Affective State Recognition in Virtual Reality from Electromyography and Photoplethysmography using Head-mounted Wearable Sensors.

by

Ifigeneia Mavridou

A thesis submitted in partial fulfilment of the requirements of Bournemouth University for the degree of

Doctor of Engineering

Bournemouth University

Emili Balaguer-Ballester

Ellen Seiss

Charles Nduka

February, 2021

Copyright statement

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Abstract

The three core components of Affective Computing (AC) are emotion expression recognition, emotion processing, and emotional feedback. Affective states are typically characterized in a two-dimensional space consisting of arousal, i.e., the intensity of the emotion felt; and valence, i.e., the degree to which the current emotion is pleasant or unpleasant. These fundamental properties of emotion can not only be measured using subjective ratings from users, but also with the help of physiological and behavioural measures, which potentially provide an objective evaluation across users. Multiple combinations of measures are utilised in AC for a range of applications, including education, healthcare, marketing, and entertainment.

As the uses of immersive Virtual Reality (VR) technologies are growing, there is a rapidly increasing need for robust affect recognition in VR settings. However, the integration of affect detection methodologies with VR remains an unmet challenge due to constraints posed by the current VR technologies, such as Head Mounted Displays. This EngD project is designed to overcome some of the challenges by effectively integrating valence and arousal recognition methods in VR technologies and by testing their reliability in seated and room-scale full immersive VR conditions.

The aim of this EngD research project is to identify how affective states are elicited in VR and how they can be efficiently measured, without constraining the movement and decreasing the sense of presence in the virtual world. Through a three-years long collaboration with Emteq labs Ltd, a wearable technology company, we assisted in the development of a novel multimodal affect detection system, specifically tailored towards the requirements of VR. This thesis will describe the architecture of the system, the research studies that enabled this development, and the future challenges. The studies conducted, validated the reliability of our proposed system, including the VR stimuli design, data measures and processing pipeline. This work could inform future studies in the field of AC in VR and assist in the development of novel applications and healthcare interventions.

Table of Contents

Table of Figures

xiii

List of Tables

List of equations

Abbreviations

1-D, 2-D, 3-D	One-, two-or three-dimensional; referring to the spatial dimensions of an image.
A.C. or AC	Affective Computing.
ADs	Action Descriptors, *used in FACS.*
AR	Augmented Reality.
AUs	Facial Action Units, *used in FACS.*
AV	Arousal-Valence
CGI	Computer-generated imagery
CV	Computer Vision.
C.V.	Cross-Validation
DoF	Depth of Field.
DOF	Degrees of Freedom.
ECG	Electrocardiography.
EDA / GSR	Electrodermal activity, galvanic skin response or skin conductance *refers to a technique for measuring the skin resistance which is controlled by the sympathetic nervous system and is attributed to emotional and sympathetic responses.*
EEG	Electroencephalography, *is a technique that measures the electrical impulses of the brain using electrodes attached to the scalp.*
EMG	Electromyography; *a technique used for measuring the electrical activity produced by muscle movements.*
EOG	Electrooculography.
ESM	Evaluative Space Model.
FACS	Facial Action Coding System; *a tool that corresponds facial muscle movements to facial expressions and displayed emotion.*
FoV	Field of View

f-EMG	Facial Electromyography.
HCI	Human Computer Interaction.
HMD	Head Mounted Display; used in VR.
HR	Heart rate.
HRV	Heart-Rate Variability
IAPS	International Affective Picture System
IBI or RR-interval	Inter-beat Interval.
IMU	Inertial Measurement Unit; *usually includes a gyroscope, accelerometer and magnetometer.*
KNN	K-Nearest Neighbour
m-HRV	Momentary Heart-Rate Variability
MAX	Maximum Value
MIN	Minimum Value
MIP	Mood Induction Procedures.
MR	Mixed Reality.
NB	Naïve Bayes
P.C.	Physiological Computing.
PC	Personal Computer.
PPG	Photo-plethysmograph.
PRV	Pulse-Rate Variability
RMS	Root Mean Square Value
SAM	Self-Assessment Mannequins.
SD / Std.	Standard Deviation.
SVM	Support Vector Machines
VE	Virtual Reality Environment; *the immersive space created in VR. There are various types of VE and they vary depending on the project's goal.*
VR	Virtual Reality.
VRA	Virtual Reality Application.
VRET	Virtual Reality Exposure Therapy
UX	User Experience.
VAS	Visual Analogue Scale.

Acknowledgements

"Γηράσκω αεί διδασκόμενος" ("I'm getting older yet always being taught more (of things)" is a famous quote by Solon the Athenian, who laid the foundations for democracy. This phrase always inspired me in the way of living, of being always inquisitive and open minded, implying that getting older does not really make you wise. The decision of taking on a postdoctoral a research engineering (EngD) position, was stirred by this saying. Yet, I realised being a doctorate student brings you more doubt in your own knowledge and your work than ever before. Working on this Thesis for this past four years made me definitely older whilst also matured my thinking. Sometimes, I contemplated over my career decisions, but there are a few people to blame ehm…acknowledge for pushing and supporting this long endeavour, especially during the Covid-19 years. Without those people, I would not have been providing you today with this mesmerizingly long and heavy Thesis - my dearest apologies to the honoured reviewers and readers. Thus, I would like to firstly express my deepest appreciation for the reviewing committee for their time, patience, and guiding feedback.

Going back to this extraordinary, intense, and hard-working past four years, I have a very clear view of those individuals and the organisations who aided towards the facilitations of my research.

Starting from my academic supervisors, Dr. Emili Balaguer-Ballester and Dr. Ellen Seiss from Bournemouth University who both encouraged, guided, and enriched my cross-disciplinary research journey. Apart from being guided in data processing and machine learning approaches, I owe my exploration aptitude and positive attitude partly to Emili, as he continually and selflessly conveyed positive composure and a spirit of adventure in regard to research. Ellen continually and deftly guided, aided (and challenged) the experimental designs, the data analysis plans, as well as the overall critical thinking on research findings. I truly appreciated her stoic patience, caring nature, and the way that she clearly and generously bestowed knowledge from the psychology field. I would also like to thank Dr. Anna

Declaration

This thesis has been created by myself and has not been submitted in any previous application for any degree. The work in this thesis has been undertaken by myself except where otherwise stated.

Chapter 1

Introduction and Thesis Summary

Virtual Reality (VR) is no longer limited to laboratory settings. The range of its potential applications is rapidly increasing, due to recent technological advancements bringing low-cost portable VR headsets to the consumer market. From entertainment, training, healthcare and research, VR delivers a highly controllable system for the design of experimental studies, while also granting ecological validity [1]. The audio-visual content is under comprehensive creative control, which renders most aspects of the user's experience assessable and quantifiable. Combining VR with the ability to track and measure the user's behaviour within it, could potentially provide the ultimate laboratory for behavioural sciences and user-experience research. The ability to understand and measure emotional responses of an individual immersed in room-scale VR scenarios could be a great methodological improvement in VR research and it could facilitate applications in health-care and well-being [2].

As VR was an emerging technology at the time when this thesis started, a practical paradigm to capture different ranges of emotional responses in VR had not been yet standardised specifically for room-scale VR settings. An exploration of the available technologies for emotion detection was needed, and of devices and sensors that could potentially perform in synergy with the existing VR technologies. Potential candidates for such integration would be non-invasive, wearable sensors, that would not hinder or disrupt the user's movement and experience. Confirming the feasibility of using a combination of sensor modalities to capture affective states, could allow subsequently the development of affect detection algorithms via machine learning. Such emotionally-intelligent algorithms for VR applications could reveal novel routes for interactive experiences as proposed in Affective Computing [3], and contribute to a better perception of various psychological or mental states (such as Flow and Presence [4]). All in all, the quantification of the user's state could contribute to e.g. identify possible pathologies, and assist in the development of well-

being tools and health-care related solutions [5] which have been increasingly incorporating VR technologies lately to provide therapies that could be delivered remotely. Multimodal real-time data acquisition and user-centred analysis would be vital to the development of such interventions.

1.1. Affect Detection in VR: Prospects & Challenges

Affective Computing (AC) and Physiological Computing (PC; [6]) emerged, in order to provide computer scientists and researchers from several fields an interdisciplinary common space for collaboration on designing responsive computing systems to the psychophysiological activity of the user. The vision of Affective Computing (AC; [3]) reflects on enriching the computer's intelligence by entailing mechanisms for realistic behaviour, that could potentially exceed the "Turing Test" requirements [7]. Remarkably, a fundamental step in forming affective dialogues between machines and humans is recognising the psychological states expressed by the human interlocutor during Human Computer Interaction (HCI).

One potential way to get closer to this ability for HCI systems is to reliably detect emotional states in different settings. Numerous theories for emotional state detection and classification which have been suggested, will be presented in Chapter 2. They link affect to human cognitive and behavioural processes as well as psychophysiological states (see brief description in section 2.3.). In addition, the advancement of technologies and the development of scientific methods, has enabled researchers to monitor and evaluate emotional reactions to stimuli from a plethora of metrics and expressive modalities, including gestures, body movements, facial expressions, central and autonomic nervous systems responses, as well as speech ([8], [9]).

More specifically, affect recognition studies in non-VR laboratory settings are currently using various methods including self-report measures (questionnaires), behavioural observations, and physiological measures, such as heart rate (HR), skin conductance (galvanic skin response; GSR), facial movement tracking via computer vision (CV) and electromyography (EMG) amongst others. These methods are used to identify and measure the manifestations of emotional reactions. However, the existing methodologies had not been widely applied in fully immersive, interactive VR settings. The majority of affective studies in VR settings rely on subjective post-

exposure questionnaires, behavioural observations, and in some cases unimodal physiological measures that primarily focus on the identification of physiological arousal. By comparison, the measurement of valence levels from physiological measures was challenging [4]. This is mainly because VR imposes some practical constraints when it comes to the integration of conventional methods utilised for affect detection (see section 2.3).

Naturally, as humans utilise multiple cues and features to recognise emotion in every-day interactions, we expect 'intelligent' computers to employ a similar approach, rendering multimodal affect recognition approaches increasingly appealing. Advantages of adapting a multimodal approach lie on the pretext that the amalgamation of multiple signals or features can provide a richer source of data. This methodological approach can reduce the limitations of unimodal approaches since obtaining truthfully accurate signals can be a considerable limitation for certain recordings [9]. The utilisation of multiple signals can alleviate this restriction and offer a more well-rounded view of the overall expressive properties of affective changes [10]. Potential challenges however can range from (a) identifying physiological signals and their features which can carry information which are useful for affect detection, (b) selecting signals and features of high quality (or less noisy) compared to others, and (c) explorer and identify which combination of those can be most informative for affect classifications.

Ideally, multimodal affect detection in VR settings could also be achieved by combining methods in an especially adapted set-up framework, which would consider system usability, mobility, portability, and wearability factors. Thus, special focus is required towards the design of unobtrusive technologies for VR. Technological solutions including cumbersome set-ups with multiple cables, can impede one's freedom of movement and distract them from the VR content. This in turn can disrupt the overall user's experience with hampering results over the effects of immersion and presence [30]. These effects are suggested to provide a measurement of the efficiency of a VR simulation to induce naturalistic responses as the user believes in the reality presented by the simulation [11]. Wearable sensor technologies for VR are currently starting to emerge [3]. The latest hardware technological advancements result into an increase of computational power which in turn assists the development of novel interfaces and software solutions. These solutions could potentially open new avenues for real-time affective state analysis in VR.

1.2.   **Research Problem and Objectives**

In every-day social settings, we are able to understand what our friends or colleagues are feeling, by drawing information from the way they express their emotions from verbal and non-verbal cues such as their tone of voice, body motion, tension, distance from others and facial expressions. The majority of current emotion detection commercial products utilising CV focus on discrete facial expression and muscle configurations, following an approach by Ekman and Oster [12]. In fact, recognising facial expressions in comparison to other behavioural measures offers some advantages related to the ingrained biological component [13] and the debated universality parameter, as supported by cross-cultural studies [14], [15] and affective neuroscience studies [16]–[18]. However, utilising unimodal approaches as cameras to analyse facial expressions to infer emotions is in fact insufficient [19]. It provides a low-grained understanding of the inner feeling of the user, which depends on cultural variables and context-related information (for example smiling slightly may mean that an individual is embarrassed or stressed). Consequently, efficient emotional state detection would require a larger range of data types in conjunction with contextual information (e.g., knowing if the condition in which the emotional response was made was threatening or friendly) and additional psychophysiological measures of for example tension (arousal) which can enrich the facial recognition side.

In practice, the detection of affective states in VR would require the careful acquisition and interpretation of data which are normally highly subjective by nature [20], [21] together with the continuous monitoring of the context in which they were collected. This would allow us to understand the response to the interaction (the 'how') to the content (to 'what' caused this reaction). Event-stimuli related information could be taken from the simulations as stimuli creation and control is feasible through VR development. The endeavour of obtaining continuous, reliable multimodal data in Virtual Environments (VEs) could be potentially achieved by integrating multiple biometric sensors on the already existing prerequisites of wearable apparatus (like e.g., VR headsets) used in VR settings, instead of being constrained by the usage of multiple individual sensory systems which can be cumbersome to use. As this research field has great potential but is still at its infancy [4], our team investigated the efficacy of using such an headset-based integrated

multimodal sensory system in VR, and the feasibility of capturing physiological changes in fully immersive VR settings.

Consequently, the main objective of this EngD work was to assess the feasibility of using multimodal VR-integrated biometric approaches for affect detection, by combining objective measures of valence and arousal in stimuli-controlled environments. The informative value of these combined measures was compared against subjective self-ratings of valence and arousal recorded continuously during the VR experience. This work draws scientific expertise from multiple disciplines (including psychology, computer science (and VR), emotion research, biology, hardware engineering, signal, and data science). Thus, our team collaborated with wearable technology company called emteq labs ltd, with whom an emotion detection interface was further developed. This interface and earlier prototypes (called originally Faceteq and most recently changed to EmteqPro) was tested and used in our experimental studies.

In order to analyse the physiological responses and map them to affective states, we defined a detection system which combines data streams deriving from surface electromyographic (EMG) sensors, photo-plethysmographic (PPG) sensors and an inertial measurement unit (IMU) for movement detection embedded on the sensor insert. As part of this system, these co-registrations were mapped against participants' self-ratings (the user's reported affect) and were combined with content related data obtained from our custom-built stimuli-presentation applications, in order to gain insight on the emotional context of the user's experience. Our hypotheses focus firstly on the affective responses collected from the facial area, including heartrate, movement data and facial muscle activations, and secondarily on the accuracy of the mapping of those responses to the affective valence and arousal ratings.

In a summary, the following research (and development) aims were investigated in this thesis:

1.   Identify the affect recognition model and potential measures from which affective states can be inferred in VR.

2.   Design a system architecture including an experimental protocol, a sensor set-up and analysis protocols, to be used for the planned EngD studies. This system will be specifically designed to validate the feasibility of multimodal affect recognition measures and algorithms in immersive technologies.

3.   Conduct quantitative studies with human participants, to identify and evaluate the link between negative, positive and neutral emotional content and emotional responses, and their impact on the selected physiological readings.

4.   Explore the feasibility of measuring affect in immersive VR settings through the following sub-goals :

    a.   Create VR stimuli material and investigate the effectiveness of VR as an affect induction tool.

    b.   Investigate the performance of our affect detection system in immersive room-scale VR.

    c.   Assess the relationship between the levels of presence and the intensity of emotional responses in VR.

    d.   Investigate the feasibility of applying automatic affect detection in immersive VR by designing classification models for both emotional dimensions (arousal and valence) and assess the levels of accuracy for each dimension that could encourage further research in this area.

1.3. **Overview of Methodology**

As part of the work, we investigated the detection of voluntary expressions and spontaneous naturalistic affective changes to affective stimuli. The systematic affect analysis and detection system proposed in this thesis includes biometric and behavioural data acquisition, in conjunction with subjective ratings from users. For the measurement of those affective states' changes we utilised telemetry, heartrate measures, facial muscle readings, and subjective annotations, as dependent measures. The experimental paradigm used in our studies revolved around the concept of independent variable manipulation from virtual stimuli and videos, including audio-visual features and interactive events which were designed (and validated) to induce various levels of valence and arousal.

The data recorded were applied into the development of a multimodal affective state recognition system, whose architecture was based on data-processing levels in conjunction with the application of machine learning approaches for the effective classification of valence and arousal levels.

As Virtual Reality imposed some technical challenges in terms of sensor placement and signal noise caused by external factors including movement, we initially commenced our experimentation with controlled "seated" experiences to

minimise movement artefacts, In the first studies audio-visual video stimuli were used. In the last study we moved to fully immersive 3D Virtual stimuli. In the latter study, we tested the set up in seated passive VR conditions (passive setting) and in room-scale interactive VR conditions (active setting).

Through continuous experimentation and improvement of both the biometric sensors' set-up and the overall system architecture, we were able to isolate and focus on a specific set of sensory data for affect recognition in VR. Results from these observations are explained in the 'Methodology' chapter and detailed information about the specific studies conducted are available in chapters 4 to 6.

1.4.   **Justification of the Research**

Although there is a large body of research work conducted using physiological sensors for affect detection, little is known about their application in VR settings where discreetness and adaptability is highly desired. Additionally, the combination of different sensory modalities integrated within a head-mounted interface specifically adapted for affect recognition from the face in VR was not ever tested nor manufactured before, to the best of our knowledge, giving our team a wealth of interesting practical questions. In addition to developing and testing the hardware, out team envisioned the design of an experimental set-up and an event tagging software which could allow the automatic and direct coupling of affective responses to virtual content in VR and to continuous self-ratings made by the user. The main concept of such system is that it would enable researchers and developers to obtain meaningful information about the user's state within VR experiences. This characteristic could support, and perhaps even help revolutionise the way we conduct event-response research e.g., in psychology, or the way we interact with the virtual content.

This work concerns of a journey through efforts and endeavours of understanding and confirming that the proposed affect detection system architecture for free-walking VR is feasible. Most importantly, this research is aimed to assist and inform future implementations and further research on affect detection using VR technologies.

1.5. **Structure of Thesis**

In this section, the chapters of this thesis are outlined together with a brief overview of their contents; Starting from the literature review and the background research, followed by the methodological developments (including software and hardware solutions utilised and developed for this research), and leading to the detailed exhibition of the experimental and data-processing studies conducted, and the results acquired. The final chapter is dedicated to the conclusions upon the results, the potential implications as well as our suggestions for future work.

- **Chapter 1**: *Introduction and Thesis Summary.*

In this chapter we present an introduction to the research scope and an overview on the main research and development objectives of this EngD thesis, the methodological framework, and the structural skeleton of this thesis.

- **Chapter 2**: *Literature Review: Adapting Affective-Computing practices in virtual reality.*

In the second chapter, we present the related background research in terms of theoretical frameworks including emotion models, existing affect detection strategies from various physiological and behavioural signals and previous attempts to detect affect with and without the use of Virtual Reality technologies. In this chapter, the psychological phenomena described as presence and immersion are discussed in relation to emotion elicitation, and the parameters required to achieve sufficient levels for naturalistic interaction in VR. Current emerging technologies for affect detection in the market are also presented. The chapter closes with a discussion on the potential limitations and considerations for affect detection technologies in VR settings, and an overview of the proposed plan of research in order to achieve the set objectives explained in chapter 1.

- **Chapter 3:** *Methodology & System Architecture.*

The experimental approaches combining qualitative (subjective ratings) and multimodal quantitative data-acquisition techniques across the selected sensor modalities are presented. Subsequently, all methods, apparatus, software, signal processing and analysis approaches which were investigated, developed and used in the following studies are outlined. The system-architecture framework with which

we approached affect detection for VR settings is illustrated as a graph. The major components of this framework are described, containing the novel sensor set-up in VR, main methods, input data, intermediate processes/analyses, and output data. This model was followed in the studies discussed in chapters 4 - 6.

- **Chapter 4:** *Feasibility Studies 1 & 2 on valence and arousal detection.*

This chapter contains the feasibility studies conducted with conventional media using the VR-adapted sensor interface to validate and inform the system architecture proposed. The chapter contains three main sections. The first section involves the validation of a selection of affective videos which was used to elicit predefined ranges of valence and arousal in the studies described in this chapter. In the second section, we describe the study designed to explore the sensitivity of the electromyographic sensor set-up, to the participants' valence changes induced by the videos presented. The data analysis and classification results are presented at the end of the section. The third section comprises the exploration of the positioning of a photoplethysmographic (PPG) sensor on the existing sensor set-up interface, in order to obtain reliable heart-rate reading from the area of the face. A study was conducted to collect PPG readings from participants alongside readings from electrocardiographic (ECG) readings, which served as the 'ground truth' for comparison purposes. The same experimental protocol was following as in the previous study. The results from the analysis are presented and discussed. These two studies served as a validation of the existing methodology and informed the design of the following studies.

- **Chapter 5:** *Development and validation of stimulus material for an affective VR study.*

Following the feasibility studies using video stimuli, out team designed the next study to be conducted in fully immersive conditions using commercial VR headsets. Four virtual environments/scenarios were created, populated by 3-D custom designed objects and interactive events, which were intended to induce variations of arousal and valence levels. We describe the creation of the environments and the stimuli together with the developed plan of interaction for triggering the stimuli activation in a non-linear fashion, based on the movement and gaze of the user. A custom-made event-tagging system was developed to track the user's interaction with the stimuli, which was outputting event-markers alongside the signal data for

synchronisation purposes. This approach was explored in order to study the physiological changes recorded by the sensor set-up in relation to the contextual information (the virtual stimuli). An online survey study was conducted to validate the designed VR environments as an affect induction tool. The results from the survey are presented per environment, reinforced by the affective ratings reported per virtual stimulus, and memory accuracy scores. Additionally, presence scores are analysed per scenario and individual differences in terms of alexithymia are presented. The chapter closes with the discussion of the results which informed the main VR study presented in chapter 6.

