Fear	510.00	50.000	5.000	6.12
6							
7							
8							
9							

Figure 3.8: Chosen movie clips. Green - TWE; Blue - VRWE

3.4 Arousal Prediction Model

The process of using, as an input, the taken ECG data and converting it to the necessary HRV metrics had to be considered a machine learning process. The first step was to ensure that the necessary data preparation was made between every step of the process so that data would be correctly formatted to be used by the next piece of software. If we take a look at those different parts, we start with the Arduino code that was responsible for receiving the ECG data and sending it to Unity, the second part of this process, and the platform responsible for calculating the HRV metrics. It was also in this platform that a saving structure was coded through which the calculated metrics and raw data were stored in a .json format so that they could then be imported by the machine learning algorithm developed in plain Python.

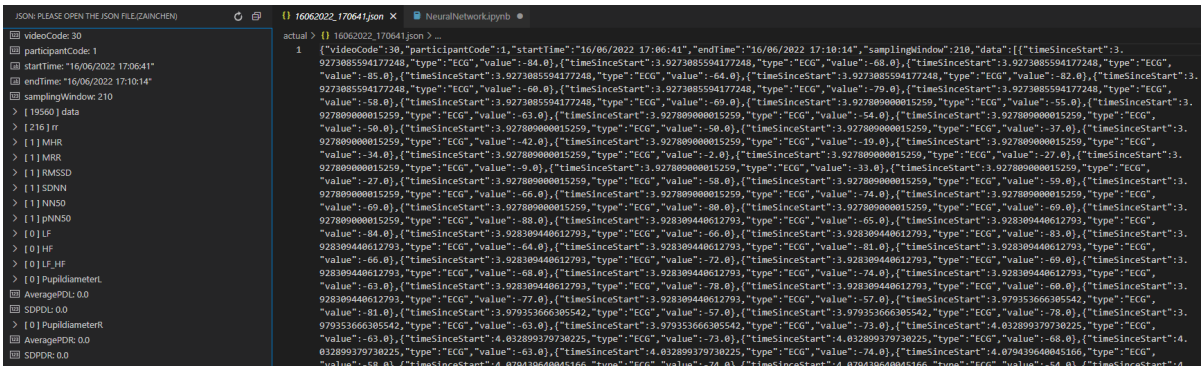


Figure 3.9: Example of a participant's recorded data file

.

It is important to note that each of these three platforms had its specific function. The Arduino environment was used due to its direct connection to the correspondent hardware, and none of the data processing needed was made through it since it would be considerably hard to debug many code lines. The only data normalization possible was applying a negative offset to the raw ECG data so that it would be centered around the same values as heart rate (around 60 to 100 beats per minute). Unity, on the other hand, had a much easier debugging platform and the flexibility provided by its coding structure also helped to ensure that the data were analyzed in real time by using the threading functionality available. This way, one thread could receive the data through Bluetooth while the data was being examined in another thread, and the watching experience was being run in the main one. After the data gathering phase, the resulting set of .json files was used by the sequential machine learning algorithm developed in Python. When trying to evaluate which type of machine learning to be used, the answer became apparent when realizing that this algorithm needed to be provided with an expected output depending on the movie clip that generated each set of data to assess if it was possible to reach the exact or approximate result with these new data points. Supervised Learning ended up representing the precise definition of this necessity. Another important consideration about the available data entries to

be analyzed is that a basic linear ”if-then” machine learning approach would not be the best option since the relation between each data point and the final result could be more complex than that. To handle this reflection, the chosen approach was to use a Neural Network, which means that the network would be organized as a set of neurons, each with inputs, outputs, and respective weight computation. The final network composition, either for TWE, VRWE, and a combination of both, is shown in the figure 3.10.

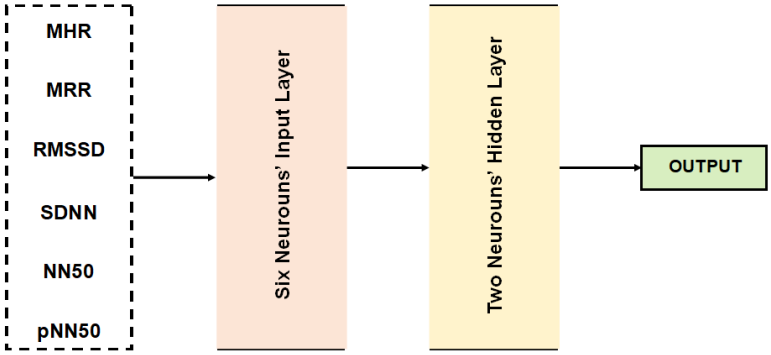


Figure 3.10: Final Neural Network

In regards to the activation functions for each layer, the choice was based on the requirement that the passed value between layers should always be continuous and positive, which made the Rectified Linear Unit (ReLU) the best alternative for all of the layers apart from the output one that, since it had only two previous neurons to consider, had as selected activation function the linear one. The data was divided into three different models, each of them with the same configuration as shown in the figure 3.10 , one that received all the data, a second one that received only the data that resulted from the TWE, and a third one that received only the data that resulted from the VRWE. Every model was tested with the same configurations, such as a validation split of 0,4, and 5000 epochs. Their compilation was also similar, with the loss function being mean squared error and an RMSprop optimizer of 0,001.