- **Chapter 6:** Affect detection in Virtual Environments

The existing methodology, including the sensor set-up and virtual reality headset, were employed in highly immersive settings, using the VEs and stimuli described in the chapter 5. An untethered version of the sensor set-up was developed to increase the freedom of movement in virtual reality, and reduce the intrusiveness of the overall set-up. A large-scale study was conducted, where participants were asked to experience three VEs (a neutral, a positive and a negative) and self-rate their perceived affect in terms of valence and arousal. These ratings together with the physiological data recorded throughout each VE, allowed the validation of the virtual stimuli, and the exploration of feasibility to detect the changes in affective responses from the physiological data. The effects of interactivity and presence on affective responses were also explored. The participants were randomly divided into two groups, an 'active' one (interactive and free-walking) and a 'passive' one (vicarious, seated experience). A simple posed facial expression protocol was also used at the end of the experience to explore the sensitivity of our sensor-set-up to detect changes in muscle activation.

The hypotheses, the methods and the overall experiment procedure are described first, followed by the results. This section is divided into four subsections. The first subsection described the findings from the analysis on the self-assessments and ratings per VE and across events/stimuli. In the second subsection, the physiological changes for each VE and event/stimulus (event-based analysis) are explored. These changes were also compared between the two groups, active and passive. The second subsection is dedicated to the exploration of the sensitivity of the EMG sensors to detect changes in posed expressions. The data recorded during the posed expressions were analysed and an expression classification experiment

was also conducted using three classifiers. The last subsection contains the classification experiments conducted on the spontaneous physiological changes recorded during the affective VE experiences. Three classification approaches are presented: a user independent mixed-users, a user-independent separated-users, and a user-dependent approach. For each approach, results from three classifiers are compared for valence detection and arousal detection. For each classifier, three classification models were trained using firstly data from both groups combined, and then for each group separately. The chapter discussion and conclusions provide a holistic overview of the findings from the study.

- **Chapter 7:** Conclusions and Future work

This chapter contain a summary of findings from building the affect detection system to the classification experiments, followed by a discussion reflecting the contribution and impact of the research conducted in this thesis on enabling affect detection in free-walking VR settings. The chapter ends with a 'Limitations and Challenges for future work' sections with outlines the potential implications the future work on affect detection in VR. This is the last chapter which is followed by the references and the attached appendices.

1.6. **List of publications**

In this section the published papers and papers in preparation are presented. They are mapped to the listed aims (see list of aims in Section 1.2) of the EngD Thesis and isted in the order of the chapters. Publications 1-7 were extracted from the homonymous published paper which were reformatted to align with the formatting and text flow of the EngD dissertation.

N.	Publications	Research Aim(s)	Chapter (section)
1	Mavridou, I., Perry M., Seiss, E., Kostoulas T., Balaguer-Ballester, E. 2019. Emerging Affect Detection Methodologies in VR and future directions. In Proceedings of the Virtual Reality International Conference (VRIC 2019). p. 35, Laval, France, April 20, 2019.	1	2 (2.3.3.3)

2	Mavridou, I., McGhee, J.T., Hamedi, M., Fatoorechi, M., Cleal, A., Balaguer-Ballester, E., Seiss, E., Cox, G. and Nduka, C., 2017, March. FACETEQ interface demo for emotion expression in VR. In Virtual Reality (VR), 2017 IEEE (pp. 441-442). IEEE.	1, 2	3 (3.4)
3	Mavridou, I., Hamedi, M., Fatoorechi, M., Cleal, A., Balaguer-Ballester, E., Seiss, E., Cox, G. and Nduka, C, 2017. "Using Facial Gestures to Drive Narrative in VR", In Proceedings of 5th ACM Symposium on Spatial User Interaction (2017 October 16), Brighton, (pp.152), ACM.	2 (demonstration)	
4	Mavridou, I., Seiss, E., Kostoulas, T., Balaguer-Ballester, E., & Nduka, C. 2018. A System Architecture for Emotion Detection in Virtual Reality. In Proceedings of EuroVR 2018 conference, London, UK, 22-23 October 2018, 2 pages.	2	3 (3.2)
5	Mavridou, I., Seiss, E., Hamedi, M., Balaguer-Ballester, E. and Nduka, C., 2018. Towards valence detection from EMG for Virtual Reality applications. In: 12th International Conference on Disability, Virtual Reality and Associated Technologies (ICDVRAT 2018), 4 - 6 September 2018, Nottingham, UK. [Best Student Short Paper Award]	3 (Valence)	4 (4.3)
6	Mavridou, I., Seiss, E., Kostoulas, T., Balaguer-Ballester, E., & Nduka, C. 2018. Towards an effective arousal detection system for virtual reality. In Proceedings of the Workshop on Human-Habitat for Health (H3): Human-Habitat Multimodal Interaction for Promoting Health and Well-Being in the Internet of Things Era (H3 '18). ACM, New York, NY, USA, Article 4, 6 pages. DOI: https://doi.org/10.1145/3279963.3279969	3 (Arousal)	4 (4.4)
7	Mavridou, I., Seiss, E., Kostoulas, T. Hamedi, M., Balaguer-Ballester, E. and Nduka, C., 2019. Introducing the EmteqVR Interface for Affect Detection in Virtual Reality. In Proceedings of 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos	2 (new version of sensor set-up demonstrated)	3 (3.4)

	(ACIIW), 2019, p.83-84 IEEE, IEEE. DOI: https://doi.org/10.1109/ACIIW.2019.8925297		
8	Mavridou, I., Seiss, E., Kostoulas, T., Balaguer-Ballester, E. and Nduka, C. 2021 Building an affective VR study: A practical guide (In preparation)	1	
9	Development and validation of stimulus material for an affective VR study: An online survey study (in preparation)	4.a	5 (5.3)
10	Affect detection in seated and room-scale Virtual Reality settings. 2021 (in preparation)	4.b-d	6

1.7. **Chapter Conclusion**

This chapter presented the general scope of this EngD research, investigating novel ways to detect affective states of participants in Virtual Reality settings, while taking in consideration methodological and practical limitations. The prototype sensor set-up and the affect detection system architecture were developed after careful examination of the existing past literature on emotion detection and VR research, which will be presented in chapter 2. The studies conducted (see chapter 4 -6) enabled us to test and validate the methodological techniques employed. The results from those studies together with our empirical observations informed our approaches and could assist future work towards affect detection using VR headsets in fully immersive conditions.

Chapter 2

Literature Review: Adapting Affective-Computing practices in virtual reality

2.1. Introduction

Virtual Reality (VR) has expanded during recent years to a popular form of entertainment and a powerful tool for a variety of applications with a diverse range of experiences. As VR has the potential to emulate environments and situations similar to the real-world, it is being utilised amongst others in training [22]–[25], and treatment applications, such as exposure therapy (VRET) and cognitive behavioural therapy (CBT) [26], [27]. The rapidly growing international consumer market led to the sale of 13.4 million virtual reality headsets (aka head-mounted displays) in 2017 [28].

Despite recent advances, researchers and developers are facing a lack of effective assessment methods regarding the user's emotional state during a virtual experience [20]. Recognising the emotional state of the user could not only assist in the enhancement of human-computer interaction or the avatar-to-avatar interaction in VR, but could also be used as additional input, enabling interaction with content and adaptive control. Additionally, from a market-research point of view, understanding if the user felt pleasantly or unpleasantly within an experience could provide useful insights on the content's impact and likeability by the user. Consequently, such metrics could be very useful for a plethora of applications in extended realities[1] whose market size is expected to explode in the next few years [29].

---

[1]Extended realities (XR) is a blanket term that encompasses all computer-generated simulations which are either completely virtual (Virtual Reality or VR), mixed with elements of the reality (Mixed Reality

As a result of this realisation, the focus on the quantification of emotion has grown steadily in the last 10 years, leading to an estimated business market worth of $20 billion [30]. The importance of assessing and quantifying emotion is clearly highlighted by various research groups in this area (e.g. [31], [32]). Nevertheless, there is only a small number of emerging technologies that combine affective state detection with VR. The majority of research labs in the field today are limited by the number of affect sensing methodologies available that can apply in VR settings and the inherited constraints some of them impose on the user's movement [4].

The research aim of this thesis is to investigate ways to quantify the emotional experience from behavioural-physiological readings when engaged with immersive, computer-generated, affective content using modern consumer VR technologies. For this reason, we systematically reviewed previous research endeavours, definitions, models and controlling parameters. This helped us design the proposed affect assessment methodology and detection system for interactive immersive technologies.

2.2.   Mapping the progress: Affect detection

There is a vast amount of literature focusing on the nature of emotions, the processes involved, how it is experienced and expressed in humans and animals. Multiple studies proposed approaches for the efficient monitoring of physiological processes which were found to be closely correlated to bodily emotional responses. Such practices made it possible for the evolution of affective and physiological computing [3], [6] and assisted in the development of integrated wearable interventions.

From these practices, the methodologies which are found to be good detectors of the two dimensions of affect (arousal and valence) are presented, together with a review of related affect detection systems developed, not exclusive to immersive technologies but applied in similar fields. Additionally, an overview of current related emerging technologies will be presented for VR applications.

One of the more important factors of the immersive experience is the feeling of Presence, a psychological phenomenon linked to the successful impact or real-like effect of VR on users. The link between presence and emotion will be explored

---

or MR) or augmented on top of the real environment (Augmented Reality or AR). For the experience of such environments the user needs to wear head-mounted-displays interface.

and discussed in section 2.3.3). Aa new research methodology will be suggested and discussed in more detail in Chapter 3.

2.3. **From Defining to Detecting Affect in VR**

A primary step into designing a paradigm for emotion detection is to select a fitting emotion model. The vast literature in psychology and other fields, of over one century now, offers a number of emotion categorisation models (see section 2.3.1) and diverse perspectives on emotion research. Some of those models, such as the dimensional model of affect, are preferred for emotion recognition applications in the area of Human-Computer Interaction (HCI) [33], as affective states can be illustrated in a space spanned by arousal and valence dimensions. As this is a high interdisciplinary area, researchers and engineers need to collaborate on a cross-disciplinary basis and exchange of knowledge in order to understand the nature of emotional manifestations, and thus detect them as accurately as possible [8]. A level of abstraction is needed in order to tailor detection models towards easier to implement practices (e.g., deducting affective states from two dimensions). However, generalised simplifications may skew the reliability of developed systems across individuals. It is therefore fundamental for computer science researchers to follow and consider the advances and disadvantages as well as the active debates in the emotion research field across multiple disciplines, while focusing also on the technical challenges and the practical implementations.

It is generally agreed that emotions comprise three components, (a) the subjective experience (e.g., feeling happy), (b) the expressive behaviour (e.g., smiling), and (c) the physiological component (e.g., sympathetic arousal) [34]. Emotions, as described by Barbara Fredrickson [35], are the personal assessment (based on subjective experience and cognitive processing) of an event that activates a sequence of response tendencies manifestations across component systems such as bodily behaviours, facial expression, vocalization, as well as physiological changes [15], [36]. Affective computing methodologies are based on the principle that by detecting the changes manifested through those component systems we can potentially infer the underlying emotional state of a user.

Generally, emotional states can be referred using different terms, such as: a) emotions, as a short-timed, intense manifestation of a feeling addressed to a stimulating source, b) moods - which are actually longer lasting and do not need an

event to be triggered, c) affect (or core affect) as the underlying instinctual response to stimuli prior to cognitive evaluation, as well as c) feelings, a less well-defined term [37]. The term of affect and emotion are widely used in the area of AC. In section 2.3.1, we describe different categories of emotion models, upon which state detection systems can be built. These models offer insights on the mapping of expressive manifestations and contextual information to defined states or levels of affect. In particular, some of the most popular models are outlined, including the discrete or categorical model, the dimensional mode of affect, and the appraisal model of emotions. In those models the same emotional state can be associated to different characteristics relevant to each model; for example, 'feeling happy' can be linked to the smiling, to positive valence, and to approach. Such defining quantifiable characteristics along with their potential advantages and disadvantages in AC research will be described in the next section.

2.3.1. Emotion modelling frameworks: From theory to practice

Although the manifestation of emotions may vary between individuals are some common principles constituting their underlying processes (e.g., physiological responses) that allow for possible for interpretation through computational means. Certain interpretations became possible with the emergence and investigation of small dynamic behaviours coupled with exciting changes in physiology (showing a united biological origin and functional resemblance [38][39]). These physiological changes can be observed via technological and computational means, via the use sensors. As such, data deriving from various bodily functions and sensory outputs can be integrated in computational systems which follow an explicit model of translating the signals into emotional connotations (a detailed discussion of the conceptual models can be found in [40]). Therefore, the role of the model used for such integration is paramount and any modelling approaches that are explored are consequently extremely interesting and imperative for the efficient design of such systems. Naturally, the construction of theoretical and computational models requires some levels of abstraction in their specification. Four main emotion modelling approaches relevant to computational modelling [37] are described next: the discrete or categorical approach, the dimensional approach, the componential or appraisal-based approach and evaluative space approach.

2.3.1.1. **Discrete Emotion Approach**

Some of the most prominent theories are classifying emotional states in distinct categories of the so-called basic or fundamental emotions. There seem to be an agreement over 6-8 basic emotions that can be expressed and recognised across cultures, including anger, surprise, fear, joy, disgust and sadness, with some theorists adding a few additional ones, such as pride and contempt ([14], [41]–[46]). Recent research argues that the basic emotions that can be detected early-on from facial expressions are actually four (happy, sad, fear/surprise, and disgust/anger) before dynamically evolving into more complex ones [47]. A large number of autonomic emotional state detection systems rely almost solely on detecting facial expressions, based on Ekman's theories on facial expressions of emotions [48].

Regardless of its straight-forward nature, this model is not always preferred for multimodal AC research, due to its strict, deterministic structure, and the complexity to identify and distinguish some states from others via physiological data [49]. This approach is however preferred for camera-based affect detection applications, where the state categorisation is based mostly on the differences between various facial muscle configurations related to the basic emotions (e.g.[50]). The expressions and individual facial configurations are distinguished and rated based on the Facial Action Coding System, aka FACS [48], by trained observers.

Indeed, the human face can convey ubiquitous emotional information in everyday interactions but only a small volume of research has investigated the formation and context of spontaneous, naturally occurring expressions [51]. As explained by Calvo and Nummenmaa [52], the interpretation of spontaneous expressions is not as easy and it relies on knowledge of context (as opposed to posed expressions) as they do not encompass fixed signs of basic emotions. Recently, a systematic review on the reliability of distinctive facial expressional configurations (of basic emotions) published by Barret et al. [19] criticised the widespread assumptions on the existence of facial expressions which can be universally interpreted into discrete emotional states (perceiver-dependent), some of which are nowadays used for applications in legal judgements, policy decisions, security and training practices. Further research is required into understanding the defining processes and interpretation of spontaneous facial expressions together with

contextual information confirmed by the persons' subjective experience, requiring little or no observer inference.

Camera based approaches in computer vision (CV) for affect detection have gained many followers as they offer an easy-to set up, non-intrusive hardware and are open to a wide area of applications [51], not limited to identifying discrete facial expressions. 'Big', full-intensity expressions have been explored predominantly as they are generally easier to detect than subtle ones [52], [53] which are complex in their morphology and more common in our everyday interactions [51]. However, a trend has recently emerged into automatically detecting dynamic and subtle facial expressions from a continuous data stream (instead of single photos) driven by every-day applications e.g.[54], a practice which may be deterred due to amount of manual coding required. Additionally, distant heart-rate detection (via skin pigmentation) is currently a novel CV approach with interesting results [55], [56]. Camera based approaches for use in Virtual Reality settings are however nearly impossible to attain as the required headsets containing mounted displays cover a big area of a person's face allowing little space for camera installation within.

Without the use of multimodal approaches and deciphering the contextual information or the person's appraisals, the detection of facial expressions alone may not account for the accurate emotional state that the person is feeling, but habitually for the emotional state that they want to express to their interlocutor (posed vs spontaneous facial expressions of emotions, see [50], [57]). Detecting emotional states would require exploring a plethora of potential emotional encounters, stimuli and covariates while also collecting data from multiple sources and modalities, so that findings could be replicated and later applied outside laboratory settings, in real-life scenarios.

2.3.1.2. **Dimensional Model**

The rich spectrums of complex, non-basic and/or subtle affective states exhibited in our everyday interactions may not be fit for discrete label characterisation or by using categorical descriptions [58]. Instead, most studies in AC utilising multimodal methods, are classifying the affect using the dimensional approach [51]. Based on this approach, all emotions can be represented along a continuum. Most popular dimensional model is the pleasantness (Valence) and activation (Arousal) model, or else the Russel's circumflex model of core Affect (see Figure 1) [58], [59], and

model of positive and negative emotional activation [60], [61]. These dimensional aspects of one's experience are often collected using scales such as the Visual Analogue Scale (VAS) [62] and the Likert scale [63]. A vast majority of studies utilises the self-assessment mannequin (SAM) by Bradley & Lang [64], a questionnaire which combines visual representation of the scale's elements into simplistic figures together with a numerical scale (see Figure 2). To avoid low inter-participant agreement, participants are often asked to self-assess their own affective states using those scales which are then considered as 'ground truth'.
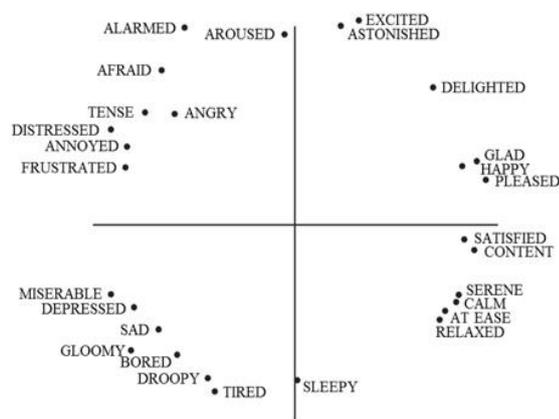
Figure 1. Russel's circumplex model of emotions.
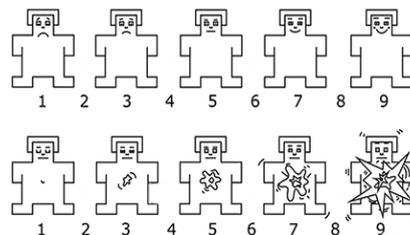
Figure 2. Example of the self-assessment manikins (SAM) for valence (top) and arousal dimension (bottom).

In short, valence can be described as the polarity of the affective state, or else, the positivity or negativity levels. Arousal on the other hand is the physiological and behavioural intensity of an affective state. For example, the corresponding increase in heart-rate or loudness on our voice, ranging from low (sleepy) to high (excited or stimulated) levels. Dominance, which is less often used, is the degree of control exerted by a stimulus or event, whether the viewer feels in control or not in its presence [65]. In the Circumflex model of affect [58], single adjective descriptions

of emotional states were rated by 343 participants in terms of arousal and valence and these ratings were then used to construct the cartesian space utilising those two dimensions (Figure 1), which are commonly used for emotional responses characterization in experiments using affective stimuli [66].

The use of this model in automatic affect detection is an on-going work which started only recently [67]. The state-of-the-art strategies involve the detection and classification of two and three levels along the valence dimension (i.e. negative vs. positive or including neutral) and the two or three levels along the arousal dimension (i.e. low versus high, or including an intermediate level) [51], [68]. Alternatively, researchers have been using the four quadrants approach (see Figure 3) which separates the dimensional space into four main areas of interest (High Arousal / High Valence, High Arousal / Low Valence, Low Arousal / High Valence, and Low Arousal / Low Valence) e.g. [69].
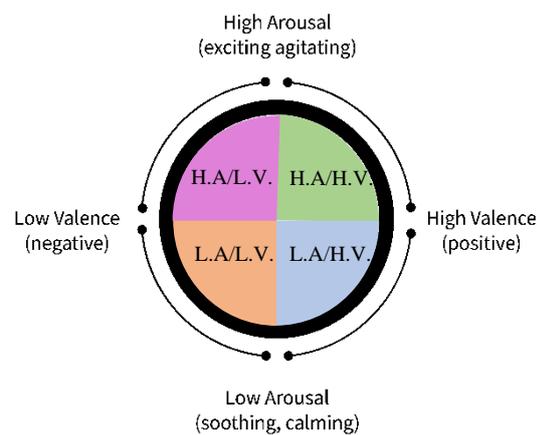
Figure 3. The four quadrants of High Arousal/high Valence (H.A/H.V), High Arousal / Low Valence (H.A./L.V), Low Arousal / High Valence (L.A./H.V.), and Low Arousal / Low Valence (L.A./L.V.).