3.5 Steaming App Prototype

After the hardware-based setup was built with the primary goal of collecting biosignals from an individual while watching a movie clip and then developing a machine learning algorithm able to map those data points into a continuous arousal level, from one to seven, using the FilmStim database as the font of expected results, it was analyzed if it would be enough. The idea generated by this dissertation was directly connected to evaluating movies and Tv shows, which was achieved with this initial part of the process. Still, if we look into the main stakeholder of this project, the individuals that watch movies and Tv shows and use structured databases of that type of content, they were never an object of study. To minor this lack of connection to the real world, it was decided to elaborate a medium-fidelity prototype that incorporated some features that directly used HRV-predicted arousal to classify, filter, and order generic movies. In regards to the process of developing this prototype, it is essential to note that, even though it is understood that a complete interface designing process, with a set of previously tested prototypes, would be needed to classify this prototype as an entirely valid one, the main objective of it was not to test usability but the utility of the present features. Arguably, a shallow designing process could influence the users' reactions and the resulting analysis, but, in the sake of time and understanding that this decision was made in a late phase of the process, it was decided to keep this new data source, since its purpose was considered unique and necessary. For this prototype, the design was based on a set of widely used streaming platforms, like Stremio [40] and Popcorn Time [41], while also taking some creative liberty regarding the new features. It is also important to point out that the chosen platform for the designing process was FIGMA due to past experiences, and the arousal was converted into Emotional Intensity (EI), with the only intent of facilitating the understanding of the users. The new features introduced were as follows:

- **EI ID**: When looking into each movie poster, it is possible to note that its EI score is present, and its background can be green, for low EI, yellow, for medium EI, and red for high EI;
- **EI information panel**: With a click on the EI ID of a movie, a new panel of information appears with important extra information such as the value of variability of EI score during the movie, from one to ten, the lowest and highest EI score recorded, and the number of records available to determine this data;
- **EI variation graph**: With a click on the graph symbol inside the EI information panel, a chart appears with each EI score calculated in five minutes intervals;
- **Filter by EI score**: By using a slider, users can filter the presented movies by their EI score;
- **Filter by EI variability**: By using a slider, users can filter the suggested movies by their EI score variability;

- **Order by EI score**: Users can order the suggested movies by their EI score;

This prototype was then used for a user testing phase to understand if the presented arousal-based metrics would provoke any acceptance or refusal reaction next to the general audience.

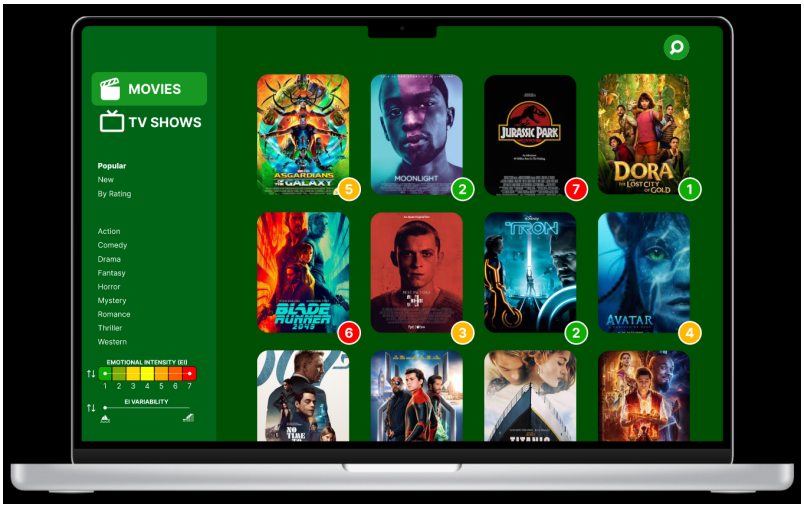


Figure 3.11: Streaming App Prototype

4

Data Gathering Phase

Contents

4.1 Description

4.1.1 Overview

A first user testing phase was conducted to acquire the necessary data to analyze and supply to the machine learning algorithm and determine if a bio-acquired arousal value is possible. The testing sessions were conducted at Espaço Cultural Fernando Augusto, in Póvoa de Santa Iria, with twenty participants.

4.1.2 Participants

With 20 participants, 35% male (7 individuals) and 65% female (13 individuals). Regarding age span, there was no age superior or inferior limit since the target audience of movies and Tv shows is quite large. The final set of participants ended up with the younger individual being 19 years old and the older one being 68. Participation was voluntary and did not require any official commitment.

4.1.3 Procedure

The participant distribution was made only attending to each participant's availability, and it was ensured that there was no participant meeting in-between sessions so that each one would experience every step of the process with a clean slate and no previous information about the experiment and, especially, the chosen movie clips. On arrival, each participant received a full briefing about the experiment's goal and how it would be conducted. Any questions were answered, and it was ensured that the participant was comfortable with advancing with the investigation. Some essential topics presented to each participant were that some movie clips could have mature content, that they would be connected to non-intrusive sensors, and that any data would always be treated anonymously. To ensure this characteristic, each participant was represented by a number and was never prompted to say their name for the record. After the ice-breaking and general introduction, each participant filled out a form with some demographic information and other necessary prompts, such as if they suffered from any heart or vision-related health problems, amount of hours weakly spent watching movies/Tv shows, and their previous degree of contact with VR. In what regards this information, it was concluded that none of the participants had any heart or vision health problems worth mentioning; most of them watched about 10 to 15 hours of movie/Tv show content weakly and also indicated a 2, from one to 5, when asked about their degree of the previous contact with VR. This form also incorporated data gathering and image recording consents. To eliminate the possibility of the order of experiences (VRWE and TWE) representing a data influence, each participant watched each set of three clips in the opposite order than the previous one. After each different clip, the participants were asked if they knew this clip before and their perceived degree of emotional intensity, even though this data was then discarded since it did not contribute to the

main goals of the experiment. After the end of the experiment, a SLATER-USOH-STEED QUESTIONNAIRE (SUS) [42] was applied. Still, since it generated a lot of confusion among the participants and most of them asked about its efficiency, this data was also discarded. With regards to the physical setup, the room lighting was lowered so that the focus of each participant could be directed to the screen of the laptop on which the experiment was run. Each participant was seated on a comfortable sofa to minimize the discomfort of being attached to sensors.

Figure 4.1: TWE setup

Figure 4.2: VR setup

4.2 Results

4.2.1 Overview

To understand the system globally and by separating the TWE and the VRWE, each of the three sequential models was analyzed apart from the others. It was also elaborated a t-test to the predicted values to understand if there is a difference between those two components. With data being recorded from twenty different users, the final pool of results ended up with 120 .json files, one for each of the six movie clips watched by each participant. The figure 4.3 shows the real values and the predictions that resulted from the validation set (10 of the 20 participants) for the general data, TWE and VRWE.

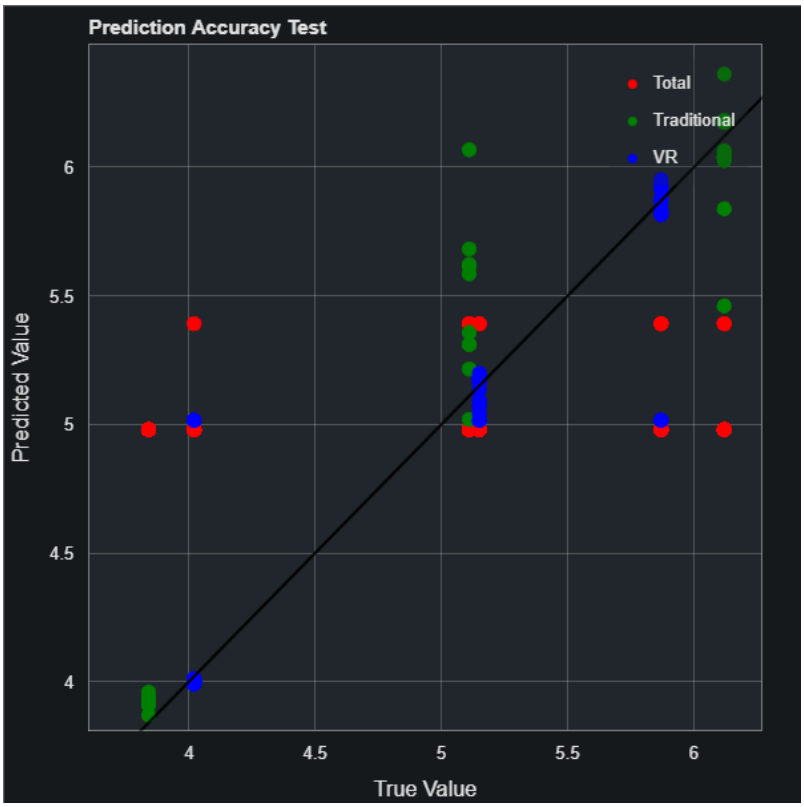


Figure 4.3: Arousal predictions that resulted from the validation set

4.2.2 General Data

Regarding the consideration of all 120 data files, the model reported a testing set Mean Absolute Error (MAE) of 0,73, an Mean Squared Error (MSE) of 0,71, and an MAEP of 15,45%. This data resulted only from 80% of the records used for the training process (validation split of 0,2). As for the other 20%, the validation set, the final results comprised an MAE of 0,74, an MSE of 0,73, and an MAEP of 15,53%. These results show a relatively low error, but it should be taken into account that, due to the capturing

window's size limitation, only three levels of arousal were targeted, which can influence the performance of this model. The figure 4.4 shows the MAE evolution during the testing phase of the model.

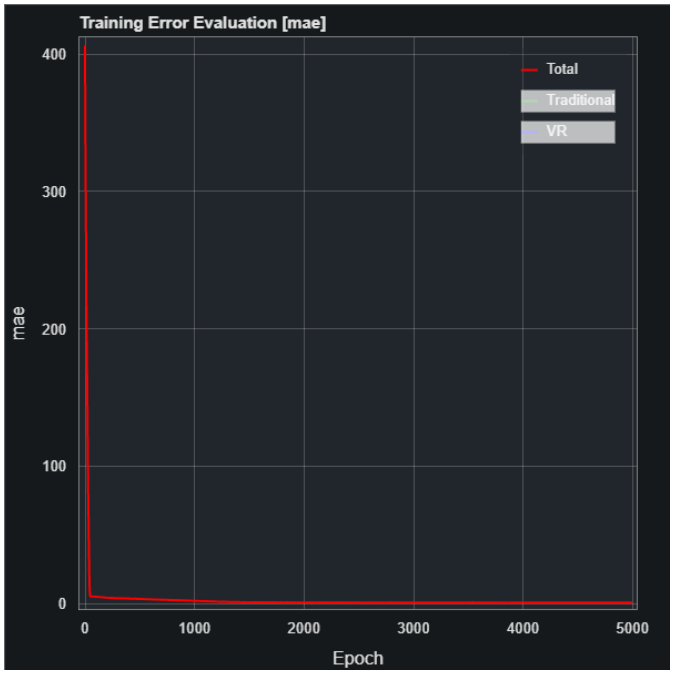


Figure 4.4: The training set MAE evolution

.

4.2.3 TWE vs. VRWE

When considering the TWE and VRWE models separately, each was supplied with 60 different records from the same 20 participants. Regarding the TWE, the correspondent model ended up with testing set MAE of 0,18, an MSE of 0,07, and an MAEP of 3,39%. As for the validation set, the final metrics were an MAE of 0,18, an MSE of 0,07, and an MAEP of 3,54%. The figure 4.5 shows the MAE evolution during the testing phase of the model.

Looking into the VRWE, the correspondent model reported a testing set MAE of 0,11, an MSE of 0,07, and an MAEP of 2,23%. Considering the validation set, the final results were an MAE of 0,05, an MSE of 0, and an MAEP of 0,90%. The figure 4.6 shows the MAE evolution during the testing phase of the model.