2.3.1.3. **Appraisal-based approaches**

Emotion theorists linked the function of felt emotions to specific components related to the reactions and action tendencies [70]–[74]. This approach of modelling emotion is also described as the componential approach [75], where the variability of emotions is regarded through changes on components related to cognition, motivation, physiology, behaviour and subjective feeling related [67]. Based on this approach, an individual can react to a stimulus (or a sensory input) based on the

context, the meaning (the connotation for the individual) and the possible consequences. According to this theory, each emotion is characterised by physiological phenomena, behavioural action tendencies [76], [77], and by the cognitive appraisal's variables e.g. novelty, valence, goal relevance, goal congruence, and coping potential [78], [79], [80]. Thus, supporting the evolutionary adaptative nature of emotions, and the idea that specific emotions are linked to physiological changes [81]. In short, the affective reaction of a participant to a stimulus is determined by the coinciding impact of a number of appraisal or evaluative processes taking place for that individual, which are relevant to the corresponding meaning's implications.

The appraisal approach originates from Arnold's suggestion (1970) and from the Schachter and Singer's [82] experiment and theory of emotion (or Two-Factor Theory), where emotional states are dependent to cognitive factors; valence (pleasantness) is linked to cognitive appraisal of a situation whereas arousal is related to the physiological body state and hence the intensity of the emotion. In other words, when arousal is induced through a stimulus, perception and interpretation of its context provide the apprehension of the specific emotion experienced. Frijda [71], [83] suggests that situations where an individual is satisfied with attaining a goal, enhancing power of survival or demonstrate her capacities is usually accompanied by positive emotions. On the contrary, painful or stressful events could elicit negative emotions, as negative emotions motivate actions towards preventing those events from happening.

The componential models allow the study emotional states changes as a by-product of various configurations of appraisal dimensions. They can therefore potentially offer a more composite view of emotions compared to the discrete and dimensional approaches [67]. Generally, this theory emphasises on the impact of the subjective experience, the between-subjects' differences in emotion elicitation and the contextual information surrounding an interaction with a stimulus. Each stimulus or event is directly appraised and therefore its affective impact can vary between people. For example, a funny video could be positively rated by a large number of viewers, but an uninterested, stressed individual could rate it negatively. As such, the importance of human-centred interpretation of affective phenomena is emphasised rather than universal, generalised categorisation approaches.

This approach is also supported by the theory of constructed emotion or the conceptual act theory (CAT) model [84]. The CAT hypothesises that the

interpretation of ones and others' emotional states are perceived and categorised in emotion-related labels based on emotion concept knowledge of the perceiver. This knowledge is originating from prior experiences with one's body and the word (e.g. cultural concepts) via 'situated conceptualisations'[85]. In affective processing a situated conceptualisation, describes the construction of a concept derived from aggregated information across multiple interactions with a category member. As such, district emotions are constructed and not biologically hardwired, whereas affect is perceived (via interoception [86]) as the ground truth, continuous bodily sensory experience.

The appraisal models emphasise the importance of context, and conceptualisations which are relevant to each individual. Therefore, for a computational affect detection model to work accurately on each individual, subject-specific processes would need to put in place in order to map the reactions to the corresponding appraisal components. The application of such models in automatic emotion detection is still an open research question [51] although theoretical frameworks are starting to emerge e.g. [67]. Most automatic systems on emotion recognition are insensitive to contextual information (e.g. task and environmental factors) [68] which are strongly related to appraisal processes. By allowing the collection of such information we could potentially apply additional, appraisal-based layers of computational processing to infer emotion.

2.3.1.4. **The evaluative space model**

With the exception of negative emotions as anger, negative events or unpleasant stimuli can produce defensive predispositions and behavioural tendencies towards withdrawal, whereas positive stimuli can produce response of appetitive predispositions towards approach [87]. The Evaluative Space Model (ESM) (also known as general model of valence of evaluative experience) [87]–[89] combines the simplicity of the dimensional model with the action tendencies suggested in the appraisal-based models. It suggests that the experience of valence comprises two affective components, an appetitive and an aversive one that impact our predispositions towards (approach) and against them (withdrawal) (see Figure 4). Based on some of the main ESM postulates, positive emotions are more similar to
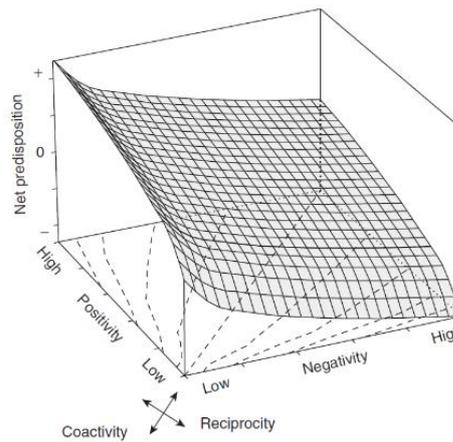
Figure 4. The evaluative space based on the ESM model as illustrated by Cacioppo et al.[87] The surface represents the net predisposition of an individual towards (positive) or away from (negative) a target stimulus.

each other than to negative ones, , the positive activation function's offset is higher than the negative one and that the motivations towards withdrawal is stronger than the motivation to approach (negativity bias) although highly influenced by individual differences [87]. Behavioural patterns in space can be monitored through trackers, thus, the detection of valenced predispositions via the calculation of approach and distancing of an individual from targeted stimuli is feasible within VR.

2.3.1.5. **Section Discussion**

The major emotion theoretical models describe the processes and characteristics from which emotional states can be inferred. From these models, the discrete and the dimensional model have been mostly used in affect detection applications. The discrete emotion model has been extensively used for automatic affect recognition and its main advantage lies on the labelling system itself, as it is intuitive and easy to match with the categorical description of emotions we use in everyday life. However, in everyday interactions we tend we express a range of dynamic emotional states and affective intensities that cannot be fully engulfed by discrete categories.

The dimensional model on the other hand, can describe a wide range of emotional valences and intensities. Although some emotional states may seem more difficult to distinguish, the dimensional model could be easily integrated in a multimodal automated affect recognition system as there is an extended body of research linking physiological patterns to those dimensions. The terms used in both

the dimensional model and the evaluative model (ESM) are similar, i.e., positivity-pleasantness and negativity-unpleasantness characterise valenced affective states, while the intensity of the emotion is characterised by the level of arousal. The observations of dynamic behavioural effects of approach or retraction from a stimulus in a virtual experience or a gaming environment is feasible via motion trackers. Hence, the incorporation of the ESM in VR-based affect detection could prove to be beneficial for multimodal valence detection, in addition to measuring the dimensions of affect via physiological measures and written or verbal reports.

The appraisal model on the other hand is more challenging to incorporate as it requires the extraction of context and user-specific information, which is traditionally omitted in state recognition approaches. Ptazynski et al. [90] highlighted the important of context in emotion recognition by arguing that such states cannot be identified in real-word settings independently from the context in which they were experienced. We however see that some of the that information could be derived from a virtual environment as all the objects and events can be carefully designed and the user interactions within can be monitored. User information could be also potentially inferred via questionnaires and the analysis of user interactions within and outside VR. In general, contextual information related to the user, the task and the nature of the stimuli within the virtual reality experiences could potentially provide ground for more advanced user-based analysis and ultimately the link between subject-specific appraisal components and affective responses.

Although the advancements of technology and science has broadened our understanding on emotion recognition, processing, and expression, the mechanisms behind the elicitations of emotional responses from individuals during the experience of affective content are yet to be fully understood. In addition, the importance of individual differences in emotional expression and its changes based on context could be preferably addressed in the future. In the area of Human Computer Interaction (HCI), the results from such research could enhance future development and improvement of affective applications and content, and assist in the creation of improved, subjectively tailored emotional experiences in XR and interactive media.

For the studies explained later in this thesis, the two-dimensional model of emotions of valence and arousal was adopted, as the most suitable model for the proposed research. A two and three classes approach per dimension was applied.

Some key aspects from the appraisal and ESM models were also used. More specifically, we analysed movement data in relation to the nature of events as additional sources of data. In addition, we extracted user-relevant information on personality, alexithymia and felt affect via questionnaires. Stimuli-relevant information for each event or object in our custom-made virtual environments as name, type, properties and positions were also collected. This way, we could continuously track what the user was seeing and responding to.

The changes along the valence and arousal dimensions for felt affect has been found in previous research to relate to changes in physiology and behaviour. In order to detect those changes, we employed similar methodological approaches. In the next section, we describe the main physiological measures and modalities used in our research, followed by an overview of the related research and the emerging affect detection approaches for virtual reality.

2.3.2. Emotion externalisation & related physiological and behavioural metrics

There is a distinct differentiation between emotion expression and emotion recognition in an interactive framework. In a minimalistic view, emotion expression involves the emotion externalization from a *Sender*, a role which can be attributed to a person or character, either real or virtual, that has the ability to be emotionally expressive. Emotion recognition concerns of a second participant, the *Recipient* who experiences and interprets the emotional externalisation of the Sender and adapts or maintains their behaviour. As a result, an iterative feedback system is developed (see Figure 5). This system enables the communication between two people or more. With the application of affect sensing techniques, it could also enable the communication between a computer and a user.

As stated above, human emotion expression is a form of communication,
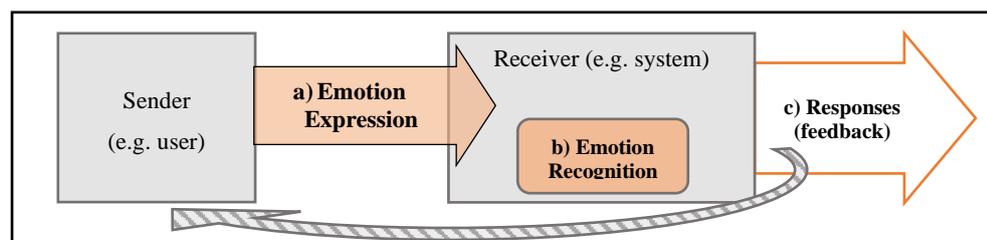
Figure 5. a) Emotion expression (Sender), b) Emotion recognition (Receiver) and c) responses to recognized emotion.

usually signified via physical mediators, such as our voice and our body language using gestural and facial expressions, coupled with physiological changes [91]. We could argue that emotion in humans is manifested on a three-dimensional mapping involving behavioural, verbal and physiological components, often coinciding with each other based on the findings from several scientific studies (e.g. [10], [38], [72]–[76]). More specifically, emotion expression could be categorised into three expressive mediators:

- Facial Expressions: related to emotional states externalization and nonverbal communication. The Facial Action Coding System [97] is commonly used to link emotional states to facial configurations. Yet, facial expressions can additionally imply mental, cognitive and physiological states [98]–[100].

- Bodily Behaviour: related to bodily responses including physiological changes [91] and body cues [101]–[107].

- Verbal Behaviour: including emotional content via vocalization [108], [109] and emotion felt self-assessment, via questionnaires and interviews.

As there is a wealth of evidence in the coalition of those mediators in emotion manifestation, ideally a complete multimodal investigation could combine and cross-validate readings from all three mediators. Features or patterns from each mediator can be associated with certain affective states. In the following sections we delve into the processes affecting each mediator in terms of their nature and how they are generated. We will discuss their homogeneity across individuals, their level of sensitivity to affective changes and the sensors that were employed in order to continuously monitor and measure their changes

2.3.2.1. Facial Expressions: Measuring muscle contractions via electromyography (EMG)

Our body, and especially the face, is considered as one primary expression mediator of the individual's emotional states. Many scientists as Duchenne de Boulogne and Charles Darwin have been investigating the production of facial expressions, their meanings and how they were used to communicate inner feelings. Duchenne believed that understanding facial expressions could reveal an "accurate rendering of the soul's emotions" [103, p.58]. Facial expressions can be triggered spontaneously, in response to emotional stimuli like a funny video (referred to as 'facial motor resonance'), or voluntarily like in a social smile or a posed expression (also referred to as 'facial mimicry') [111]. Although facial expressions can be self-controlled or suppressed, they can appear consciously or unconsciously in various settings (e.g. [13]). In our everyday life, facial expressions as other bodily responses can be expressed within the interactive framework as a form of communication, but they can also occur as an externalisation of the internal affective state when exposed to physical stimuli or mental images. Various external factors including emotional contagion, experimental bias, and cultural factors, can affect the externalisation of our facial expressions because of our high awareness of our own facial movements.

The combination of high awareness together with our innate ability to control certain facial expressions, makes us able to use posed expressions on our own volition when wanting to express a specific emotion or suppress another.

However, for applications to real-life scenarios, the detection of posed expressions of emotions may be less informative as they differ from those that spontaneously occurring in naturalistic settings [112], [113]. In other words, the facial movements intrinsic to posed expressions can display an emotional state that the sender intends to convey, while the expressions resulted by spontaneous facial movements resemble to the sender's/user's real, unmitigated affective experience. For example, when a person is genuinely happy, apart from the muscles activated when smiling to raise our cheeks (such as the 'zygomaticus major'), the muscles surrounding the eyes ('orbicularis oculi') are also contracting (resulting to the formation of the genuine or 'Duchenne smile'[114]). A genuine smile is less susceptible to be falsified, as the muscles around the eyes are more difficult to be activated voluntarily [115].

However, the majority of past expression detection research focused on voluntary posed expressions [19], with few studies investigating spontaneous expressions e.g. [116]. Assuring that the facial expressions of a person are sincere and spontaneous is a difficult task for researchers. For example, an inattentive study design could give away the main research goals and apply bias to the participants from simple details and/or instructions. Researchers in this area often resolve into concealing information relevant to the goals of the study (and the measures used for the detection of facial muscle activation) in order to collect sincere responses from the unwary participants.

Generally, facial expressions are the result of the contraction of set of facial muscles. Arising from the bones to the skin, the muscles' orientation of the fasciculi or else the muscle's fibres can be parallel (linear), oblique or spiralized relative to the direction of pull at their attachment [117]. There are more than 20 muscle sets on our face, from which we can extract the foremost muscle regions which impact on the formation of facial expressions of major emotions. Some of those are the zygomaticus major, corrugator supercilii, frontalis, orbicularis oculi, depression anguli oris, orbicularis oris and the levator labii superioris alaeque nasi [118](see Figure 6). From those, the major muscles situated under the area of a standard VR HMD are the zygomaticus, frontalis, corrugator and orbicularis oculi, whose positions are shown in Figure 7.
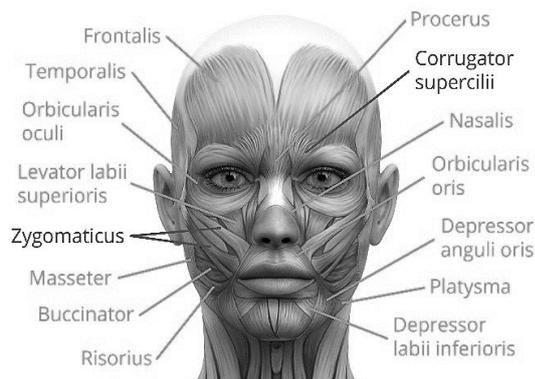
Figure 6. Facial muscles, adapted from [562].
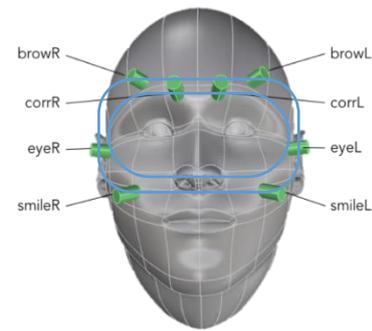
Figure 7. The EMG sensor-locations used in our studies.

Relevant to the dimensional model, the zygomaticus major and the corrugator have been extensively investigated for the distinction of positive versus negative valence, as they are responsible for smiling and frowning [119], [120]. There is however an on-going debate on (1) whether these recordings are specific enough to detect spontaneous positive or negative valence, and (2) whether reading from those muscles alone can suffice for valence detection or whether other muscle groups should be added [120]–[122]. Concerning the first point, one of the key-issues is that readings from sensors positioned on the cheeks (over zygomaticus major) can detect higher activations during smiling but also during other expressions/facial configurations as the muscle lies within close proximity of other muscles like the buccinator, masseter and zygomaticus minor [123] inducing high rates of cross talk between them. Additionally, aversive and negative expressions can also induce motion on the cheeks [120]. On the other hand, the corrugator muscle, which is responsible for frowning, can be activated in negative valence conditions but also during cognitively difficult tasks.

Concerning the second point, emotional facial expressions can be more accurately detected when recording from several facial muscle groups simultaneously. For example, during a 'Duchenne smile', readings from both the muscles surrounding the mouth and the eyes could serve as a better indicator of spontaneous positive valence[124]. The frontalis or "brow" muscle is responsible for the raising movement of our eyebrows, stretching on top of our forehead. Such movements are attributed to dynamic expressions of generalised fear, anger, but also of surprise, which could be of ambiguous valence. The measurement of the activation of multiple facial muscles in parallel could allow the discrimination between facial activations from which affective states can be inferred.

The detection of affect via the inherent facial muscle activation of a user may be superior to the detection of their facial expression by video capture, as it is the underlying mechanism that creates those facial configurations and as the detection of those does not entail a subjective characterisation (e.g., forming a smile by lateral mouth movement equals to positive affect). In other words, the detection of individual muscle activation does not require their translation by a human expert observer as in the case of FACS (it is perceiver independent) and therefore a more objective way of evaluating facial movement all-together. Additionally, facial muscle contractions can be of different intensities (as in micro-movements) which are sometimes difficult to perceive with cameras and also difficult to 'fake' [125]–[127]. Researchers supported that during facial motor resonance, the changes on the orbicularis oculi muscle (around the eyes) during a genuine smile were not visible to the eye, but detectable by electromyographic sensors (EMG) [128], [129], which makes EMG a superior methodology compared to video and computer vision.

Although, some facial muscles like the zygomaticus major are easy to activate voluntarily, other muscles like the corrugator supercilii which is bilaterally innervated hinders fine voluntary motor control [120]. Nevertheless, there is a large number of facial muscles whose correlation to affect has not been investigated in depth yet. As the majority of facial muscles are extremely close to one another (especially crowded in the area of the cheeks [118]), their activations can be correlated, resulting to difficult discretisation of their activation. Perhaps detecting activation simultaneously from a group of areas superimposing major facial muscles and detecting affect based on their dynamic relative function, may prove advantageous compared to detecting activation from a specific muscle.

Surface EMG is commonly the method used to measure underlying muscle contractions using sensors applied directly on the skin. These sensors can detect changes in surface voltages on the skin when muscle activations occur. Nowadays, measuring facial expressions and emotional responses using EMG is a fundamental tool for researchers in media, marketing, gaming and psychology [32], [115], [120], [130]–[134] and medical practitioners i.e. for the assessment of facial palsy [135], [136]. EMG has been used for the measurement of affect [91], facial expressions [19], gestures [137], fatigue or tiredness [138], stress [139] and pain [140] amongst others.

Traditionally, two EMG sensors are placed along a muscle while also a third sensor is used as a reference or 'ground' sensor, placed on an area of the body or

face which is less likely to be affected by movement, like on the top of the forehead. Tethered, adhesive-based sensors are typically used in past research, which may require the additional application of conductive gel [141]. Such sensors can be used on various parts of the body. However, the application of multiple sEMG sensors on the face can unavoidably be intrusive and cumbersome to use. Data collected from those sensors are amplified and converted to microVolts (μV).
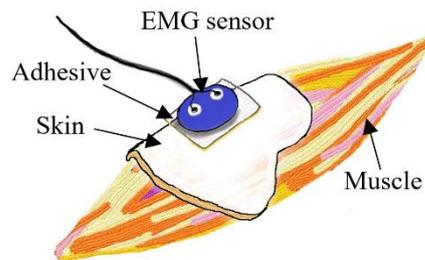
Figure 8. Representation of surface EMG sensor on skin.

The activation of a muscle in focus is visible via the visualisation of the EMG filtered data-stream. An envelope of the signal is commonly computed from the root-mean-square (RMS) values of the EMG signal to facilitate interpretation and the visualisation of the activation power [142]. An example of filtered signal and their corresponding RMS envelopes from multiple EMG sensors recorded simultaneously from the face is shown the Figure 9.
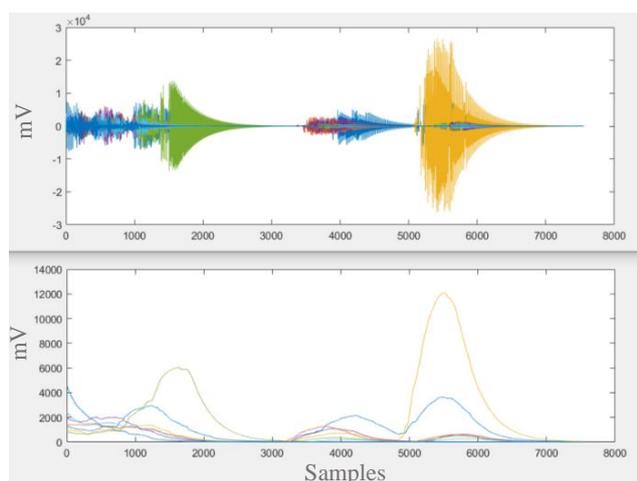
Figure 9. Up: Filtered EMG signals from seven sensors. Down: the RMS values of the sensors. The EMG-derived values here were multiplied by 10^4 for these visualisations.

2.3.2.1. **Bodily behaviour: Measuring arousal levels with skin conductivity and heart-rate changes**

Various studies investigate the activation of the autonomic nervous system (ANS) during emotional stimulation. For this, they have used different cardiovascular, electrodermal, and respiratory measures [40]. The most common cardiovascular measure is heart rate (HR), followed by blood pressure (systolic and diastolic BP), heart rate variability (HRV) and temperature. Electrodermal responses as a parameter of the sweat gland function (either phasic or tonic, see [143]) measured are the galvanic skin conductance and resistance (level, response rate and amplitude). Respiratory measures such as breathing rate, variability and period are also less frequently included. As electrodermal and respiratory measures were not included in our studies they will not be further discussed.

Heart-rate activity can be measured with electrocardiographic (ECG) and photoplethysmographic (PPG) sensors, both of which are non-invasive and exosomatic, placed on the skin of the wearer. During an ECG recording, usually two (e.g. in a biometric scenario) or more electrodes (in clinical settings) are placed on the chest of a subject, in order to capture a clear and precise signal of the heart beats (namely R-wave, including the QRS complex[144]). This signal is achieved by measuring voltage changes on the skin resulting from the cardiac beats. The electrodes are normally applied on the upper abdominal and thoracic area [145]. Standard ECG acquisition practices typically require a careful preparation of the skin area before the signal acquisition; some systems even require the application of gel to the area under the electrode. Alternatively, ECG belts can be used more effortlessly compared to applying individual electrodes on the skin of a user. Unfortunately, both methods can constrain the motion of the wearer if tethered.

A photoplethysmograph (PPG) measures the reflection of the illuminated light from an oximeter onto the skin, thus monitoring the changes of the absorption of light resulted from blood-volume's changes. Consequently, PPG is able to detect the cardiac cycle from the area of the skin where the PPG is applied to. PPG sensors need to be applied on a close proximity to the surface of the skin. As these sensors do not rely on electrical activity, they do not require gels or adhesives upon application. Thus, integrating PPG sensors on wearables is cost-effective and the application of those sensors is more effortless than the application of ECG electrodes or the usage of ECG belts. However, the signal acquired by a PPG measure is less

detailed than the one monitored by an ECG, and it can be subjected to noise artefacts resulting from the user's movement, wrong placements on the skin, and the effect of other physiological parameters that can affect the blood flow, such as respiration and cardiovascular circulatory conditions [146], [147].

Overall, as PPG sensors are typically inobtrusive and can be placed anywhere on the body where the blood vessels are close to the surface of the skin, they make excellent candidates for wearable integration; especially when considering the development of a VR wearable solution for heart-rate detection. In the experiment presented in Chapters 4, a PPG sensor was placed on the wearable VR insert. To assess if its quality was sufficient, the PPG signal was examined and compared to the signal from an ECG-belt. Examples of both signals are presented in Figure 10.
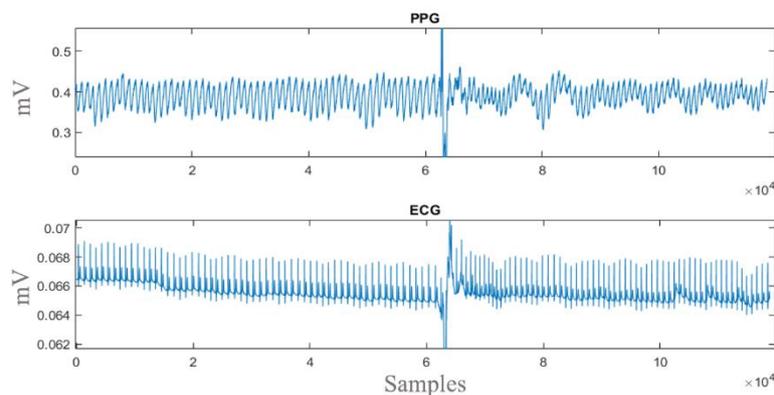
Figure 10. A segment of simultaneous recordings made with PPG and ECG sensors from a user who is initially is seated and stands up (middle of the recording).

With the right set-up, both methods, PPG and ECG, can be used to measure heart-rate as beats-per minute (BPM) and heart-rate variability (HRV). Heart rate is commonly used in medical settings and it is integrated in an abundance of modern wearable devices, e.g., smart watches. If BPM is detected to be above or below the person's 'normal' or 'healthy' range, it can indicate illness and other serious conditions [148]. HRV indexes on the other hand are considered one of the most promising markers of the ANS regulation with links to adaptive emotion regulation and related processes. Both BPM and HRV can show changes in physiological arousal between resting and active conditions. For example, in high-arousing, anxiety-inducing conditions BPM is expected to increase and HRV tis expected to decrease [149]. Heart-rate variability (HRV) is the study of the variation of the

successive R-R intervals, or else the time distances between continuous heartbeats in a specific time window. HRV can be affected by sympathetic and parasympathetic processes of the autonomic nervous system, which can be affected by changes in physiology and environmental factors. A such it has been used to assess emotional arousal [150], valence [151], mental workload [152], [153] and stress or anxiety states [154]. That being said, HRV and well as BPM can be affected by confounding factors unrelated to emotional changes such as age and medical causes, including physical, breathing and cardiovascular conditions [155].

The average BPM can be computed from a series of two or more peaks, thus allowing for its computation within short recordings. On the other hand, long-term signals recordings (in terms of duration) are suggested to be most informative for HRV[151], [156]. However, when it comes to active recordings, in different studies where the participants are engaged on a specific task that requires movement, short time-windows were also used, showing adequately rich information for every part of the active task [150], [157], [158]. Short time windows may vary from 10 seconds to 5 mins and more [159], provided that the signals have high signal-to-noise ratio.

HRV indexes contain features which can be computed via time-domain and frequency-domain analysis. The most common methods include the root mean square of successive differences (RMSSD), the standard deviation of beat-to beat intervals (SDNN) and the proportion of the number of R-R interval pairs that differ more than 50ms divided by the total number of R-R pairs (pNN50) (see [158], [159]). As these features measure the variability of the distances of the successive NN peaks, they depend on oscillations of high frequencies thus being impervious to the use of short time windows [159]. The proportion of the signal within certain frequency bands (and their ratios) identified such as, typically low frequency (LF: 0.04 - 0.15Hz), high frequency (HF: 0.15-0.4Hz) and very low frequency (VLF: 0.0033 – 0.04Hz), can show different fluctuations caused by the parasympathetic nervous system and breathing [149], [160], [161]. For example, changes in HRV during acute stress may show a reduction in HF power from baseline [156]. However, accurate lower frequency detection requires long recordings (min. 4 minutes long) [112], [156]. The features selected based on the available recording durations and environmental conditions for the analysis of the experiments presented in this thesis are presented in Chapter 4 and 6.

2.3.2.2. **Bodily Behaviour: Measuring body movement**

The body itself, can mediate parts of the person's emotional experience though its movement ([15], [41], [162]). There were many attempts to correlate specific body movement's characteristics or postures with emotional states (e.g. [163], [164], [165]) . Such changes in our body movement (or the movement of our body parts e.g., hands) can be measured in terms of changes in orientation, velocity and shape of the movement.

Motion or limb tracking can be attained via the use integration of inertial measurement units (IMUs) comprising of gyroscope, accelerometer, and magnetometer sensor along three axes, x, y and z. Such sensors can be easily integrated in wearable solutions and are non-invasive. A wealth of research is utilising inertial sensing for activity recognition in active experimental protocols (e.g. [166]) and for inferring the underlying emotional state of the user [167][168].

We are constantly using motion tracking sensors with current VR technologies, especially when using HMDs and hand controllers with 3 and 6 degrees of freedom (DOF). In HMDs, motion tracking systems are responsible for the synchronisation of head movement and display angle. In other words, the user's actual physical movement is reflected on the user's point of view within VR, and this is attained via integrated sensors onto the headset and motion sensing cameras when using room scale motion tracking (based on the degrees of freedom each system is allowing). Users in room-scale VR with 6DOF can walk and move as they would in the real-world using their bodies. Based on the ESM model (approach-withdrawal) the analysis of movement patterns to specific contextual information in VR could offer a rich source of valence information regarding the affective state of the user.

2.3.3.   Affect detection approaches using VR

VEs and simulations offer many benefits as part of the experimental process. With VR we have the opportunity to create any environment or scenario that is programmatically and computationally possible, while being able to track and monitor the user. The advantage of VR technologies is that users often get deeply immersed in these virtual realities, that they often feel they are actually existing in those virtual spaces (feeling *presence*), maintaining reduced awareness of the real space and, therefore, reacting in a naturalistic way. There are various factors that can influence this feeling of presence such as technology and content (e.g. head mounted display (HMD), unobtrusive input technologies, the environmental attributes of the simulation, the task's nature, interactivity, etc.), as well as subjective traits of the user which can enforce or reduce the effect of VR [169]. Research studies have suggested that those participants who feel a high level of presence in VR, will interact in VR in a naturalistic way as in real-life [170]–[172], which supports the argument concerning VR suitability for psycho-behavioural related research.

Therefore, realistic emotional responses, of behavioural, physiological, and vocal nature, are expected to be elicited in VR settings. However, the development of methodologies focused on the detection of those in VR settings is still in its infancy (with the exception of eye and body tracking technologies). In recent years we have seen a drastic increase in the development of emerging wearable technologies integrating biometric sensors, which can be worn simultaneously with an HMD for use in VR. This sudden increase in prototype development is coupled with a further increase on the number studies focusing on detecting affect from physiological signals in VR, published in the last five years. A review of the related studies will be discussed in the next sections.

2.3.3.1. **Virtual Reality as an experience: Presence and affect**

VR as a medium is highly advantageous over traditional experimental protocols. VR technologies have the potential to simulate real-world interactions which can be highly immersive and activate intense emotional reactions [173]. Emotional elicitation within VR is linked to the level of presence, the impact of immersion and the overall involvement of the user with the mediated experience [174]. VR is widely accepted as a medium whereby the experience of presence is able to occur [175],

[176]. Presence is described as the subjective "sense of being there" and among others, the "perception illusion of nonmediation" [177], [178]. In other words, while feeling present in VR, the user's senses and cognition are preoccupied with the virtual environment in a level that her awareness of the outside world disappears, and the virtual simulation becomes the reality. In this context, the person is expected to feel and react to the simulated situations in similar ways as in the real-world. However, the level of presence that a person can experience in VR is dependent on various factors whose impact may vary between individuals e.g. personality and emotional state [11], [179].

Definition and assessment of Presence – There are many existing theoretical models for measuring presence in VR. These models can be divided into descriptive (defining the components of presence) and structural ones (focusing in the process of the generation of presence, cognitively and mentally) [180]. In those presence can be measured as a result of several factors or properties of the VR experience. These factors were derived from self-report acquired from participants and notes made by external observers. In 1988, four factors were suggested by Witmer and Singer [181] in terms of control, sensory, distraction and realism levels. Additionally the factors of vividness, interactivity/influence on content, and user characteristics were emphasised by Steuer [177] in 1992. The division to exogenous and endogenous factors were later suggested by Slater [182] in 1993. Exogenous factors are related to the fidelity of the simulation system and are the necessary conditions for presence to occur, such as the quality of the interface. Endogenous were those factors that affect the subjective experience of the user, related to the overall interactivity, virtual body, anticipation of action effects and consistency. Since then, many questionnaires were constructed to engulf and measure those factors, and additional ones were also suggested related to the social nature of the experience [183] and the engagement or enjoyment of the user [184](also see [185]).

Presence is however by definition highly subjective [186], and therefore the effect that all these factors could have on the subjective experience of presence for each user can vary. For example, a VR simulation where all its objective components are prespecified, may be perceived as more or less immersive by different people. To disassociate the subjective feeling of presence from the objective characteristics of the simulation and the technology used, we followed the terminologies of presence and immersion as suggested by Slater ([187], [188]).

Immersion is one of the contributing factors of presence [176]. Slater and Wilbur [189] defined immersion as the objective technological qualification of the VR related capabilities and equipment employed e.g. HMDs and trackers equipped. Therefore, the feeling of presence is the human subjective response to the virtual reality experiences and to the level of immersion attained by the system [188]. Immersion has been found to increase emotional responses in simple neutral environments [190]. The level of immersion can be increased by several factors, as for instance: the number of human senses for which a medium provides stimulation (i.e. media sensory outputs), the consistency of sensory outputs, content features, resolution, field of view, virtual lighting, motion, dimensionality, camera techniques, aural presentation characteristics, interactivity, obtrusiveness of a medium, and the capabilities of the system for social interaction [178], [191], [192].

With today's technology it is possible to create highly immersive experiences due to the high-quality portable and less obtrusive VR inputs/output technologies (HMDs, controllers, tracking sensors) and the additional processing and graphic power of computers. Great consumer-ready examples used in VR research are systems like the HTC Vive [193], Oculus Rift [194], PS VR [195] which include high-quality displays, head and limb tracking locomotion, as well as low latency and high-quality audio. Generating experiences with high interactivity and control mechanisms, ecological validity (term referring to the content's richness/vividness[196]) and rich narratives is easier than ever. From this perspective, it is easier with today's technological capabilities of the VR systems to build the necessary foundation for presence.

However, this may not be the case for all VR users. Specific personality characteristics such as imagination and empathy[197], and certain emotional states such as fear and stress [176], [198], [199] have been found to be linked with higher elicited presence levels in VR. The experience within a virtual reality can become more realistic when a person in able to suspend their disbelief [177] and allow for expectancies to be generated [200]. The ability or willingness to suspend disbelief [201] and the suppression of the actual feeling against the expected interoceptive emotional states [202] are suggested to alter and intensify the effect of presence in virtual experiences. Measuring a person's susceptibility to feeling presence in VR is challenging because to our best of knowledge, there is no questionnaire that specifically addresses presence proneness or suspension of disbelief in VR settings. As a way around this issue, researchers often tend to use fantasy proneness or

creativity experience questionnaires and absorption scales, such as [203]–[205]. However, absorption is not as clearly defined as presence because it contains aspects that are related to the big-five personality traits [206]. Witmer and Singer [127] designed an immersive tendencies questionnaire that measures the ability to get involved/immersed in everyday activities as a proxy to the virtual experiences. However, researchers argue that scores from this questionnaire accurately correlate with presence levels only in settings of high presence [144], and that items of this questionnaire in fact measure the engagement of the user [20] .

Although predicting the subjective ability for increased presence in VR is challenging, presence can be measured within and outside the VR experiences. Up to this date, the level of presence in VR is commonly measured via observations and questionnaires, which are answered by the user retrospectively of their experience (post-experience). Frequently used questionnaires are the Presence Questionnaire by Witmer & Singer (QEP), Slater-Usoh-Steed Questionnaire (SUS), Krauss et al. Questionnaire, Presence Questionnaire (PQ), Igroup Presence Questionnaire (IPQ), ITC-Sense of Presence Inventory (ITC-SOPI) and others (list of questionnaires can be found in [208]).

Alternatively, changes in physiological signals are suggested to show changes in presence [209]. Such physiological indicators of presence include galvanic skin response, heart-rate responses, skin temperature, muscle tension, and pupillometry [210]. However, as physiology can be affected by various parameters e.g., motion, emotion and stress, the reliable investigation of the true relationship between physiological changes and presence may be more complicated than expected. It is possible that emotional responses and by extension physiological changes are the by-product of highly immersive experiences. Freeman et al. [196] alerted researchers that such correlation could be limited to arousing stimuli, suggesting a presence model where arousal / alertness contribute to higher presence. Although this theory has not been systematically tested, significant links between physiological measures of arousal, such as heart rate and electrodermal activity, were found to accompany higher presence in immersive scenarios [27], [211].

Just like with emotion assessment, most 'ground-truth' methods of presence evaluation are focusing on written methods and ratings, either by the participants themselves or by external observers/perceivers. One of the caveats of using only self-rating questionnaires for the evaluation of presence levels and emotional states is that, in most cases, individual ratings take place before, rarely during and/or after the

experience, at a dedicated and specified time window. Since a VR experience by default requires a duration of interaction (similar to watching a video or completing a task), interrupting the experience or completing the questionnaires at the end could provide attenuated and totalized results which are not representative of the moment-by-moment feelings. For example, if a positive event happened within a generally negative experience, the overall rating of emotional valence could be affected. Therefore, the various levels of affect and presence may be better investigated in a fine-grained temporal level of analysis. This way, we could also account for the effects of highly memorable events over others e.g. events with high emotional impact and events that happened at the beginning and towards the end of the experience (based on primary and recency memory effects) [212],[212]. Ideally in the investigation of the link between affect and presence, ratings and reading should be recorded at shorter time windows or continuously while also respecting the flow of the experience of the user and avoiding any breaks in presence.

The relationship between emotion and presence – It is considered that during an immersive virtual experience users may react in a similar way as in real-life conditions [213], a phenomenon that potentially derives from the effect of presence [213]–[215]. Overall, the level of presence and the intensity of emotional reaction in VR are appearing to be correlated [1], [216], [217]. In exposure therapy examples using fear inducing stimuli, presence is found to be strongly linked to the emotional responses (e.g. [216], [218]). However it is unclear whether presence is a pre-requisite for enhanced emotional responses or whether emotional stimulation enhances the feeling of presence (as in [170]). Perhaps those phenomena can enhance each other in an iterative way, as long as there is an initial minimum level of presence.

These unresolved specifics of the relationship between presence and emotion are exciting and deserve further investigation. Research findings show that the level of presence (and immersion) plays a significant role in the development of emotional stimulation in virtual experiences [180]. Parameters that can potentially increase presence, the technologies used, the content and the congruent narrative scenario can influence the elicitation of emotional experiences of users in VR [215]. In comparison pilot studies using immersive and non-immersive mediums found that negative visual stimuli were rated as more emotionally arousing and more negative, when presented in VR settings compared to a screen [219]. Yet many support that

presence constitutes a requirement for emotions to be elicited and it does not play a direct role into their intensities [220]. Nevertheless, presence in most cases is recorded as a one-off point at the end of the experience. It would be interesting to explore the relation of moment-by-moment presence levels to intensities of the emotional dimensions is we develop the means to. For example, would be interesting to assess presence using multiple assessment measures within multiple randomised scenarios with various valenced and arousing events.

Section synopsis

VR is a medium, that unlike others, has the ability to make the users believe they 'are' part of it, existing within its simulations. In these experiences, the users' cognition, perception, and various sensorimotor exigencies are activated resulting into the feeling of presence. In this section we explained the terminologies of presence and immersion used in this study, and the current issues behind their measurement. The link between emotional stimulation and presence was described through the prism of the related research.

In this thesis, the link between subjective valence, arousal and presence ratings will be investigated, via objective physiological measures and self-ratings. In addition, potentially mediating subjective factors will be also collected and considered for the interpretation of the findings. These factors are, for example, personality traits and emotional elicitation (e.g., alexithymia). Identifying the nature of the correlation between presence and emotional stimulation could provide valuable insights for future research.

Congruent content narrative, interactivity and naturalistic movement (in the psychical space as in the virtual one) are some of the factors suggested to increase presence [221]. Within the last five years, VR technologies have hugely improved, allowing for capabilities which were not as easy and cost-effective to use in research in the past. These capabilities and controlling inputs which can set high immersion levels. They now allow room-scale interactive experiences for large scale studies, meaning participants can freely walk within the physical space. This can be done using cameras that track the user's position and rotation in space. Keeping the immersion level persistent across participants, it will be interesting to investigate the effect of first-person, 3-dimensional, room-scale experience on presence and emotion against static-passive, vicarious experience.

2.3.3.2. **The effects of emotion on cognition, and the link to subjective factors**

Emotions are affecting cognitive processes such as perception, attention, memory, judgement and risk-taking when making decisions, but this effect is not always straight forward. These cognitions play an important role when processing mediated stimuli [222], [223]. For example, presence together with intense emotional stimulation can affect our perception of the environment around us (exteroception), our proprioception (sense of our body in space), locomotion, and kinaesthesia (sense of movement). The Yerkes-Dodson Law suggests that performance increases with physiological or mental arousal up to an optimal level. However if that level is surpassed and arousal increases higher, then performance decreases [224], [225] negatively affecting cognitive processes like attention and memory. This effect is shown in behavioural studies using highly distracting stimuli, which capture the viewer's attention and result in attenuated performance during cognitive control tasks such as task switching, inhibitory control and memory target detection tasks [226]–[230]. The same effect applies to memory related process. For example, affective stimuli are generally more memorable compared to neutral ones [231] (with the exception of negative traumas [232], [233] where the memory is inhibited). In VR for example, Sutcliffe and colleagues suggested the use of memory recall tests to find negative effects caused by usability constrains, perceptual distortion and interactive controls [234]. Other studies, using VR, have reported that in healthy populations negative arousal and anxiety can negatively affect the storage and memory retrieval content of temporal and spatial information [235], and memory retrieval related to the actual physical world [236]. In decision making, the appraisal of a situation related to the avoidance of negative and the expectancy of positive emotions can drive our decisions [237]. However, as with memory retrieval and attention, high arousal and anxiety can hinder advantageous decision making [238], [239].

Appraisal theories suggest that emotional elicitation can be also influenced by subjective evaluations of occurring events (e.g. [72], [74], [240], [241]). Since subjectivity forms a decisive factor on emotional experience, any sustained variability in personality, state traits, or mood could impact on emotional processing. For example, negative valence and high arousing states (such as anxiety) have been postulated to affect processing effectiveness and attentional control [242], [243]. Certain personality traits, such as extraversion and neuroticism, have been found to

correlate with valence responses [244]. More specifically, people with high extraversion tend to self-rate stimuli more positively and neurotic people tend to rate more negatively.