4.3 Level By Level

One of the hypotheses considered for this experiment tried to evaluate if there is a considerable difference between TWE and VRWE. This could be achieved by performing a paired sample t-test with the

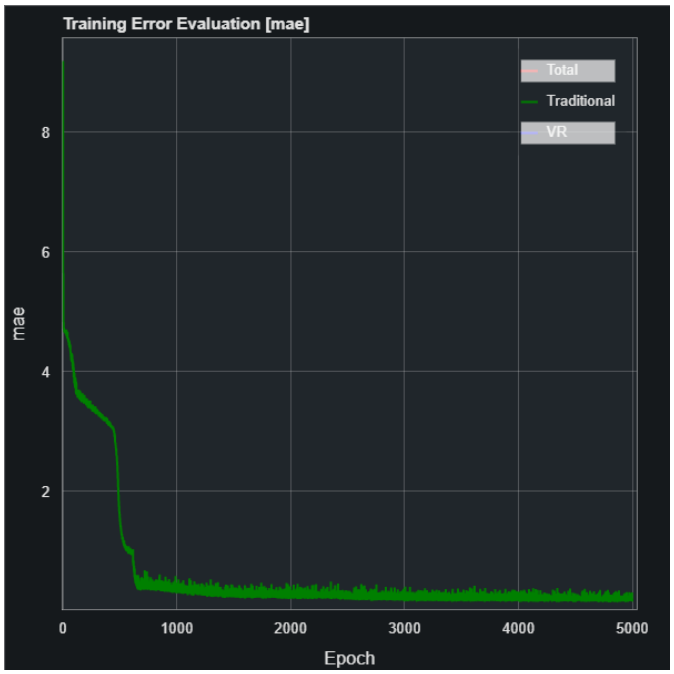


Figure 4.5: TWE training set MAE evolution

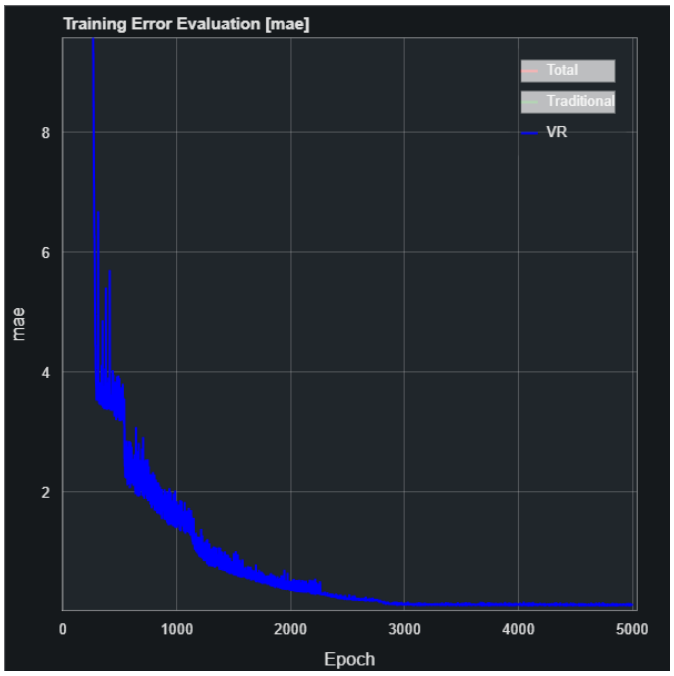


Figure 4.6: VRWE training set MAE evolution

120 records gathered. Still, since, due to the specific capturing window's size, only arousal levels from 3,840 to 6,120 were able to be predicted, it was decided to separate this t-test into three levels: Level 1 (arousal level of 4, rounded), Level 2 (arousal level of 5, rounded) and Level 3 (arousal level of 6, round).

	Level 1	Level 2	Level 3
Sample size	10	10	10
Difference Mean	-0,282287	0,369446	0,321086
t	-2,00215	3,60907	4,54473
Df	9	9	9
P-value (one-tail)	0,0381438	0,00283377	0,000698215
P-value (two-tail)	0,0762876	0,00566755	0,00139643
Lower 95.0%	-0,601233	0,137878	0,161264
Upper 95.0%	0,0366586	0,601014	0,480908

Table 4.1: T-test result metrics for all three levels.

This split was also supported by the fact that, as shown in the figure 3.8, both TWE and VRWE were comprised of a movie clip from each of these arousal levels. For these t-tests, the considered α was 0,05 and the Df 9, since the total amount of participants for each level was 10, the same ones used for the validation of each model. The table 4.1 shows the results of these tests.

As can be concluded from the data in the table 4.1, and after taking into account that to reject the null hypothesis (the difference between the paired population means is equal to 0), the P-value (two-tail) needs to be less than α (0,05), only level 2 and 3 respect this rule and prove that there is a significant difference between TWE data and VRWE data. Regarding level 1, the conclusion is the opposite: there seems not to be a considerable difference between TWE and VRWE. This contradiction may seem confusing and counter-productive, but if we look into the evolution of the P-value (two-tail) from level 1 to 3, it is possible to understand that it seems to be lowering with each group. This data may conclude that the stronger the expected arousal for a movie clip, the more accurate the VR related predictions are. This conclusion needs, though, to be taken with a grain of salt since only three targeted levels and nine participants can be analysed.

5

User Utility Test Phase

Contents

5.1 Description

5.1.1 Overview

This second stage of user testing was not directly related to the first one. This means that the data acquired in the first phase was not used for this one. The main objective of this part was only to evaluate if people would find the usage of supposedly biosignal-based classification metrics in a movie streaming app ambient to be justified and worth it. This experiment was conducted online via Zoom. It was easier to find participants and correspond to their availability concerns since the developed prototype could easily be interacted with via an online platform with a shared screen for observing purposes.

5.1.2 Participants

There was no prior participant selection process apart from availability and the necessity of neither one being also recruited for the first phase. This guarantee was established so that each user would not be influenced by the technical part of the proposed solution since it would never be presented to its final users on that form. Eight individuals participated in this study, six female and two male. Regarding the age span, since we are talking about a streaming app and the experiment was done via a web-conference app, the younger participant was 18 years old, and the oldest was 23 years old.