Additionally, alexithymia levels and interoceptive awareness (inability to identify and describe one's emotions) could hinder subjective emotional appraisals and therefore affect emotional reactivity [245], [246]. The term alexithymia defines one's ability to interpret, process and describe the emotions of themselves or of others. High alexithymia may reflect deficits in cognitive processing and regulation of emotion [247]. Although, high alexithymia was not found to strongly correlate with changes in valence, multiple studies have shown that alexithymia is linked to reduced physiological arousal (referred to as 'hypo-reactivity'), assessed via hear-rate and skin conductance measures [248]. By comparison, individuals with intense behavioural expressive tendency or else, emotional expressivity, tend to self-rate and externalise their emotions effortlessly [249], [250].

Therefore, the synergy of cognitive evaluation during emotional processing with subjective covariates such as trait factors could provide crucial additional information about the user's expected emotional responses.

2.3.3.3. Current Emerging Affect Detection Methodologies in VR & future directions

Understanding the emotional state of the user in VR could assist in a range of use cases. It would aid real-time continuous affect recognition and the awareness of the user's state changes, affective design and adaptive control of the surrounding environment. Adaptive control is when specific signals can be utilised to alter the environmental parameters, which in turn can possibly alter the user's affect, as a feedback loop. In this section, we will present the current emerging affect detection technologies for consumers and researchers that are developed to either specifically mitigate practical issues of combining physiological sensors with modern HMD VR devices, or to provide wireless wearable capabilities with can used in VR settings.

Apart from conventional unimodal methods such as camera tracking or heart-rate sensors, recent software and hardware prototypes have emerged that combine multimodal approaches and affective read-outs specifically adapted for real-time applications. Commercial technologies including, Emotiv Epoc, LooxidLabs, Enobio, Neurable and EmteqVR [251]–[255] have emerged in recent

years to provide real-time emotional feedback and affect recognition readings in VR. Although only a small number of studies using these technologies in VR are published, we were able to gather some of the more relevant findings as well as the practical implications of each technology.

Arousal detection in VR, and especially the detection of stress, has been synonymised with analysing heartrate and electrodermal activity (EDA) changes [187]. The Q sensor by Affectiva [256] a wireless wearable biosensor has been used on a wide variety of studies, including one which investigated the levels of stuttering whilst in anxiety provoking VR environments [257]. Although the Q sensor is no longer available on the market, Affectiva has designed and developed software solutions for affect detection, offering a software development kit (SDK) for developers using the Unity3D game engine [258].

For valence detection in VR, researchers and developers can utilise technologies that incorporate electroencephalography (EEG) sensors and electromyography (EMG) sensors. Generally, the number of portable technologies using EEG is higher at the moment than EMG for gaming and VR purposes. EEG readings are extremely sensitive and can be affected by the user's movement. However, the majority of studies using EEG are stationary, not leveraging the full potential of VR for spatial interaction and freedom of movement.

A recent study aiming to assess emotional responses induced in virtual reality found statistically significant correlations between the reported valence and arousal picture ratings and the EEG bands outputted from the Emotiv EPOC+ 14 channel EEG headset [259]. The system is light and easy to use, involving a short preparation of hydration of the sensors before usage. A limitation when using this headset alongside the HTC Vive VR system is the difficulty of ensuring precise localization of the electrodes which can increase variability of readings between participants but also between sessions of the same participant. Therefore, the Emotiv EPOC+ should be used in the correct context to ensure accurate affect detection. Similarly, the Neurable headset combines EEG sensors with the HTC Vive [193] HMD to ensure consistent localization, allowing user intent to be detected and used as interaction input in virtual environments [260], [261]. Further to this, Neurable have also developed an SDK for Unity 3D for developers [262]. Combining the SDK with the ability to measure gamma waves means there is potential for real-time affect detection in VR, as it has been found that gamma waves correlate with emotionality [263]. In 2016, a study examining the effect of body ownership in virtual reality

using a different EEG sensor technology, Enobio (32 sensor set-up) noted that both augmented and virtual reality produce higher brain activity in beta and gamma waves than when present in the real world, which is something to consider when using EEG sensors in Virtual Reality research [264]. Another technology that came out in 2018 is the LooxidLabs headset, which combines 9 dry EEG electrodes and built-in eye tracking cameras into their own VR HMD. Unfortunately, we have little evidence of the system's accuracy of detecting affective states as it has not yet been used in an emotion related VR research study.

Currently, emotional valence is difficult to measure in room-scale VR (non-seated experience) and the current EEG approaches may add additional movement constrains to the user. The method of measuring electromyographic signals (EMG) from the face of the user in VR could give us a reliable indication of their affective state [128]. In this context, another recent example of multimodal affect detection technology is the EmteqVR interface (including formerly know sensor-mask as 'Faceteq'), whereby EMG and PPG sensors are embedded on a foam VR insert, allowing its use on commercial head-mounted displays (HMDs). Studies investigating the detection of valence and arousal using this device have shown promising results [265], [266]. EmteqVR and the aforementioned technologies could be improved further by the addition of eye motion tracking, to monitor the individual's gaze while in the virtual environment, thus, allowing a fully rounded analysis of the individual's affective state when experiencing an emotional stimulus.

All the technologies presented, showcase the growing need for multimodal signal analysis to understand the user's emotional state in VR. As sensors become smaller and easier to integrate, we expect a rapid growth of affect-detecting technologies in the next years. The importance of heir unobtrusive wear-ability and usability in VR is paramount for VR research, as low levels of immersion and presence are correlated with hardware related distracting factors and reduced freedom of movement [267]. Ideally, researchers and developers in VR would benefit from the combination of metrics for simultaneous arousal and valence recognition, in user-centred hardware approaches that promote free movement and easy integration with HMDs.

Table 1. List of commercial systems for physiological and affective sensing and their compatibility with VR technologies. The table is divided into two sections, one for the arousal related methods and one for the valence related ones. ('HMD comb.': Compatibility with HMD headsets, 'VR cond.': VR conditions tested, 'e.o.u.' : ease of use, 'n.t.': not tested).

Method/Physiological sensor	Body area	Commercial system	HMD Comb.	VR system	VR Condition	Room scale (e.o.u)
Arousal						
Photoplethysmography (PPG)	Wrist, palm	E4[268]	Y Y	- -	Seated [269]	n.t (high)
Electrodermal activity (EDA)	Wrist, palm	Q sensor[256] E4[268]	Y Y	- -	Seated [270]	n.t (high)
Valence						
Electroencephalography (EEG)	Head	Emotiv EPOC [259] Neurable [254] Enobio [253] Looxidlabs [252]	N Y Y Y	- HTC-Vive - Own system	Seated [259], [260], [264]	n.t. (low*)
Electromyography (EMG)	Face	Faceteq/EmteqPro [271], [272]	Y	HTC-Vive		To be tested
*high sensitivity to motion artifacts						

2.4. Limitations and considerations on current affect detection approaches in VR settings

[Extending to *Context*: Subject-Medium-Object]

For this project, it was important to explore:(1) ways to induce affect in VR settings, (2) investigate the adequate methodologies for affect assessment assessed during a VR experience (subjective and objective), and (3) identify the state-of-the art classification techniques for affect detection (4), while attempting to ensure high presence and naturalistic behaviour. In this section we will outline relevant affect induction and detection studies using virtual reality technologies and express our considerations for the future emotion studies using interactive VR experiences with HMDs.

2.4.1. Affect Induction in VR settings

In section 2.3.2, we presented various studies in which VR environments were used as an emotion induction tool. Just like videos exceeded static images, immersive interactive experiences can exceed conventional audio-visual stimuli. Due to this highly immersive nature, virtual reality is an active method of stimulation which when coupled with high levels of presence can potentially induce more realistic emotional responses than conventional passive methods [170], [171], [273]. The methodological challenge for using VR settings is that the induction of naturalistic emotional responses to stimuli, and the related activation of an enhanced level of presence, requires careful experimental design and control. However, the generation of virtual immersive stimuli for affect induction is a relatively new area for exploration and we are aware of only one database so far that has been developed. This database, created by the team of [274] consists of immersive (360° degrees) videos for VR with corresponding arousal and valence ratings. With a good variety in video situations, easy replicability, and with the corresponding affective ratings, such databases can work as an excellent assistive tool. However, as mentioned before, attaining high levels of presence require high-interactivity options which are not covered by a standard, passive 360° video perspective. In fact, 360 videos may be more immersive and generate presence, but their emotion stimulation effect is

closer to videos than interactive, close-to-real-life experiences. Therefore, developing and using a validated database consisting of 3-dimensional, interactive virtual environments could potentially yield the true potential of virtual reality as an emotion induction tool. Developing such a stimuli library could facilitate research from different fields and little experience with VR development to use ready-made content in their research with little effort.

Regardless of the available stimuli databases, the effect of virtual reality for emotion induction has been mostly explored for two main purposes: stress/arousal induction and stress reduction or meditation. Virtual environments as a psychological research tool have been used in conjunction with physiological responses acquisition in various application areas. Recent studies in anxiety and VR research showed that VR can be used in exposure therapy (VRET) to induce relaxation and to reduce anxiety [27], to distract users from pain (see review by [275]) and for stress reduction in clinical contexts [276], [277]. VR scenarios were also used to induce negative emotions and stress (negative arousal), for example in public speaking scenarios [278] and by introducing phobic elements into the VE [279]. Presence levels were reported higher in stressful VEs than neutral VEs scenarios agreeing with the existing literature [218], [280]. Other examples include phobia therapy [281], meditation and relaxation [282], [283], training and exergames ([25], [284]–[286]), spectrum disorders [287], [288], mental health therapies [289], computer games [290], presence and aesthetic experiential research [171], [291]. In the majority of those studies, the spectrum of arousal and valence is skewed towards either highly arousing negative experiences or low arousing of neutral/positive valence. The identification of a generalizable stimuli-presentation protocol that can induce similar positive responses across individuals is challenging, as it is difficult to detect positive responses without knowing the context [5]. High positive responses are much dependent on the context and the subjective appraisal of the stimulus (e.g. non-life-threating situation, rewarding) [292], which can have a different effect on individuals based on their previous experiences, cultural background, personality, expressivity, and sense of humour. Overall, the exploration of the full spectrum of the affective dimensions and the balanced focus on positive high-arousing experiences (e.g. generating laughter) as well as negative is scarce [293].

Arousal and valence affect induction research involving 3-dimensional stimuli and spatial interactive elements in VR, is limited yet very promising. One of the most

recent projects utilising 3-dimensitonal immersive content (3D spatial structure, as static empty rooms) [294] where EEG and heartbeat dynamics were employed to detect three affective states in VR. The authors collected physiological signals and SAM scores per scenario of which visual aesthetic parameters (changes in lighting, colour, and textures) were altered to induce a certain affective state (negative, positive, neutral). Analysis and binary classification of arousal and valence were made for each one of the virtual environments, which showed promising results for the capability of immersive VEs to induce affective states. As the effects of each VE were analysed as a whole, the individual effects of the individual aesthetics properties used in each one of the environments were not tested. The effects of geometrical and aesthetic properties of virtual environments on affective states (mostly for valence and arousal) were tested against post-experience self-ratings in previous studies.

In other studies, certain audio-visual parameters were altered to study emotional arousal and relaxation using different version of a virtual park [295], [296] and valence properties of aesthetic parameters like colour and room length-width-height ratio in indoor virtual spaces [297]–[299]. Results showed promising correlations of controlling parameters with certain affective values (e.g., ceiling height, number of windows, enclosure/room openness, brightness, colour hues) which could be used for the development of VR stimuli in future studies. However, the effects of these parameters on physiological measures (with the exception of room openness) were not tested in these studies. The effects of room enclosure on heart-rate responses was tested in a virtual Trier Social Stress Test (commonly referred to as VR-TSST), but however not significant differences were found between the groups who performed the task in the open space compared to the ones in the enclosed space [300]. On this note, these spatial manipulations alone may be insufficient to induce a wide range of spontaneous emotional responses along the dimensions of affect.

In a realistic simulation, such parameters may change over time, or after certain user's actions (moving from one room to another, closing light, interacting with an 3D object etc.). As such, the user interaction in Virtual reality with potentially stimulating content is predominantly non-linear and dynamic. The effects of those audio-visual surrounding parameters of the virtual experience on affective responses could be studied by analysing the moment-by-moment changes of the content/stimuli in conjunction to physiological changes. As a VR user can move their

view with a minimum of three DOF when using a VR headset, the continuous tracking of the user's gaze or eye view (for instance via eye tracking methods) to the environmental properties would be greatly beneficial. Such tracking could allow us to correlate contextual information of the events/tasks happening in VR to the user's physiological responses. Additionally, benefitting from latest VR technological proliferations, allows for active experiences, enhanced graphic representations, the addition of interactive stimuli and enhanced room-scale user movements. These improved properties of the platforms could potentially enhance the user experience, and result into stronger emotional elicitation and generally more naturalistic responses [301].

The incorporation of additional modalities in VR was limited until recently, due to the constantly iterative nature of these emerging VR technologies [302]. Today, additional measures like pupil diameter (detected with eye tracking technologies especially adapted for VR settings) and speed of limb movement (measured with from motion tracking sensors) are also starting to emerge as measures in studies of affect detection [303], [304], and are expected to be very promising for the future of emotion detect in VR.

2.4.2.   Spontaneous Affect classification studies using VR

Currently, only a small number of studies utilise immersive VR as an affect induction tool. Proportionally, there is a scarce number of studies combining continuous physiological data acquisition within a VR experience with Machine Learning (ML) for the automatic detection of valence and arousal. The majority of past research on spontaneous affect detection used other types of stimuli presentation, either in the form of static images, text, audio and videos [68], [305]–[307]. From those studies, when utilising physiological data to detect dimensional affect, the most common strategy found is the simplification of the dimensional space to dimensional ranges of interest, thus streamlining down to a binary or three-class problem [308], [309]. In the case of valence, the classes commonly detected are 'negative' versus 'positive', and in some cases with an additional class 'neutral' (e.g. [307], [310]). For arousal, the dimension is usually divided in 'low/passive' versus 'high/active' areas (e.g. [311]), with the introduction of a 'middle' area (e.g. [312]). In cases were both dimensions are detected simultaneously, four classes are detected corresponding to the quadrants of the AV space (e.g. [313]).

Today, there is an increased interest on spontaneous affect detection methodologies from continuous physiological signals using virtual environments and immersive VR devices. In the last years alone, big technological advances were accompanied by advances on affect detection methodologies developed specifically for immersive VR settings using HMDs. In 2010, Wu and colleagues utilised skin conductance, respiration , ECG and EEG signals features to classify three levels of arousal in a VR Stroop Task [314]. A support vector machine (SVM) classifier was used. The paper does not define whether a highly immersive set-up, as a virtual reality headset or cave system, was used in the study. It is expected that a multi-monitor set-up was used instead to allow the positioning of the EEG sensors which by design pose physical constraints. Shumailov and Gunes in 2017 classified two levels of arousal (low/high) and valence (negative/positive) from an EMG (positioned on the arm) during a VR gaming condition using a commercial VR HMD (HTC Vive) [315]. The paper demonstrated high accuracy in valence and affect detection (> 85% for both dimensions) with a SVM classifier from arm muscle patterns using data from eight participants. This approach applies to immersive VR gaming, where users are actively interacting with the content using hand controllers. Interactive VR applications may however utilise other types of tracking methods e.g. gaze or eye tracking [316][317]. Similarly, in 2019 Pinto and colleagues [318] combined physiological readings from the hand and chest (BVP, EDA, ECG and respiration) and an SVM classifier to detect low and high levels of arousal and valence across individuals. The teams utilised seven 360º videos (from the first VR video library created by [274], viewed in VR HMD) to induce different levels of affect and collected post-experience SAM ratings, which were used for the labelling of the classes (using the median as the class divisor: 1-5 for negative, 6-9 for positive). The system yielded 58% accuracy for arousal and 57% for valence, with accuracies across models being higher for arousal than valence similar to e.g. [319]. More recently, Marin-Morales and colleagues [294] utilised 3D immersive environments (viewed using a commercial VR headset) in conjunction to EEG and ECG sensors to detect two levels of arousal and valence using SVM algorithms. Their models reported achieved 75% accuracy for arousal and 71% for valence across participants.

Table 2. Table listing studies on affect detection performed using physiological sensors in VR settings ('Condit.': Recording Condition).

Physiological sensor	Detection (levels)	Stimuli	Labels	VR system	Study	Sample	Condit.
GSR,RSP, ECG, EEG	Arousal (3)	VR stroop task	Performa-nce scores	Not known	Wu et al., 2010 [314]	18	Seated
Myo[320] EMG (both arms)	Arousal (2) Valence (2)	VR game	Post-gaming annotation	HTC Vive	Shumailov and Gunes, 2017 [315]	8	Standing
BVP, EDA, ECG RSP (hand, chest)	Arousal (2) Valence (2)	360° videos [274] + YouTube videos	SAM (post-video)	HP Windows Mixed-Reality	Pinto et al., 2019 [318]	18	Seated
EEG, ECG	Arousal (2) Valence (2)	Four 360° panoramas	SAM (post-VE)	Samsung Gear VR	Marín-Morales et al., 2018 [294]	36	Seated

The trends on emotion classification mention the necessity for the development of models that can apply to real-life data by collecting naturalistic-spontaneous responses [10]. However, the data recording settings commonly arranged for the affect detection VR studies described are closer to controlled, *laboratory* settings rather than *real-world* ones [321]. This is mainly because the user's body movement and location/seating were constricted, and that the participants are requested to perform a set of predefined tasks/ actions, rather than left to 'freely' explore and interact with the content. This may also be explained due to the non-interactive stimuli/content used, such as passive videos and non-interactive spatial elements. As Virtual Reality allows for naturalistic behaviour, free head/body movement and interaction, as well as adaptable task/scenarios, data recording settings within virtual reality settings could potentially approach 'real-word' settings. As such, virtual reality settings could offer the medium between strict laboratory and challenging real-word settings.

Affect classification studies often mention the high variability between individuals' ratings (e.g. [318]). These may be related to effects of external factors (e.g. the environment, the technology involved, the stimuli presentation interface, DOF), while also on the perception ambiguity of the content/stimuli, which may be attributed to various subjective differences (such as difficulty in self-rating perceived emotions, habituation/desensitisation and cultural differences) [75], [322], [323].

One may suggest that difference in affect ratings for example for a long VR video, may be deriving from smaller events or details that occurred that may have generated different affective qualities for each individual. As a result, these secondary details may be experienced and remembered in different intensities by the individuals (see section 2.3.3.2).

Still, the most common ground-truth used in past research for the data annotation/labelling is driven by affect self-ratings after the main experience (post-experience) [301]. With the introduction of tools such as the Feeltrace annotation tool [324] affectRank [325], Gtrace [326] and the Continuous Measurement System (CMS) [327], continuous self-ratings could be recorded throughout an experience. Continuous self-ratings made during the VR experience could offer a higher resolution on the changes along the affective dimensions which could contribute to the development of finer-grained classification models for affect detection.

A current disadvantage in emotion detection VR research, is the lack of common methodology, and thus reduced ability for generalisation and repeatability. Special focus could be on listing potentially influencing factors. For example, the different types of media experienced and the degrees of motion freedom (e.g., lying, standing, walking) may naturally have a different effect on physiological data recorded (see section 2.3.2). As such, different classification approaches may be more suitable for different conditions.

In summary, despite the recent progress in the field, it appears that state of the art in affect detection using VR content is still limited to:

1. Intrusive physiological sensor modalities that constrict use motion.
2. Lack of spatial stimuli datasets, that could be used in an interactive set-up.
3. Affective classifications based on post-experience self-ratings of an environment rather than dynamic changes of continuous in-VE ratings.
4. Limited contextual information regarding the user's characteristics, external factors, and in-depth annotation of the VR-content's qualities.
5. Use of VR for passive stimulation, not utilising full potential of current VR systems (e.g., interactivity, dynamic changes, room-scale locomotion).

2.5.   **Chapter Discussion and Conclusions**

In summary, the area of affect induction and detection in VR is in early stages although there is wealth of potential applications including in a wide range of fields. The dimension of arousal has been extensively investigated with physiological measures in VR and other interactive media. However, for the spherical assessment of affect, valence evaluation should be equally measured with continuous physiological or behavioural measures. Current practical and technological challenges due to the obstruction of the face (the richest source for valence sensing), underlined the need for an adapted approach towards affective modelling in high-fidelity VR settings. Specifically, these challenges are triggered by commonly use of VR headsets and the motion difficulties in existing cumbersome, tethered sensor devices, designed for sedentary/seated conditions.

The combination of the detection methodologies for both affective dimensions in one sensor set-up could provide the solution for affective recognition in VR settings. Until recently, the practical implications of HMD-based VR did not allow the use conventional valence measures, as used in previous lab-based experiments (e.g., wet facial EMG, cameras, EEG sensors), as the face and head were covered by the VR headset and its head-strap (see the area covered by the VR headset in **Figure 11**). In addition, VR headsets provide high-resolution head motion-to-vision mapping, which aspires for naturalistic behaviour and interactions as those in the real world. However, in a classic lab-based sensor set-up such whole body movements are avoided or minimised. The level of ecological validity of VR, together with high levels of subjective presence and immersion are suggested to produce naturalistic responses. Thus, designing a sensor set-up that could be applied in interactive VR settings, could also open avenues for physiological and affective sensing in real-world scenarios without the use VR headsets (e.g., in free walking and head-moving physical tasks, or even during social interactions).