5.1.3 Procedure

On arrival, each participant received a briefing about the experiment but was not informed of the source of the biodata. They were only told that the presented streaming app introduced some new features based on heart rate-related data and that it is entirely symbolic and inaccurate. No form was presented to each individual regarding consent and conditions since their need was eliminated with a more direct approach. Each participant was, instead, asked about their will to participate; the support for image taking was not even presented since it was preferred not to record the sessions to provide more freedom and comfort to the user. The observer and conductor of the study recurred to taking notes about each user's participation. After the general ice-breaking and introduction section, each user was presented with the prototype to be tested. The test conductor followed a script that followed the guidelines of Google's UX Certificate [43], to provide an external theoretical background. Participants were asked for specific tasks and instructed to follow a Think Aloud perspective. At the end of the experiment, there was a direct communication window to understand users' concerns and suggestions better.

5.2 Results

5.2.1 EI ID

The script for this testing phase contained a simple identifying question that prompted the user to find the presented movie with the highest EI score. This question led to a percentage of correct answers of 87,5%, corresponding to one incorrect answer. Regarding the wrong answer, the user overlooked one of the available movies and its corresponding EI score, looking instinctively to the red background and not the number itself.

Every participant responded affirmatively when asked if this feature should be considered useful.

5.2.2 EI Information Panel

For this feature, it was asked for the user to identify any specific information that they could consider the less and the most important, with 62,5% chose the number of records as the essential info point and the EI score as the less important one.

When asked if this feature was useful, every participant answered affirmatively.

5.2.3 EI Variation Graph

Regarding this topic, for the utility-related question, 50% of the participants answered that they did not find this feature very useful and would not use it often.

5.2.4 EI Score Filter

When asked to filter the presented movies by EI score, even tho every participant eventually succeeded and found this feature useful, some did not get it right the first time.

5.2.5 Order by EI Score

It was also asked the participants to order the movies shown by EI score, which proved to result in a unanimous conclusion that this feature, although practical, was not directly accessible to every user.

5.2.6 EI Variability Filter

Lastly, for the EI Variability Filter, every user managed to fulfill the given task, but only 37,5% found this feature useful. The majority argued that it could represent too much information and a high degree of freedom that could lead to confusion.

6

Discussion

Contents

As represented in the chapter 1, the goal of the new approach for movie and tv show classification proposed in this document was to introduce a unique data point about this type of media directly related to the bioreaction of participants when watching it. This new type of classification is intended to prove itself useful next to streaming apps' users and viable, returning trustable and correct data. The hypotheses expected to be considered are:

- **H1)** This new classifier will reach an average of, at least, 60% regarding it being useful or not to the targeted public;
- **H2)** The data acquired will return valid results in real-time and isolated from any conscious mental processes, resulting in an Mean Absolute Error Percentage (MAEP) of less than 10%;
- **H3)** This approach to movie/TV show classification will be perceived by, at least, 60% of the target audience as more scalable and trustworthy than a traditional and widely used feedback questionnaire;

These hypotheses will be approached individually to analyze their possible rejection or acceptance, considering all previous parts of this document.

6.1 Hypotheses' Analysis

6.1.1 "This new classifier will reach an average of, at least, 60% regarding it being useful or not to the targeted public."

For this hypothesis, we ought to look to the second testing phase, as this was the one direct opportunity to evaluate if a biosignal-based classification of movies and TV shows would effectively prove itself useful to adequate stakeholders.

In a simple quantitative look, from the six different features presented to each participant, we reached an average of 72,9% when considering the one common question for all those, which aimed to understand if the participant found that specific feature useful. The only feature that did not reach the classification of "useful" was the EI Variability Filter. Several different reactions concluded that the main problem with this classifier is that not every person can quickly identify the exact mean of "Variability" when applied to this concept since it is a hardly instinctively measured characteristic. Another exciting perspective presented by a participant about this feature took into account the relation between user freedom and its consequent confusion degree:

I think that this feature, even tho it may be naturally confusing to a less informed population, respectively, an older one, is also not helped by the fact that it represents new paths for the participant and a lot of further information to assimilate, which can lead to the general confusion. - Participant 3

This consideration represents a vital data point regarding the specific utility of this feature, even though it may not significantly influence the general utility of biosignal-based info used by it.

Another vital source of information is the final discussion part of this testing phase. In this session, each participant was asked if they would, if asked, want to contribute with their data for this type of system regularly. Even though the majority (six out of eight) pointed out that, if ensured of their privacy and if this data were taken in a non-intrusive way, they would not have any problem sharing their measured bio reactions, the general reluctance was palpable.

I think I would participate, but I would not do it without thinking about it just because data leaks are much more frequent and extensive; who knows what could happen to my data. - Participant 1

To analyze possible solutions to this question, a look into the suggested characteristic of voluntary participation in the data-gathering process is needed.

Participation in supposed sessions to obtain the data for the proposed system would always have to be voluntary. This is a pretty simple conclusion since forcing people to provide personal data without knowledge about the inner processes of such a system would highly influence the resultant data and would not even be according to the ethical rules for this type of research [44]. On the other hand, a voluntary approach could lead to a small number of participants and, consequently, less data than necessary for a correct classification of a movie or TV show. But here is a half-measure that deserves a deeper look: a rewarded voluntary method. Suppose this system would be somewhat implemented by a streaming company, for example. This company could implement data-gathering groups where participants would sign up and be recruited to provide the necessary data instead of a user-by-user data-gathering process. These participants could then be rewarded to increase the general will to participate in future sessions, both for the current participants and for new ones. This approach would also allow for targeted artifacts, making content with fewer available records be directly provided with new data.

The final question presented to each user asked if they think that, if their current preferred streaming app would to implement a system like the presented one, they would use it on a regular basis. Every participant said that this new type of classification would, without a doubt, be helpful on a day-to-day basis, and some even made a not asked comparison with other already available classifiers:

I would look for the EI score more frequently than for an IMDB rating. - Participant 8

Attending to the analysis done in this subsection, we can conclude that a bioreaction-based system of movie/TV show classification can be helpful next to the audience, validating the hypotheses **H1**.