Figure 11. Two photos of users wearing two different consumer VR headsets being. The majority of the face is covered by the headset, while the bottom of the face is shadowed. The head strap and headphones also cover the top and side of the head.

Today, immersive technologies provide multiple input and output tracking capabilities, which combined with the virtual experiences simulating reality, offer a powerful tool for researchers. The data recording settings in modern high-immersive, interactive, room scale VR would potentially exceed the controlled laboratory settings (commonly used in affect recognition studies) by simulating environments that are closer to the real-world. This in turn could induce more naturalistic and spontaneous responses, thus providing rich affect data and improving the current affective modelling/recognition approaches (see section 2.3.3).

.

To leverage the potential of current commercial VR technologies (e.g., for physical movement and interactivity), one also needs to take into consideration the unnecessary distraction and movement constrictions that tethered and chunky physiological sensors can introduce, as well as related movement artefacts that can result in the data. Today, more and more emerging HMD-adapted technologies employ physiological sensors. However, the Faceteq prototype is one those emerging technologies which could be applied in highly immersive settings without hindering the interactive experience of the user. As our team is looking at ways to provide affect-related metrics into one hardware interface especially adapted for VR, it is our aim to combine arousal and valence detection according to the dimensional model of affect, using multimodal approaches. Therefore, for this research project, focused on facial muscle activation for valence (using a configuration consisting of

multiple EMG sensors underneath the HMD), in addition to arousal measures as PPG and ECG.

In the next chapter, Chapter 3, we present the affect detection system that we propose, along with more details regarding the methodology used in three experiments.

Chapter 3

Methodology & System Architecture

3.1. Introduction

The research questions of this thesis were investigated following an experimental approach combined with quantitative[2] data acquisition techniques across various data-collection modalities. The combined multimodality aspect together with the controlled experimental approach allowed for validation and reinforcement of any results. Based on our review, we designed a system architecture for affect detection within virtual reality experiences that is illustrated as a graph model. In this chapter we will outline this proposed system architecture, and discuss the methods, techniques and apparatus used in the experiments discussed in chapters 4 and 5.

The system architecture's key components comprise of the input data, the main processing engine including the mapping of data to specific custom affective labels (e.g., positive or negative valence), and the detection or prediction output. These components will be explained in the following sections (3.2 – 3.5). All 'user-dependent' data types (data obtained from the user) including physiological, behavioural, and subjective data (self-reports and questionnaires) are described in section 3.3 – 3.4. Additionally, as VR offers the freedom to coordinate and track the interaction between the user immersed in VR and the content, the suggested paradigm for the presentation and control over stimuli's effect in VR will be explained in section 3.5. In that section we will also discuss the acquisition of additional contextual information that could potentially assist future content-

---

2 Quantitative approaches were preferred over qualitative for comparative research, due to their ability to define cause-effect relationships, and strengths of associations between numerical dependent variables, while maintaining validity and reliability in the interpretation of the responses ([559], [560]).

dependent affective studies using custom-developed content. In section 3.6 we discuss the data processing pipeline and analysis methods used. The processed data combined with self-ratings of affect are used for the mapping and classification of the participants' affective responses to different levels of valence and arousal.

The methodological tools and protocols shared between the feasibility experiments were progressively refined based on observations and results from each experiment. The specifics on the methodological changes will be discussed in the *Methodology* sections of each study in Chapters 4 to 6.

3.2. **System Architecture**

In brief, the system architecture suggested for this project was based on the distinction of input data streams associated with either valence or arousal, based on the concept that both affective dimensions can deliver a better understanding of the affective state of an individual. Data modalities linked to valence such as electromyography and movement sensors, and those linked to arousal such as heart rate and electrodermal activity [93], [328] were combined in one data stream and were recorded synchronously.

Additionally, VR-specific settings needed to be considered for the unobtrusive acquisition of those signals. The majority of interactive VR experiences nowadays, either seated or standing, require the user to be able to freely move the limbs and/or body to interact with the content. Example types of those can be exploratory, educational, gaming, social, creative etc. Therefore one major goal for us was to utilise a variety of biometric signals while keeping a simple, unobtrusive set-up, thus reducing the numbers of sensors placed on limbs [328]–[330]. Further, in order to reduce the volume of cables surrounding the user (which can distract and create breaks in Presence [331], [332]), we utilised a wearable multimodal interface prototype, called Faceteq by Emteq Ltd (especially designed for use in VR). This prototype design allowed us to append additional sensor modalities and test them on the go in collaboration with the development team of the company [4], [329] (Figure 12). Faceteq, comprised of EMG sensors positioned based on the protocol proposed by Van Boxtel, 2010 [122]. The prototype we first used was published and presented in 2016 [272].

From 2015 to 2017, we utilised the Faceteq interface in various prototype versions including custom GSR, ECG and PPG sensors incorporated on the

interface. Through extensive testing, we found that this interface offered promising potential for unobtrusive affect detection in VR. Certain aspects of the interface such as the materials used, the number of sensors and the data collection paradigm were defined in the duration of this EngD. As part of the EngD programme, the findings originating from extensive pilots and studies assisted the further development of the interface (hardware and software). The outcome of these experiments rendered specific recommendations for:

(a) shape and elasticity of the wearable interface,

(b) width and type of foam utilised in its construction,

(c) number, type and positioning of sensors utilised,

(d) the development of a protocol pipeline for new users on how to utilise the interface, apply sensors and record data,

(e) the implementation of an event-marker based data collection system, and

(f) the development of a dedicated affect detection algorithm for offline analysis.

Figure 12. Prototype version of Faceteq interface in 2016 and 2019.

For the proposed affect detection system entailed an input data stream provided from wearables interfaces (such as Faceteq) and the successful mapping of their activation patterns to the two dimensions of affect. arousal and valence (core affect dimensions [64], [333]). Following this principle, the suggested system architecture is presented in Figure 13. There are three streams of input data; a) continuous streams from physiological and movement sensors, b) the user-specific traits recorded via self-ratings and questionnaires, and c) streams related to the stimuli content within a simulation environment (boxes numbered (1) and (2) in left upper and lower side of the figure).

The continuous streams of raw data are recorded while participants are exposed to stimuli with affective content, and during that time they are also self-rating their levels of valence and arousal on a continuous scale (bottom orange box including our custom-built continuous affect self-rating application 'CASR'). Collecting the levels of arousal and valence on a continuous scale from the user is

an essential step for acquiring information on the levels of emotional stimulation, or else 'emotional flow' of an experience [334]. These data can also be used in order to verify the content's affective impact and therefore the success of the experimental design in attaining the intended emotional manipulations.

Once the multimodal data-streams from our sensors are collected (predominately EMG and PPG, movement specific metrics were added later on) they become subject to signal processing (schema 'Data Processing' No.3). These physiological signals are denoised and divided into epochs (e.g 512 seconds was used for Feasibility Study 2, Chapter 4). Next, various components or features of the signals are extracted through the 'feature extraction' step. From those components (e.g., the calculation of the variability of Inter-Beat-Intervals (IBI) or peaks distances, and the room mean square (RMS) of the f-EMG signal) we can infer the level of arousal/stress felt and valence expressed by the user. Detailed lists of the features and specific data analysis processes are included in the experimental design description of each study in Chapters 4 and 6. The features together with the participants' self-ratings are then sent to train a classifier to categorize the physiological patterns and affective responses into arousal and valence levels. Finally, the outputs of the classifier can then be used to produce a two-dimensional
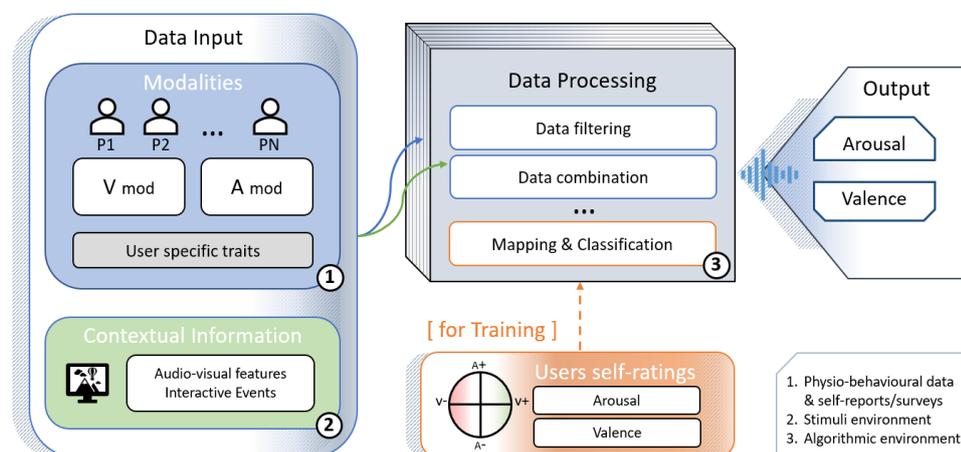
Figure 13. Flowchart of the proposed system architecture for affect detection in VR. Physiological data streams are entered (1, left) together with contextual information of the experience (2). The data are then processed through filtering and normalisation before being fed to a classifier (3, centre up) Using the user's self- ratings (centre bottom), the classifier model is trained. This model can be used to detect arousal and valence from new data (output, right).

representation of the user's state, which can be used for adaptive control and feedback within VR.

3.2.1.   Feasibility studies

In order to evaluate the system's architecture for off-line analysis, we designed a series of feasibility studies with human participants (Feasibility studies for the detection of emotional responses (FEDEM)). In the studies, physiological responses together with continuous self-ratings from each participant were collected initially using conventional content-presenting technologies and later on using virtual environments, see Figure 14.

For the first study our plan was to record affective responses in controlled conditions using already validated stimuli, with minimised head movements and without a VR headset. The stimuli were selected to induce five affective states, which populate the four quadrants of the affective space plus the neutral (high valence – high arousal (HVHA), high valence -low arousal (HVLA), low (negative) valence-high arousal (LVHA), low (negative) valence – low arousal (LVLA) and neutral – medium arousal (NMA)).

From then we moved to the integration of supplementary modalities and the recruitment of additional participants. Finally, after performing adjustments on the main apparatus and the processing pipeline we designed the next study (Fedem 3) using virtual stimuli in custom-built 3-D dimensional environments using commercial VR headsets. In this study the participants could either watch videos of pre-recorded explorations from matched participants (passive group), or actively explore three virtual environment (VE) scenarios (active group): three different versions of the same environment containing stimuli of positive, neutral and negative valence and of various levels of arousal. In the VE scenarios, stimuli were presented as 'events' were predefined although triggered according to the user's interaction (gaze) and time spent in the environment.

The detailed experimental protocol and information regarding the stimuli used for each study are presented in Chapter 4 (for studies Fedem 1 & 2) and Chapter 5-6 (for study VR study Fedem 3).
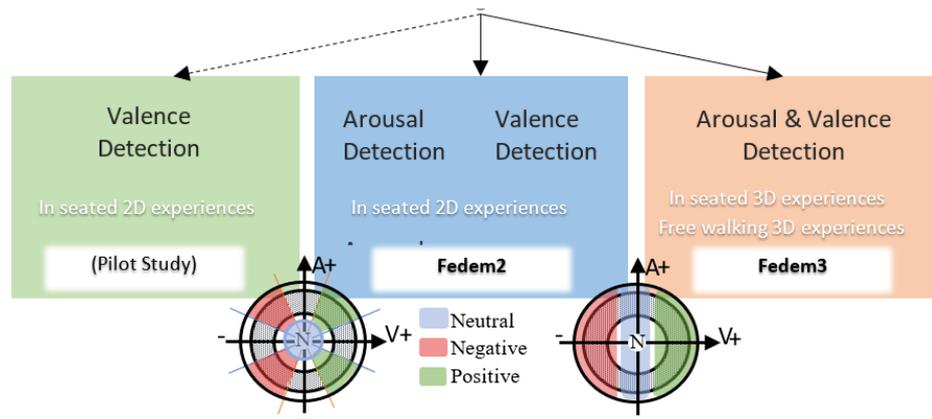
Figure 14. Feasibility studies designed to detect arousal and valence levels using affective videos and virtual stimuli. For each study the affective stimuli were chosen from the corresponding-coloured areas on the VA space (bottom left and right); blue are stands for neutral, red for negative and green for positive valence. Three valenced VR scenes were designed for the Fedem3 study, containing stimuli/events of mixed levels of arousal.

3.3. Discrete & Continuous Self-Rating Data

Subjective data, including questionnaires and interviews, are a fast and low-cost data-acquisition method. In this thesis, questionnaires were administered in all experiments conducted. We were interested in the individual differences between participants (e.g., alexithymia, expressivity levels) that could have an effect on the responses of each participant to the stimuli used in our studies.

The subjective data sources can be divided into two categories: the 'discrete responses' collected per participant using questionnaires, and the continuous affective self-rating (CASR) data which includes Valence and Arousal ratings collected throughout similar emotion detection studies, like in [324], [335].

3.3.1. Discrete Responses

As such, in the first category we included questionnaires on demographics, SAM mannequins and screening questions, as well as personality traits, alexithymia, and expressivity as they can correlate with changes in sympathetic activity, emotional expression and inhibition [244], [249], [336]–[341].

The demographic questionnaire included questions on age (as aging can increase expressive control and reduce HRV ([342]–[345]), gender (as sex difference were found on emotional expressivity [346], [347]) and experience with immersive technologies (see Appendix D). Additional screening questions were added in the Feasibility study 3 study regarding the existence of phobias (related to the stimuli, for instance arachnophobia), and extreme neural, medical, cardiovascular, mental and psychological conditions which could affect our data. Examples of those where given, e.g., facial paralysis, stroke, aortic aneurism, anxiety, depression. The participants were recommended to discuss this further with the experimenter if they were unsure of their response.

For the personality we administered the OCEAN (Big 5) questionnaire consisting of multiple questions calculating 5 personality traits of extraversion agreeableness, neuroticism, conscientiousness and openness to experiences [348]. More information will be found in chapter 5.

We also used the Toronto Alexithymia Scale (TAS-20) [339] questionnaire which quantifies one's ability of people to interpret, process and describe the emotions of themselves or of others. High alexithymia can affect emotion regulation of emotion with effects of physiological responses during affective stimulation using audio-visual stimuli (see section 2.3.3.2).

In cases where we needed one rating per stimulus, video or Virtual Environment scenario for valence and one for arousal we used the SAM scales [64], excluding the dimension of control.

Additionally, the Berkeley's questionnaire on Expressivity [249], the Depression Anxiety and Stress scales (DASS) [349], [350] and the Basic Empathy Scale (BES) [351] were added as optional questionnaires in the last study (Study 3).

The responses from all questionnaires were anonymised and were associated to the raw data using numerical codes to allow for the between groups comparisons.

3.3.2. Continuous affect self-rating tool (CASR)

As the affective impact on an experience can vary during its duration, we recorded moment-to-moment ratings of valence and arousal rather than one rating overall. For this purpose, we developed out own version of self-rating tool called 'CASR' (continuous affect self-rating tool) using Unity game engine [352], based on an

existing tool for continuous affect rating 'Feeltrace' [324]. The CASR interface was integrated within our custom-built stimuli presentation software (see Figure 15).
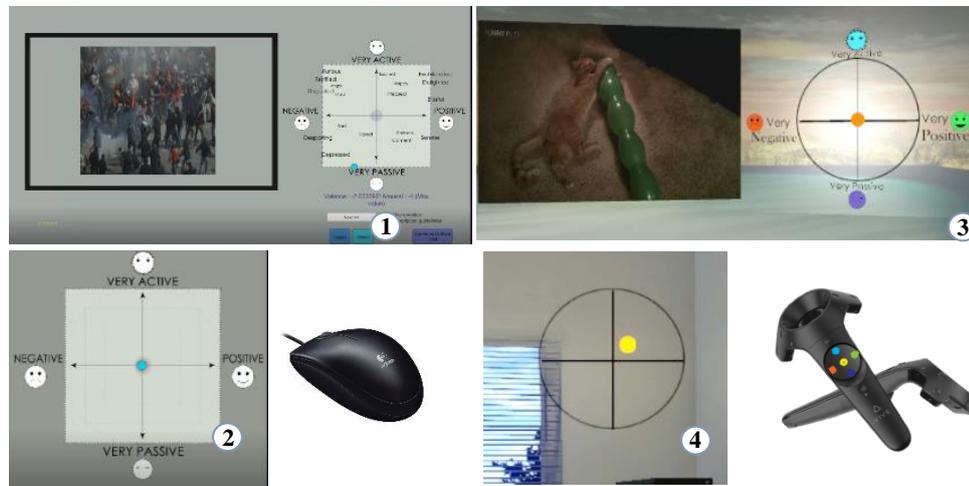
Figure 15. Screenshots from the Continuous Affect Self-Rating (CASR) interface; First version (1, left side) during training and (2) in main study using an optical mouse. Last version (3, right side) in training and (4) during main study using VR controller.

The participants would first get trained into rating their affective state using the mouse. The pointer of the mouse appeared as a blue dot appearing onto the CASR space (with or without emotion labels integrated within the cartesian space, pictures 1 and 3 of Figure 15). Then, within the study they would use the same space by hovering over the regions of interest throughout the experience (pictures 2 and 4 of Figure 15. On the x-axis we added the valence dimension ranging from very-negative ("-1", left) to neutral ("0", middle) and to the very positive ("1", right). Similarly, we added the arousal dimension on y-axis ranging from very low/sleepy ("-1", down) to very aroused/active ("1", up). The users were able to use the in-between spaces, using the entire cartesian space. Space limits were set to control or out of space-border ratings. In the latest version of the interface, the participants used an input wireless controller with an integrated circular trackpad instead of a mouse.

3.4. **Physiological Signals & Additional Modalities**
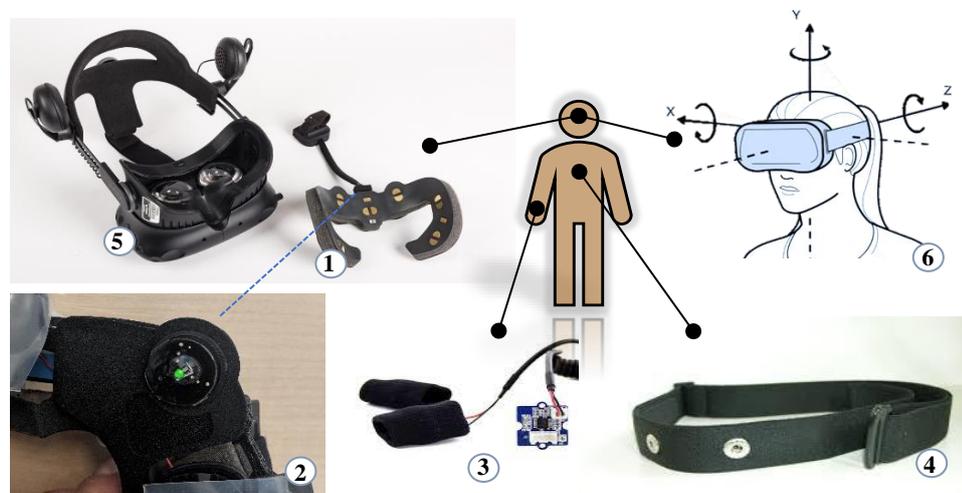
Figure 16. Measures used; (1) the Faceteq mask with EMG and with (2) PPG sensors, (3) external GSR, (4) ECG belt, (5) VR headset from which we extracted (6) IMU data (rotation & position).

The proposed metrics used in this project that can be linked to valence changes are EMG and distance from stimulus (approach/withdrawal), while the ones linked to arousal changes are the ECG, PPG and GSR. The measures used together with the location where they were positioned on each user are depicted on Figure 16.

Utilising the Faceteq prototype [255] gave us the opportunity to use 'dry' EMG sensors, no requiring gel or special skin preparation, fixed onto a mask-frame. This interface (newest version known as EmteqVR) can be adjusted on commercial head-mounted displays (HMDs) like the HTC Vive [193]. It utilises EMG technology for muscle tone detection to determine facial muscle activations. Unlike standard surface EMG electrodes which require skin preparation, conductive gel and adhesive pads, this prototype consists of eight dry integrated EMG sensors (24-bit signal resolution, sample rate: 1000 Hz, signal bandwidth: 20-450Hz). Remarkably, there were no inter-sensor latency variations, thus allowing for simultaneous sampling across sensors. The sensors are individually shielded and connected to an ADC box (Analog-to-Digital Converter) which digitalises and amplifies the sensors' streams (Figure 17, the outputted values correspond to micro-Volts, where typical outputted value looks like this: e.g., 0.0981mV.

Figure 17. The EMG sensors (N = 8) on the Faceteq device are connected to an ADC box. The signals are amplified and streamed simultaneously with a 1000 sampling rate.

Complementary sensors – Additional sensors integrated onto the Faceteq interface are a 9DOF accelerometer-gyroscope and a PPG sensor (positioned on the left temple). The device (the ADC box) is connected to the computer via USB cable although in the last version we used Bluetooth connection instead. This multi-sensor setup can enable the system to record information about the user's movement, gaze, facial tension, and heart rate simultaneously while wearing a VR headset. Another advantage of using this prototype was the existing live data streaming solution to Unity3D [352] (the game engine we used for the development of the stimuli presentation environment/applications used in our experiments) via an API provided by Emteq Labs.

The recordings of raw data were exported and stored in ASCII (text) format for post-acquisition data analysis. The recordings were then analysed and filtered in MATLAB, and the root mean square (RMS) values between time-windows was be calculated for visualisation and further analysis. An example of the EMG trace after transformation into RMS values while a user is performing three facial expressions (3-open smiles, 3-frowns, 3 surprises) is shown on Figure 18.

The choice of the hardware allowed for quick application on the user while also allowing freedom of movement in VR. Especially easy to integrate on wearable solutions are the PPG sensors, which unlike the ECG sensors (both measuring heart-rate activity) can be adjusted on anywhere on the body where blood vessels exist close to the surface of the skin, requiring little to no skin preparation. EMG and PPG readings were provided from the Faceteq Interface. A custom-built ECG belt consisting of two sensors was also used for the first experiment.
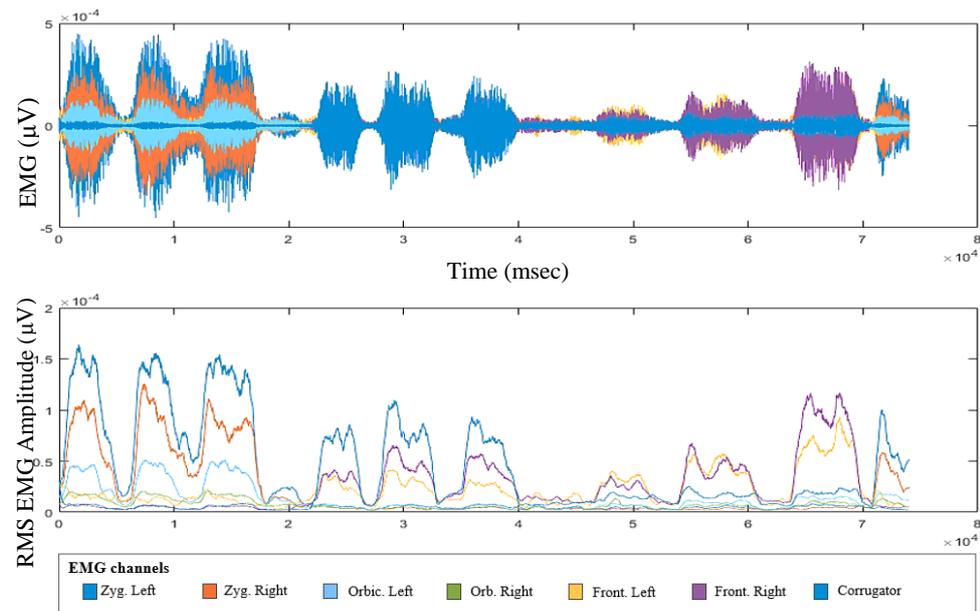
Figure 18. Example of EMG signal traces from a user performing 6 expressions (3 consecutive smiles, 3 frowns, and 3 surprises). Upper figure shows the filtered signals, and the lower figure shown the RMS trace of the signals, across all sensors (*Zyg.:* Zygomaticus, *Orb.:* Orbicularis oculi, *Front.:* Frontalis, and Corrugator).

The ECG design we followed a 2-lead configuration (2 electrodes and 1 ground sensor) using an elastic belt. The belt holding the two electrodes (negative and positive) is positioned symmetrically around the heart, in this case on the upper abdominal area, on the floating ribs on the left and right side of each user, with relative distance between each electrode (approx. 7-20 cm, Figure 19). As ground electrode we used the ground electrode from the Faceteq interface, used for common mode rejection (used to prevent power line noise from interfering with the bio-signals) located on the upper forehead.
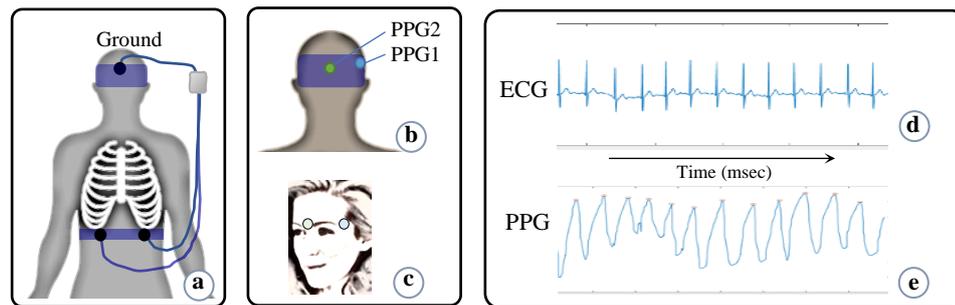
Figure 19. (a) An ECG was applied in a 2-lead configuration using a belt with the ground sensor within the headset. (b) A PPG sensor was intergrated within the headset, and (c) the two positions explored were over the left temple (blue point) and the lower middle of the eyebrows (green point). Graphs (d) and (e) shows the traces of the ECG and the PPG signals. The successive peaks for each steam were used to calculate inter-beat-intervals.

As an additional measure of valence, we measured movement tendencies of approach and withdrawal to/from stimuli the within the virtual environment (for Fedem 3 study). For this purpose, we calculated the distance between the user and an object/event in real-time as the user was walking within the virtual scene (Active group, 'Fedem3' experiment, Chapter 5). The user-object distance was extracted from within the VR simulated experience expressed as 'distance' (distance from point a (position of user in VR) to b (position of stimulus in VR), as in Figure 20).
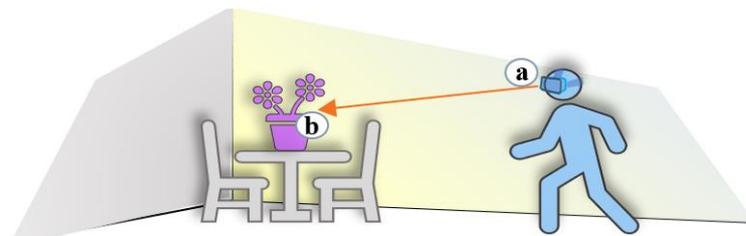
Figure 20. The distance variable was calculated between (a) the virtual viewpoint of the user and (b) an object's center's position (or 'pivot point') in the virtual environment. This metric was added in the last experiment 'Fedem3', where dimensions and user's starting point are constant between scenarios.

3.5. **Stimuli Presentation and Event design in VR**

Although humans perceive and interpret affective stimuli subjectively (e.g. related to their semantic knowledge and associated memories) [353], certain audio-visual parameters have been found to impact on affective stimulation. Such parameters vary

from the low-level features to high-level attributes, dynamic movements and interaction/narrative design [354]–[361]. Low-level features include colour, size, shape, texture, and opacity, while high-level attributes involve quality, composition (level of complexity/simplicity), orientation and functional content parameters such as, lighting and visibility (e.g., contract with background). For the FEDEM 3 study we created the VR environments (VEs) inspired by these parameters, with special focus on dark versus bright lighted compositions (negative-positive), with intense colours and textures for the arousing VEs. These parameters' specific impact within immersive experiences could be investigated via systematic studies in the future. As all constructed content, 360° VR environments provide the 'canvas' for the construction of worlds and scenarios, falling under complete artistic and engineering control. Investigating the composition elements that create emotional experiences could assist all future VR applications but was outside the scope of this EngD work.
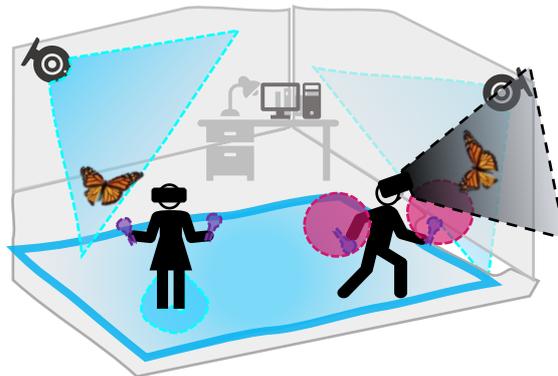
Figure 21. Four models of tracking using VR: Experience tracked outside ((1) cyan shapes represent the motion tracking, and (2) magenta shapes represent the action tracking using input devices such as controllers) and within VR (3) by tracking the view of the user (black shape) and (4) the individual objects information of the surrounding virtual environment (e.g. the butterfly appearing), enabling the flow of contextual information (e.g. he smiled - because he saw a butterfly).

In interactive 360 experiences, the view of the user is not following a strict linear-narrative (e.g. seeing event 1, then event 2 etc.), but can vary based on the spontaneous action-motion of the user (user-controlled exploration) [362], [363]. In order to know which stimulus was visible to the user at any point throughout the experience, we need to know what the user is seeing in real-time. This can be achieved by tracking the gaze of the user through their point of view (commonly using a technique known as 'ray casting') and through 'eye-tracking' via eye-tracking technologies, e.g. [362], [363], see also Figure 21.

In the experimental paradigm developed for FEDEM 3 (VR study), we developed an event-marker algorithm (using Unity3D [352] and the HTC VIVE input utility plugin [364]) (explained in Section 5.3.1). Practically, the algorithm is tracking the gaze of the user continuously (via ray-casting) and adds event-markers in the data corresponding to the visible stimuli within the field of view of the user (Figure 22). Using this algorithm, a VE designer can link objects or actions to specific event-markers. These event markers are continuously tracked and save in ASCII format so that they can be inspected at the end of the experience. These markers were synchronised via system-time timestamps with the rest of the collected data adapted from the EmteqVR application.

Figure 22. Integrating a custom-built event-marker system for VEs.

This way, not only we can track the users' movement (via camera tracking), or their actions (via their controllers' gestures), but also the contextual information within a VR experience. Without this algorithm, it would be difficult and time-consuming to know what the person reacted to and which event applied the most affective impact, as it would require to annotate numerous screen-recordings. Coupling the information regarding the content with emotional responses could allow us to better understand the context of an affective experience.

3.6.   Experiment Variables & Data processing protocol

Chapter 4 shows a feasibility study for the inference of affective responses with a system ready to be incorporated in VR setting. In Chapter 6 we describe the experiment using VR technologies designed based on and the findings reported in Chapter 4 and 5. Both experimental designs allowed us to investigate the effect of affective changes (cause by emotion induction via audio-visual stimuli) on the metrics (dependent variables) obtained by our sensors, and the sensors' ability to detect those changes in VR settings without impacting heavily on the overall immersive experience of each individual (levels of presence).

3.6.1.   Variables and study designs

As independent we define the variables that can affect and cause changes on the dependent variables across participants. For our designs we used audio-visual stimuli (pictures, videos and 3-D spaces/scenes) as independent variables, which we manipulated to create different affective conditions; either following the quadrant model of affect or following a 3-level valence/arousal model with (shown in Figure 14). The dependent variables (also described as input data in the system's architecture section) were recorded using continuous self-ratings, post-experiential surveys and physiological data. The values from those dependent variables were then analysed in conjunction with explanatory variables, in order to explore effects due to the impact of the independent variables.

To account for individual differences and the effects of certain variables on the dependent variables with little or no-relation to the independent variables effect, we recorded additional explanatory variables. As such, those variables were recorded in the form of a demographic pre-acquisition questionnaire, the personality traits of the participant, alexithymia levels, expressivity levels and an objective baseline measure of the participants physiological metrics.

In the first preliminary experiments, our goal was to measure changes between affective states induced by stimuli across participants. We therefore followed a repeated measures within-group design with various stimuli (conditions), using 2 factors (Arousal and Valence) with 2 or more levels (high/ low or positive/negative), following a 2x2 factorial design (see Table 3). Thus, each

participant experienced all conditions in a randomised order. This design is very common in other affect-detection studies [69], [365]–[367]. The main advantage of this design is even a small sample size of participants can yield rich data for all conditions, that could potentially divulge a more coherent narrative. A possible disadvantage is that the participant may develop expectations between conditions and that the effects of each condition may be carried over on the next one (especially for the carry-over effect of negative stimulation) [368] if an adequate break between condition is not considered. In our designs, breaks and neutral stimuli were used between conditions.

Table 3. 2x2 Factorial design user for Feasibility study 2 using video stimuli as conditions.

	Group A - Factors	
	Positive Valence	Negative Valence
Low Arousal	Condition 1	Condition 2
High Arousal	Condition 3	Condition 4

The last experiment described in this thesis (Fedem 3 VR study) was designed based on a between-group and a within-group analysis approach. In the within-one-group design, data collected from our conditions (affective VEs) were analysed to determine the effect of the independent variables on the dependent variables (valence and arousal ratings and physiological measures). Yet, we added the independent factor of interactivity, by having two independent groups in total; one that experienced all the conditions (randomised) actively (Active group) and one passively, vicariously (Passive group), see

Table **4**. Subsequently between groups comparisons were explored.

Table 4. Independent group factorial design used for Feasibility study 3 using VE as conditions.

	Group Active			Group Passive		
	Negative	Neutral	Positive	Negative	Neutral	Positive
Arousal	Condition 1	Condition 2	Condition 3	Condition 1	Condition 2	Condition 3

3.6.2. User-dependent and User-independent classification approaches

Both a user-dependent approach and user-independent analysis approach was conducted in our experiments.

User-dependent design allows for comparative research across conditions for each participant separately using the user's self-ratings as ground-truth. The main advantage of this approach is that since an emotional experience is highly subjective, we can observe changes on the dependent variables which are characterising a particular user, allowing for the development of personalised models. User-dependent affect classification models are expected to perform more efficiently than user-independent models [323] but they have low potential to generalise to new users.

The user-independent approach or the across-subjects approach is a cumulative comparative analysis treating all subjects in a concatenated fashion. The average changes on the dependent variables between conditions or groups are used, assuming that the manipulation of independent variables had a holistic effect on the majority of the participants. Other researchers [69], [365], [369], [370] have used this approach to create affect detection models that can be applied to new users. High individual variability in emotional responses make the creation of such models a difficult task. The validation accuracies decrease especially when the testing data originate from 'new' users (not previously used for the training of the model) instead of using mixed-subjects cross validation (where part of users' data can be found in both training and testing set) [323].

3.6.3. Data Pre-Processing

The classification experiments and statistical analysis tests were made after the pre-processing of all data, including the processing of the physiological data. For each experiment exhibited in the following chapters, specific steps for the pre-processing of the raw data were followed including data filtering and feature extraction. The processing steps conducted from the Feasibility study 1 and 2 are explained in Chapter 4 '*Data processing'* section and for Feasibility study 3 VR study in Chapter's 6 '*Signal Data Processing'* section. All data including questionnaire

responses, continuous-self-ratings and data streamed from our sensors, were linked together, concatenated, and parallelised for each participant, into one large dataset. Subsequent datasets were extracted for the different user groups related to the analysis goals of each experiment. Final steps included the pre-processing of sensor data (incl. cleaning, filtering etc.) and the organisation and concatenation into a format for further statistical and classification testing. The raw data from our sensors and the virtual environment were stored in either .txt/.csv(ASCII, comma delimited) or json format.

3.6.4. Quantitative analysis of responses data

The appropriate statistical methods were adopted to assess relationships between the independent variables and the continuous dependent variables, which are experiment dependent. Generally, for both experiments, we were interested in the comparative evaluation of the effects of the factors for each study on our measures, e.g., group type (if any), conditions (stimuli), signal features on our physiological measures and on the self-ratings of arousal and valence. Kolmogorov-Smirnov and Shapiro-Wilk tests of normality were performed on the data to assess whether the data were following a normal distribution. Analysis of variance (ANOVA) was used to test the significant differences between means of groups and/or multiple conditions (factors) within subjects which followed a normal distribution. Significant effects found from ANOVA ($p < .05$) were examined with subsequent Bonferroni-corrected post-hoc tests. In cases where the number of observations were above 25 and there was slight departure from normality, ANOVA was still used as analysis method because it is robust to small violations to the normality assumption [371]. The ANOVA was followed up with non-parametric post-hoc tests where necessary. For highly non-parametric distributed data we used the Friedman's ANOVA test to test for significant differences between conditions (e.g., positive, neutral, negative stimuli), followed by multiple comparisons Bonferroni-corrected post-hoc tests in case of statistically significant results, such as Wilcoxon signed rank test for within-users pairwise comparisons, and Mann-Whitney U for comparisons between two independent groups. The tests were performed in SPSS® [372] and MATLAB® [373]. The significance 'alpha level' value for all tests was $a = 0.05$ (5%) (and thus any p values which was equal or less than the value were considered statistically significant).

Additional tests included Simple and Multiple Linear Regressions to investigate the ability of the independent and explanatory variables to predict changes of the continuous, dependent variables such as the level of arousal and valence. We tested whether regression models were significant and concentrated on the r-squared values (or else coefficient of determination) because it is related to the goodness-of-fit of the regression line. The r-squared value is expressed between 0%- 100% (the percentage of the dependent variable variation explained by the model).

Correlation analysis were also computed to measure relationship between two measured variables e.g., correlating the root mean square (RMS) values of one EMG channel with the self-rating on valence. Pearson correlations were selected for data following a normal distribution (or when sample size N>=25) while Kendall's tau or Spearman's correlations were used for non-normal distribution. The correlation coefficient r indicates the strength and direction of the relationship between the variables.

3.6.5.   Classification algorithms

We explored both user-dependent and user-independent approaches in both experiments. In the feasibility study 2 (FEDEM 2: Chapter 4) we use a regularised Support Vector Machine (SVM) classifier and in the feasibility study 3 (FEDEM 3: Chapter 6) we used three classifier methods: Support Vector Machine (SVM), Naive-Bayes (NB), and k-nearest neighbour (KNN). These classifiers have been suggested [327]–[328] and used in emotion recognition studies before showing promising results (KNN on f-EMG and heart-rate signals [376], [377], SVM used for discrimination of facial muscular activations [369], [378]–[381], and NB for affect detection from physiological signals [310], [382], [383]). Automatic hyperparameter tuning was applied to optimise the penalty parameter γ and the kernel function parameter σ for the radial basis function (RBF) of the SVM [384], the distance metric and the k variable of the KNN, and the 'width' parameter of the NB. To develop the model we used the internal functions of MATLAB® and the libSVM® [385] library.

Signal features were extracted from the pre-processed EMG and PPG signals. The features selected for each experiment as specifically described in the corresponding section of each chapter. These features were used as the input variables (predictors).

The arousal/ valence responses from the CASR self-rating tool were rounded to create 2 and 3 categories (for 2-classes: negative and positive for valence, low and high for arousal, and for 3-classes: negative, neutral, positive for valence and low, average, high for arousal).

The classifier was then used to map the features ranges with the responses during training. A 10-fold cross validation (CV) approach (used in FEDEM 2 and 3) and a leave-one-subject-out (in FEDEM 2) were adopted as explained in Chapter 6. These methods are commonly used for supervised learning to assess the predictions ability of one model using data which were not previously used in the training of the model, thus avoiding overfitting [386]. K–fold CV obtains k number of subsets having equal number of members from the dataset which are randomly distributed in the subsets. In practise, each of the subsets is used as a testing set in turn, and the remaining data as the corresponding training set. The resultant prediction accuracy rate is calculated by the average accuracy rates of all k predictions (out-of-sample accuracy). In the leave-one-subject-out (LOSO) CV approach, the data from one or more subjects (the number used is experiment-dependent) are used for the evaluation of the model, while the remaining data (from all other users) are used for the training of the model [294]. Therefore, the data from each participant can either be used for the training or the testing of the model. The reported accuracy is the ratio of correctly detected labels against the total testing labels expressed as percentages.

3.7. **Ethical Considerations**

All experiments reported in this thesis were pre-approved by the Bournemouth University Research Governance and Ethics Committee. Specifically, the feasibility study 2 (FEDEM 2) was approved on 26/01/2017 (ID: 12025) and the feasibility study 3 (FEDEM 3) was approved on 13/02/2018 (ID: 18848).

In this final section we will address a range of practical and ethical considerations:

Data protection: The requirements of the Data Protection Acts 1988/ 2018, the General Data Protection Regulation (Regulation (EU) 2016/679) and the data processing requirements of Bournemouth University have been complied with. This includes the secure collection, storage, dissemination, and retention of data.

Numerical IDs were randomly selected, to link pre-trial surveys, main-study data and post-experiential self-ratings which were collected from different devices.

Voluntary informed consent and participation: Sufficient explanation of the experimental process and the potential risks were given to the participants before they were required to sign a consent form either in paper or in digital format, confirming their approval in taking part and allowing their anonymised data to be collected.

Emotional Stimuli: Affective audio-visual stimuli were selected to induce various intensities of valence and arousal. We anticipated that they may induce negative emotions to the viewers, and in some cases stress. To reduce the level of stress as much as possible, we carefully selected the stimulus material that was used based on a) the content of the stimuli used in similar studies and validated databases and b) subjective evaluations from individuals in our pilot and validation (survey) study. However, since the affective impact of content is highly subjective, we were always in communication with the participant, allowing for breaks within the experiment and early withdrawal from the study in the case of fear for negative symptoms.

Reduce potential harm: We informed the participant about the nature of the stimuli and entered screening questions before the experimental studies to check for different health and psychological conditions (e.g., anxiety, depression) or phobias related to the selected stimuli, such as arachnophobia (excessive fear of spiders), together with well-being questions (e.g., motion sickness) which was administered before and after each block of sessions throughout the experiment.

Audio-visual recordings: For the evaluation of our measures related to the facial muscle activity we collected visual evidence of the facial responses. The audio-visual materials were stored in a secure UK based digital environment with suitable password protection. Transcription was undertaken in a secure and confidential environment and was not used in any outputs (publication, dissemination, etc.)