6.1.2 ”The data acquired will return valid results in real-time and isolated from any conscious mental processes, resulting in an Mean Absolute Error Percentage (MAEP) of less than 10%.”

For the second targeted hypothesis, it is necessary to look into the first testing phase and the machine learning algorithm analysis results.

Regarding the gathered data validity, it is possible to note that the model with the biggest MAEP is the one that merges data taken in the TWE and the VRWE, with a value of around 15%. This could lead to a general accuracy of about 85%. Still, it is essential to note that because this model uses data gathered in different setups/situations, it must not be taken with deep concern. This model served, majorly, as a ground zero, a basis for the whole experiment, making it a lot more interesting to look into the data from the TWE and the VRWE separately.

The TWE and the VRWE reported an MAEP percentage of 3,54% and 0,90%, respectively, regarding the validation set constructed immediately before training. Due to their low values, these percentages can support that the resultant predictions are valid and can be used for movie/TV show classification.

These results can also be considered real-time since the source data was taken directly from the user while watching the respective movie. They are also isolated from conscious mental processes since data was taken via biosensors and did not require any direct user data supply.

These three factors, being all respected and valid, prove that the hypotheses **H2** is verified.

Another critical reflection is the difference between the TWE and the VRWE and their respective data. As pointed out in the chapter 5, a paired sample t-test was developed to clarify if there is a considerable difference between data from these two types of watching experiences. The general conclusion is that this test proved this difference for the highest two of the three arousal levels able to be targeted and rejected for the lowest one. This fact, merged with the consideration that the t value seems to increase considerably when upping the expected arousal value, can hint at a possible conclusion that the bigger the experienced arousal, the more precise VR-related data, but this conclusion needed a more extensive set of data to be fully explored and confirmed.

6.1.3 "This approach to movie/TV show classification will be perceived by, at least, 60% of the target audience as more scalable and trustworthy than a traditional and widely used feedback questionnaire."

This final hypotheses can be somewhat trickier than the previous ones to be analyzed. Starting with the trustworthiness of this system, it is advised a look into the second experimental phase, more precisely, into the final discussion part. The last question asked to the participants about this system was developed specifically to address this need for validation. Each participant was asked if and how, after being told that this data would instead result from a set of questionnaires, this new information would influence their degree of confidence in the whole system. From a set of exciting considerations, the consensus was that the proposed approach should result in more reliable data, even though it may not be able to provide all the necessary info for a classification system:

I think those types of questionnaires could suffer significantly from misunderstandings and language barriers and be skipped. - Participant 5

Speaking by experience, I think that if I were presented with a questionnaire asking how I felt about the movie I had just watched, I would skip it or answer as quickly as possible with no regard for the importance of the information I was providing. - Participant 1

Wouldn't personal favorites or hated actors, directors, etc., end up influencing the results? - Participant 7

These considerations and the fact that every participant considered the proposed system a better alternative to questionnaires support the hypothesis **H3**. Still, it is essential to point out that the only mean of traditional feedback gathering asked about was questionnaires. Even though it is considered one of the most used, further comparison with other users' means of feedback retrieval could be made.

6.2 Future Work

6.2.1 Hardware Advancements

One of the main intentions of the developed setup for the data gathering setup was to minimize intrusion as much as possible. Intrusion can represent significant data interference and, apart from this central issue, can also influence the willingness of the participants to provide their data since comfort is the main decision point. With the growth of the streaming industry, which is predicted to keep its pace after the pandemic, new ways of consuming content will become gradually more frequent.

VR is one of the new ways of watching movies/TV shows and one with a similar predicted path of increased popularity. With the virtualization of content and the growing appeal of VR environments, watching movies/TV Shows is expected to become part of this new way of interacting with digital content. Another advantage, and one directly related to the approach proposed in this document, is the introduction of biosignal recording hardware in VR equipment. These sensors are initially designed with the single purpose of being seemingly integrated with the hardware needed for a VR experience. This makes them invisible to the users and almost reduces the voluntary participation question to a single checkbox or definition preference in a streaming app UI. One currently existing solution that could be directly applied to this situation is the HP Reverb G2 Omnicept Edition [45]. This headset already provides, apart from eye tracking, pupillometry, and face tracking, heart rate information measured directly through the provided headset. It even offers a development suite that automatically analyses HRV and allows this info to be extracted and used by developers when developing their apps. This way, all the data needed for this solution would be directly available with zero to no intrusion.

6.2.2 Other Industries

Although the proposed system was directly aimed at the movie/TV Show streaming industry when its concept was presented, more industries could profit from using a bio-acquired EI score classifier. A unique look will be taken into the Marketing and Music industry to explore this possibility further.

In this case, the music industry can be directly connected to the movie/TV show industry. An EI analysis could be helpful to not only when providing the general mood of the song to the users but also to creating new understandings next to the artists and producers about new releases and the intensity of the general song, or even the evolution of this score throughout the music. One could argue that emotions play an even more significant role when speaking about music and its purpose, so a biometric analysis of the audience could increase the proximity between the artist and the public.

Looking into the Marketing industry, this is one of the industries whose main target is the general public and how they react to visual content. As established before, the proposed EI score can be a direct path to user reactions without any conscious interpretation or interference. Using this type of classification/audience reaction meter, one of the possible new approaches would be to apply it to a user research session and directly compare the verbal feedback of each participant to their unconscious bodily reaction. This way, it would be possible to acquire more data and detect possible incoherence related errors with the verbal feedback given.

7

Conclusions

Contents

7.1 Overview

With this thesis, a new approach for movie/TV show classification, based on HRV, was presented. This solution started with theoretical research about possible links between emotional status, bodily reactions, and visual content provided. Once these links were justified, an Arduino-based system was developed to gather the ECG data necessary for the HRV analysis. A user study with 20 participants was conducted to gather the bio-data needed to be fed to a machine learning algorithm that took twelve different HRV features and supplied a constant arousal value directly related to the watched movie clip. This data ended up showing a low error percentage. Lastly, a second user study was conducted, a usability test directed to a streaming app prototype that implemented some HRV-related features, aiming to understand if a system like the proposed one could be useful to the audience. This system proved valid, returned unconscious and real-time results, and was also considered useful by a population of users. Future developments and expansion perspectives were also discussed.

7.2 Main Conclusions

After all the work developed during the process of creating this thesis, both experimental and theoretical, some main conclusions can be taken:

- It is possible to measure the emotional intensity, also known as arousal, experienced by a person when watching some movie/Tv show;
- HRV has a valid set of metrics able to be used for measuring arousal;
- The audience of streaming is opened for new classifiers for movies/TV shows based on bio signals;
- The target audience of streaming is trusts classifiers for movies/TV shows based on bio signals more than others that use feedback questionnaires as main data source;

7.3 Limitations

Along with important conclusions, and due to the highly prototypical nature of the developed work, there are important limitations associated with the approach presented in this thesis, such as:

- The use of a more Open Source solution with low to no isolation and proper filtering processes made the use of HRV metrics limited to the ones related to the Time Domain, what can influence the accuracy of the predictions made by the machine learning model;
- The use of clips of movies instead of the whole material could lead to incorrect classification of it since a three minute clip can not account for the whole movie;
- The streaming app prototype made did not comprise a full prototyping process with previous instances and respective tests, and this fact can support the idea that the way data was presented to the users could end up affecting the validity of the provided feedback;

Bibliography

[1] M. Wasserman, X. H. T. Zeng, and L. A. N. Amaral, "Cross-evaluation of metrics to estimate the significance of creative works," *Proceedings of the National Academy of Sciences*, vol. 112, no. 5, pp. 1281–1286, 2015. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1412198112

[2] J. G. Navarro, "Theaters vs. streaming: First time movie viewing preferences in the u.s. 2020," Aug 2021. [Online]. Available: https://www.statista.com/statistics/947757/theaters-streaming-watching-movies/

[3] L. O. Molinero, "Cinema: audience emotions under the spotlight," Dec 2019. [Online]. Available: https://www.mediametrie.fr/en/cinema-audience-emotions-under-spotlight

[4] U. E. R. d. Avila, I. C. Braga, F. R. d. F. Campos, and A. C. Nafital, "Attention detection system based on the variability of heart rate," *Journal of Sensor Technology*, vol. 9, no. 4, p. 54–70, Nov 2019. [Online]. Available: http://www.scirp.org/Journal/Paperabs.aspx?paperid=96583

[5] R. L. van den Brink, P. R. Murphy, and S. Nieuwenhuis, "Pupil diameter tracks lapses of attention," *PLOS ONE*, vol. 11, no. 10, pp. 1–16, 10 2016. [Online]. Available: https://doi.org/10.1371/journal.pone.0165274

[6] "Global film and video services market report 2021 - opportunities and strategies to 2030 - researchandmarkets.com," Sep 2021. [Online]. Available: https://www.businesswire.com/news/home/20210910005333/en/Global-Film-and-Video-Services-Market-Report-2021---Opportunities-and-Strategies-to-2030---ResearchAndMarkets.com

[7] "Movie market summary 1995 to 2022." [Online]. Available: https://www.the-numbers.com/market/

[8] T. Richards, "Predicting the future of the entertainment industry post-covid," Mar 2021. [Online]. Available: https://news.usc.edu/183870/future-of-entertainment-after-covid-movies-tv-streaming-usc-experts/

[9] C. Arkenberg, D. Cutbill, J. Loucks, and K. Westcott, "Digital media trends," Dec 2020. [Online]. Available: https://www2.deloitte.com/us/en/insights/industry/technology/future-of-the-movie-industry.html

[10] "Global video streaming market report 2021-2028 - researchandmarkets.com," Sep 2021. [Online]. Available: https://www.businesswire.com/news/home/20210923005761/en/Global-Video-Streaming-Market-Report-2021-2028---ResearchAndMarkets.com

[11] S. S. Tomkins, *Affect, imagery, consciousness*. Springer Pub. Co, 1962, vol. I.

[12] ——, *Affect, imagery, consciousness*. Springer Pub. Co, 1962, vol. II.

[13] W. M. Wundt, "Classics in the history of psychology." [Online]. Available: https://psychclassics.yorku.ca/Wundt/Outlines/

[14] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.

[15] "Conviva's State of Streaming Q4 2020," Tech. Rep., 1 21. [Online]. Available: https://www.conviva.com/state-of-streaming/convivas-state-of-streaming-q4-2020/

[16] "US VR and AR Users, 2018-2022 (millions)," 3 2020. [Online]. Available: https://www.insiderintelligence.com/chart/234406/us-vr-ar-users-2018-2022-millions

[17] Statista, "SVOD services reach worldwide 2022, by country," 9 2022. [Online]. Available: https://www.statista.com/statistics/813698/svod-reach-by-country/

[18] f. given i=J., given=Jeremy, "How Oculus Story Studio Learned Storytelling in Virtual Reality," 6 2021. [Online]. Available: https://spectrum.ieee.org/how-oculus-story-studio-learned-storytelling-in-virtual-reality#toggle-gdpr

[19] "SKYBOX VR Video Player on Oculus Quest." [Online]. Available: https://www.oculus.com/experiences/quest/2063931653705427/?locale=pt_PT

[20] f. given i=E.S., given=Ed, *Emotion and the Structure of Narrative Film: Film As An Emotion Machine (Routledge Communication Series)*, 1st ed. Routledge, 12 2011.

[21] M. Slater and S. Wilbur, "A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments," *Presence: Teleoperators and Virtual Environments*, vol. 6, no. 6, pp. 603–616, 12 1997. [Online]. Available: https://doi.org/10.1162/pres.1997.6.6.603

[22] M. Slater, "Immersion and the illusion of presence in virtual reality," *Br. J. Psychol.*, vol. 109, no. 3, pp. 431–433, Aug. 2018.

[23] V. T. Visch, E. S. Tan, and D. Molenaar, “The emotional and cognitive effect of immersion in film viewing,” *Cognition and Emotion*, vol. 24, no. 8, pp. 1439–1445, 2010. [Online]. Available: https://doi.org/10.1080/02699930903498186

[24] f. given i=S., given=Stephen, “Are film critics losing sync with audiences?” 7 2021. [Online]. Available: https://stephenfollows.com/are-film-critics-becoming-out-of-sync-with-audiences/

[25] “The Truth About Test Screenings,” 10 2019. [Online]. Available: https://www.youtube.com/watch?v=Fnvk9MNokPw

[26] f. given i=J., given=Jason, “DC Fans Don’t Understand What Criticism Is — And Film Media is Partly to Blame,” 12 2017. [Online]. Available: https://www.flavorwire.com/605957/dc-fans-dont-understand-what-criticism-is-and-film-media-is-partly-to-blame

[27] f. given i=R., given=Ramya, “How to Build a Movie Recommendation System - Towards Data Science,” 8 2022. [Online]. Available: https://towardsdatascience.com/how-to-build-a-movie-recommendation-system-67e321339109

[28] f. given i=M. and f. given i=A., “Stress generation and non-intrusive measurement in virtual environments using eye tracking,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 12, pp. 5977–5989, 3 2020. [Online]. Available: http://dx.doi.org/10.1007/s12652-020-01845-y

[29] f. given i=T. Y., given=Tsz Yan, f. given i=M. Y. E., given=Man Yi Erica, and f. given i=H., given=Hakwan, “Between-subject correlation of heart rate variability predicts movie preferences,” *PLOS ONE*, vol. 16, no. 2, p. e0247625, 2 2021. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0247625

[30] f. given i=A., given=Alexandre, f. given i=F., given=Frédéric, f. given i=X., given=Xavier, and f. given i=P., given=Pierre, “Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers,” *Cognition Emotion*, vol. 24, no. 7, pp. 1153–1172, 11 2010. [Online]. Available: http://dx.doi.org/10.1080/02699930903274322

[31] f. given i=S., given=Shalom, f. given i=H., given=Himanshu, and f.-H. given i=A., given=Allison, “A Survey of Affective Computing for Stress Detection: Evaluating technologies in stress detection for better health,” *IEEE Consumer Electronics Magazine*, vol. 5, no. 4, pp. 44–56, 10 2016. [Online]. Available: http://dx.doi.org/10.1109/mce.2016.2590178

[32] Wikipedia contributors, “Autonomic nervous system,” 8 2022. [Online]. Available: https://en.wikipedia.org/wiki/Autonomic_nervous_system

[33] ——, “Sinoatrial node,” 8 2022. [Online]. Available: https://en.wikipedia.org/wiki/Sinoatrial_node

[34] f. given i=B. M., given=Bradley M. and f. given i=L. J., given=Linda J., “Heart Rate Variability as an Index of Regulated Emotional Responding,” *Review of General Psychology*, vol. 10, no. 3, pp. 229–240, 9 2006. [Online]. Available: http://dx.doi.org/10.1037/1089-2680.10.3.229

[35] ECG Echo Waves, “The ECG leads: electrodes, limb leads, chest (precordial) leads, 12-Lead ECG (EKG) –,” 6 2021. [Online]. Available: https://ecgwaves.com/topic/ekg-ecg-leads-electrodes-systems-limb-chest-precordial/

[36] “OpenSignals (r)evolution (Download),” 12 2021. [Online]. Available: https://support.pluxbiosignals.com/knowledge-base/introducing-opensignals-revolution/

[37] f. given i=F., given=Fred and f. given i=J. P., given=J. P., “An Overview of Heart Rate Variability Metrics and Norms,” *Frontiers in Public Health*, vol. 5, 9 2017. [Online]. Available: http://dx.doi.org/10.3389/fpubh.2017.00258

[38] A. Bernardes, R. Couceiro, J. Medeiros, J. Henriques, C. Teixeira, M. Simões, J. Durães, R. Barbosa, H. Madeira, and P. Carvalho, “How reliable are ultra-short-term hrv measurements during cognitively demanding tasks?” *Sensors*, vol. 22, no. 17, p. 6528, Aug 2022. [Online]. Available: http://dx.doi.org/10.3390/s22176528

[39] F. Shaffer, S. Shearman, and Z. M. Meehan, “The Promise of Ultra-Short-Term (UST) Heart Rate Variability Measurements,” *Biofeedback*, vol. 44, no. 4, pp. 229–233, 12 2016. [Online]. Available: https://doi.org/10.5298/1081-5937-44.3.09

[40] “Watch videos, movies, TV series and TV channels instantly.” [Online]. Available: https://www.stremio.com/

[41] Wikipedia contributors, “Popcorn Time,” 8 2022. [Online]. Available: https://en.wikipedia.org/wiki/Popcorn_Time

[42] M. Usoh, E. Catena, S. Arman, and M. Slater, “Using presence questionnaires in reality,” *Presence: Teleoperators and Virtual Environments*, vol. 9, 04 2000.

[43] “Google UX Design.” [Online]. Available: https://www.coursera.org/professional-certificates/google-ux-design

[44] Office for Human Research Protections (OHRP), “The Belmont Report,” 9 2022. [Online]. Available: https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html

[45] Hp reverb g2 omnicept edition. [Online]. Available: https://www.hp.com/us-en/vr/reverb-g2-vr-headset-omnicept-edition.html