Compensation: Participants involved in our laboratory-based studies (FEDEM 2) were compensated for their time with voucher at a value of £10 per hour. Participants in our last study (FEDEM 3) were volunteers from the Science Museum in London

(UK) who participated without the expectation of a compensation. All participants were given a debrief postcard with the details of the study, and the contact details of the experimenter.

3.8. **Chapter Discussion, Conclusions and Interconnections Between Chapters**

In this chapter, the overarching methodology used to investigate physiological responses to affective stimuli was descripted. The methods proposed were adapted for integration with VR technologies. The experiment set-up (sections 3.3 - 3.5), the study factorial designs (section 3.6), and experimental procedures pursued in the studies were also described. The overall system architecture was described in section 3.2. This system was comprised by the input methods, which recorded physiological and self-ratings data from the users, the data processing and classification methods, which were used to generate models that mapped the physiological features to the self-reported levels of arousal and valence. The system outputs were the detected levels of valence and arousal in discrete, two and three classes (Figure 13).

As part of the input methods, we presented the prototype sensor set-up which was developed and tested in feasibility study 1 comprising EMG sensors along the frame of the VR headset for easy integration in VR settings. In the next studies, the prototype was enriched with additional sensors, including a photoplethysmographic (PPG) sensor for heart-rate detection and an IMU, for movement tracking. Usability factors were also investigated as part of this EngD, including comfort, fit on the skin, sensor contact during facial movement and integration within commercial VR headsets. As a result, multiple iterations of the prototype were designed and informed from the studies described in the next chapters (chapters 4 and 6).

The SAM ratings were used as the main method of reporting valence and arousal levels and was later adapted into a custom-build tool, called CASR. This tool developed for capturing the continuous valence and arousal ratings from the users in VR was described. The CASR tool was developed for the HTC Vive controller, and was used during the feasibility studies. The position of the user's finger on the controllers' trackpad were mapped into the 2-dimensional space and were outputted in ASCI-formatted file together with the signal data for data synchronisation purposes.

For each feasibility study a custom-build stimuli presentation environment was built, which included event-marker in-build systems for automated data labelling. These systems were used with linear video presentation used in the studies described in chapter 4, and also in the interactive, VR experiences used as part of chapter's 6 VR study. The stimuli (video and virtual stimuli) were validated by

online survey studies. The statistical analyses tools and methods used were described in section 3.6. The tests were used to investigate the relationship of the dependent variables and independent variables of our studies.

This system architecture allows a robust recording and analysis of physiological responses to affective responses in VR settings. This setup could allow the integration of additional sensors as EEG for VR (see overview of emerging technologies for affect detection in VR in section 2.3.3.3), which could enhance the multimodality of the set up even further. Saying that, this endeavour is outside the scope of this EngD thesis.

The practical and ethical considerations addressed in our studies were described in section **Error! Reference source not found.** including data protection, informed voluntary participation, compensation, the selection of emotional stimuli and measures taken to reduce risk and protect the participants in our studies. The studies reported in this thesis were reviewed and approved by the Ethics panel of the Bournemouth University.

Figure 23 shows the design iterations of the sensor set-up described in this chapter used in the first feasibility studies (1 & 2, explained in chapter 4) and the progression of the experiment set-up towards the main VR study, described in chapter 6. Chapter 4 discusses the two preliminary experiments for assessing the feasibility of valence and arousal detection from the sensor set-up proposed. Experiment 1 focused on the use EMG sensors around the area of the face for integration with commercial HMD devices and the development of the experimental paradigm and event tagging system which was used in the subsequent studies. Experiment 2 added the integration of additional PPG sensors into the Faceteq device. These were tested for feasibility. Many of the methods (and hardware used) described in the current chapter were developed for these two preliminary experiments. The final version of the device comprising the types and number of sensors suggested by those experiments was used in the VR study described in Chapter 6.
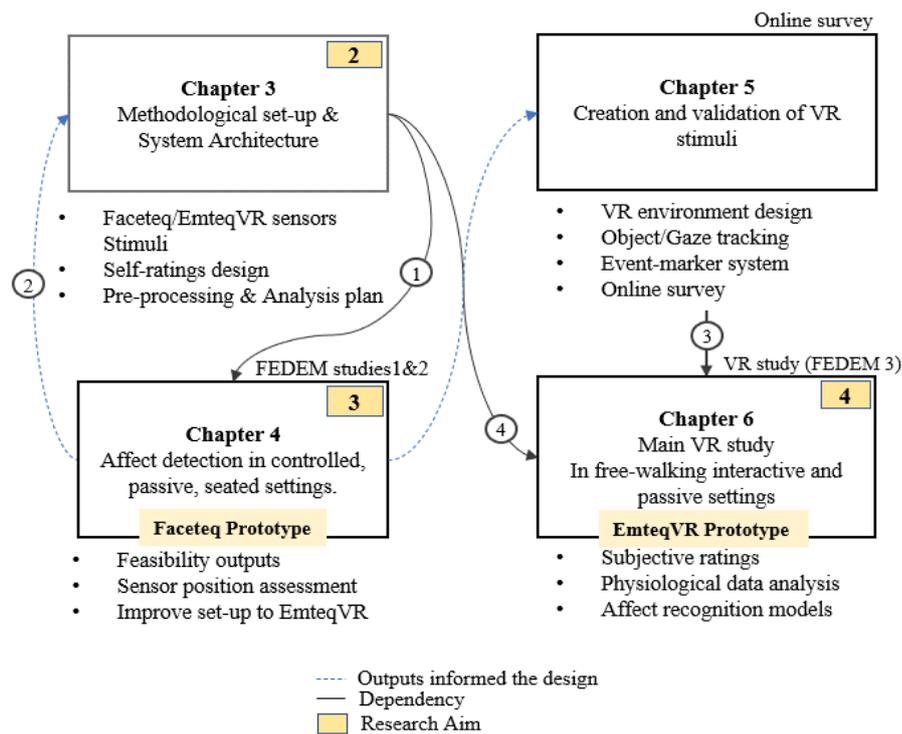
Figure 23. Diagram showing the connections and dependencies between chapters. Each box outlines a separate chapter. A high-level overview of the contents of each chapter are described under box. Each chapter is connected to a research aim whose corresponding number is added on the top right corner. The continuous lines reflect dependencies between chapters and the dotted lines represent information flows between chapters. For example, the methodology and system set-up described in chapter 3, was applied to the studies described in chapter 4 and in chapter 6. Chapter 5 was informed by the outputs of chapter 4 and was prerequisite to chapter 6 as it provided the validation of the VR stimuli used in VR study.

Chapter 4

Feasibility Studies on valence and arousal detection with the Faceteq prototype (FEDEM) 1 & 2

4.1. Introduction

Experiencing Virtual Reality environments involves a naturalistic user behaviour, i.e., free movement of the body. Recording high-quality physiological responses to affective Virtual Environments is still a challenge because acquiring good signals from physiological sensors rely heavily on the sensor quality, the contact with the skin, and the ability to deal with enhanced levels of artefacts in the data caused by the interactive nature of the VR set up [387]. The experiments reported in this chapter, do address this issue by evaluating the feasibility of detecting valence and arousal levels with our sensor set-up as described in Chapter 3, section 3.2. For these initial feasibility experiments, it was decided to conduct studies which required no head movement by the users, using a standardarised affect induction paradigm. For this purpose, we used 2-D stimuli on a virtual screen which allowed the user to look in the same direction throughout the experiment. Videos from validated libraries were used as stimuli for these experiments. Similar stimuli have been used by the majority of affect stimulation studies as described earlier in Chapter 2.

One online questionnaire and two feasibility studies are reported in this Chapter. The feasibility studies investigated a) the effect of positive, neutral and negative valence on the f-EMG signals, and b) the effect of low and high arousal on PPG signals are described. Thus, we designed these studies to test sensitivity of our dependent variables, the physiological signals derived from our prototype set-up, by inducing high and low activations varieties along the dimensions of valence and arousal. The position of the sensors (on the mask insert), the materials used, the

experimental paradigm used and the fit of the mask to the face of the users were also aspects we wanted to determine for future studies.

For these feasibility studies we decided to use affective and neutral videos as stimuli because the emotional effect on physiological signals from audio-visual videos exceeds often the effect elicited by static images [388]–[391]. Indeed videos engage auditory and visual senses simultaneously, and variably over time [180]. We also expect VR 3D stimuli to induce more intense responses than 2D stimuli. Another reason behind the choice of videos, is that they are one step closer towards a VR experience than static images. The narrative, dynamic nature of an unravelling event and the experience of a continuous situation in which emotion can be dynamically modulated, can also exist in an experience within virtual reality. These dynamic changes over time and the contextual background often enhance emotional experience [218], [392], [393].

To ensure that our affect induction would induce the different categories of valence and arousal (positive, neutral, negative valence with low-high arousal levels), we decided to use videos from a previously validated video library. The library had videos which were new at the time and, hence, we expected that they were also new to our participants (based on the evidence that habituation can decrease physiological reactivity [394]–[396]). However, because the video library was rather new, we also decided to revalidate the stimuli in our feasibility studies using a video validation survey.

After careful examination of the database, we found that there was a high percentage of videos consisting of babies in the positive category, while in the negative category there were some videos that could be perceived as very distressing. In our effort, to choose the best combination of videos to represent our affective categories we conducted this video validation survey, asking participants of our target age group (18-35 years) to rate the videos in the terms of arousal and valence, indicate if they have ever seen the video before, and flag videos which they found very distressing. From this survey, a list of 40 videos were selected and were used in the following feasibility studies measuring valence and arousal.

In short, this chapter will describe these three studies: (a) the affective video validation study using an online video validation survey, (b) an experimental feasibility study on valence detection, and (c) an experimental feasibility study on arousal detection using the Faceteq prototype. For these studies we followed an affect induction protocol that used video-clips on a screen; a protocol which has been

used in numerous affect detection studies, e.g. [69], [365], [397] (see an explanation of the protocol on section 4.2.1).

In Study 1, a selection of videos originally extracted from a film database by [397] were rated by 82 participants in terms of valence and arousal using an online survey. Based on the average ratings per video, 20 affective ones and 20 neutral videos were selected. In Study 2, data from eight EMG sensors were recorded from participants using the Faceteq insert (without wearing an HMD). In Study 3, a PPG sensor together with an ECG belt were embedded on the insert. Apart from the additions made on the apparatus, the experimental set-up and procedure were kept identical between the two feasibility studies, including the video-clips, the video-presentation software, and the self-rating method.

4.2. Study 1: Affective video validation study: An online survey

To ensure the reliable induction of affective states within our studies, we started investigating the available stimulus databases, in particular the video databases. Videos from a film database by [397] published in 2016 were chosen for our online validation survey study because the database had short clips of various affective states that were previously rated on arousal and valence rating scales by 411 participants. It contained crowdsourced clips which were collected from online video hosting services. The film database was published just before the start of our survey design development.

Through initial viewing of the videos, we identified that a large portion of the positive low arousing and high arousing videos contained faces of infants. Although the effect of infant faces has shown certain correlations to positive attributions, and higher attention on mothers [398], we were unsure whether the same effect would be prominent in our targeted participant population, i.e. female and male university staff and students between the ages of 18 and 35 years. Therefore, we decided to assess the affective impact of selected videos. As part of the conditions for the creation of the video database, we also wanted to find out whether those videos have been viewed by participants beforehand because they have been circulated on online media portals in the past. Previous exposure to the videos is likely to impact the affective ratings due to habituation effects (causing weaker

affective responses overtime compared to new viewers [305]). Therefore, we decided to exclude the video clips with high familiarity scores from the stimulus material used in future studies.
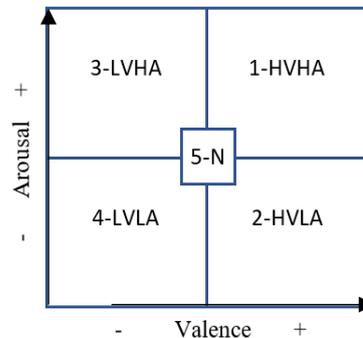
Figure 24. The five affective states represented in the affective space. HVHA: High Valence (positive) and High-Arousal, LVLA: Low Valence (negative) and Low-Arousal etc. N: Neutral.

Therefore, a survey was created to revalidate the evoked valence and arousal levels by the selected videos. The videos were originally divided in categories that could elicit negative, positive, neutral and, and mixed emotional states (see [399]). The valence and arousal ratings per video for each category provided by the authors were used to identify videos that could induce distinct low and high arousal responses as well as negative, neutral and positive valence. The categories for the video selected were five in total: four based on the four quadrants of the affective model (High-Arousing Positive, High-Arousing Negative, Low-Arousing Positive and Low-Arousing Negative) and one for the neutral-medium arousal state, as shown in Figure 24**Error! Reference source not found.**. The most representative and distinctive videos for each of the five categories were carefully chosen. Specifically, eight of the most representative videos for each targeted affective category and 25 neutral videos (total of 57) were selected for this survey.

As stated in chapter 2 (Section 2.3.1), the choice of these video categories was based on the dimensional model of core affect. As explained for the studies reported in section 4.3 and 4.4, establishing these distinct videos categories as stimulus material will allow us to record distinct physiological measures (facial EMG, and PPG) which can be continuously and simultaneously captured, together with the subjective arousal and valence ratings of the videos. For example, the recording of heart-rate changes in non-affective (neutral) and affective (negative or positive) settings. In addition, the creation of a database specifically designed for the study of affective responses from physiological signals will be valuable for future

studies where, for example, effects of the extreme polarities of the two dimensions could be investigated.

In summary, the aims of this survey were to:

1. assess affective impact of each video using subjective arousal and valence ratings and select five clips per affective category as well as 20 neutral videos as stimulus material for future feasibility studies.

2. identify videos which have been already seen by a large percentage of our participant population. These videos would be excluded from the studies precautionary so as to account for the effect of habituation and thus reduced affective response.

We predicted that the selected videos would generate the expected valence and arousal ratings in our target group as per the ratings reported by Samson et al. (2016). We also expected that the negative videos would be more arousing that the positive ones (as seen in studies described in section 2.3.3).

4.2.1. Methods

Participants

For the survey, 82 participants from the student and staff population of Bournemouth University (78% females, 22% males) with an age range of 18-40 years (M_{age} = 24.22 (±6.64)) were recruited. Participants were screened for mental/psychological disorders (e.g., clinically diagnosed anxiety and depression). The majority of the participants had not participated in a similar study before (78,05%). Participants were given either SONA credits and/or vouchers (£5) as compensation for their time.

Materials and Procedure

The 54 selected videos used in the survey are listed in Table 3. The videos belonged to one of the following five categories: positive-valence high-arousal (PH), positive low-arousal (PL), negative low-arousal (NL), negative high-arousal (NH), and neutral-medium arousal (NEU). The valence and arousal ratings from the original study by Samson et al. [247] are visualised on the cartesian affective space in Figure 25 (the individual scores are also displayed in the Appendix C, Table). Since the original mean SAM ratings are given on a range from 1 to 6, we rescaled the values

to 1-9 using the following formula: $x^{new} = ((x^{old} - 1) / (5)) \times (8) + 1$, where x^{old} is the value which is converted, and x^{new} is the rescalled value.

Survey design. The videos were embedded within the survey and all information that could bias the participants i.e., titles of videos, were deleted. The survey was designed in Qualtrics and administered via 'anonymous link'. A short demographic questionnaire and a training session on how to rate videos using the SAM scales were added at the beginning of the survey using videos which were not used in the rest of the survey. We asked participants to view the videos using speakers or headphones. A quick audio check was added using an audio file to ensure all participants had the sound volume adjusted so that they could hear the sound from the videos.

The survey was programmed to randomise the categories of the videos and the videos within each category. A white, blank page was dedicated per video and the self-rating scales for each were not made visible unless the participant had watched the video in full. A 9-point slider was included per valence and arousal scale with the original SAM figures on top of each scale, and text descriptions were added above numbers '1', '5' and '9' per scale. These descriptions were taken from the instruction manual provided with the SAM scales [64]. Under the SAM scales, a question on whether the participants had seen the video before was added. Participants could answer with 'yes' or 'no' by clicking in the respective button on their screen. Feedback from each participant were requested on whether there was an issue with the flow of the survey and general feedback on the videos used, asking specifically for any very distressing videos found in the survey. The option for feedback was provided through the free text boxes within the survey.

4.2.2. Data analysis plan
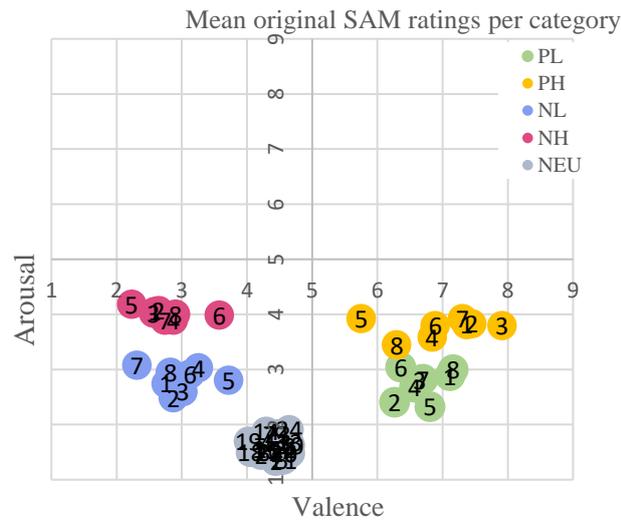
Figure 25. Rescaled valence and arousal scores (range: 1-9) based on ratings from the film data base [247].

The criteria below show the conditions which the selection of videos had to satisfy for the creation of the video database that could induce the five categories of affect. As part of the criteria, the selected videos per category had to fall within specific valence and arousal score ranges, but also show reduced familiarity by our target age group. A threshold of 30% was chosen for the familiarity (i.e., if three out of 10 participants had seen the video before), above which the video would be excluded. The ratings of the videos from the survey would indicate their ability to induce the respected levels of valence and arousal. Since we had three categories for valence, negative, neutral and positive, the valence scale was divided into three parts: from 1 (very negative) to 4.5 (where 5 was neutral) was indicated as 'Negative' area, from 4.5 to 5.5 were indicated as 'Neutral' and from 5.5 to 9 (very positive) were indicated as 'Positive'. The videos of the preselected valence categories were required to be induce the corresponding mean valence SAM score. Similarly, the videos were classified into two main categories of high and low arousal (based on the original ratings, see Figure 25) whether their SAM ratings exceeded 3.5 (high: 3.5-9 (very intense), low: 1 (very calm, sleepy)-3.5. For the neutral videos, mean arousal scores were expected to be lower than 2.5 since their affective impact was intended to be lower than all the other categories. All the criteria for including the videos in the corresponding category as described below:

1. Percentage of participants familiar with the video less than 30%.

2. PH (Positive, High arousing): Mean Valence>5.5, Mean Arousal>3.5

3. PL (Positive, Low arousing): Mean Valence>5.5, Mean Arousal<3.5

4.   NH (Negative, High arousing): Mean Valence<4.5, Mean Arousal>3.5

5.   NL (Negative, Low arousing): Mean Valence<4.5, Mean Arousal<3.5

6.   NEU (Neutral, Low arousing): 4.5<Mean Valence<5.5, Mean Arousal<2.5

7.   Mean Valence scores of PL and PH are equal

8.   Mean Valence scores of NL and NH are equal

9.   Mean arousal scores of PL, PH, NL, NH > Neutrals

10.   Similar difference of arousal scores between the two positive and the two negative categories.

To control for changes in arousal within valence categories, our team added criteria no. 7 and 8. This equality between valenced categories permitted us to study arousal changes when valence scores are of similar value. This equality could be expected for the arousal scores between the positive and negative groups since we did anticipate negative videos to elicit higher levels of arousal [400], [401] (additional studies are described in section 2.3.3). The physiological impact of arousal for the neutral videos was expected to be lower than that of the other categories (see criterion 9) [401], [402]. For each couple of categories per valence hemisphere (positive and negative), the difference in mean arousal was expected to be of similar value, so as to allow for the study of arousal changes independently of valence. This design is related to the affective model theory where valence and arousal dimensions are not correlated, and high arousal can be found in either positive or negative (see affective model, p. 16).

4.2.3.   Results

Familiarity with the videos: The question 'Have you seen this video before?' was testing the familiarity with the presented videos. The overall familiarity was very low (M: 4.92% ± 0.10). Only two videos scored more than 30% (clips named 'PL1' and 'NEU5'). Detailed percentages per video are presented in Figure 26.
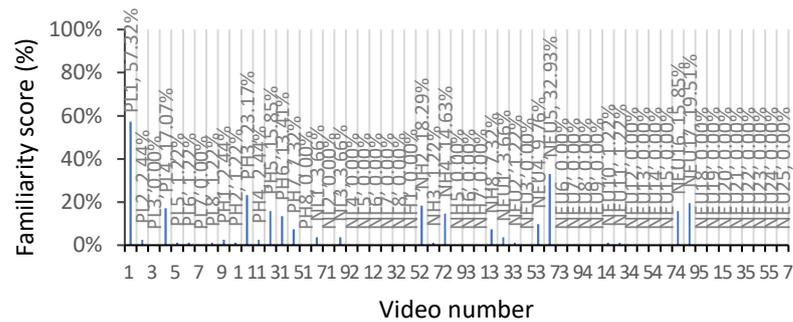
Figure 26. Percentage of familiarity with video per clip from the survey. Acronyms indicates the name of the video (see Table 4), accompanied by the familiarity score (%).

Valence and arousal ratings: The mean arousal and valence scores and standard deviations for each of the video clips are displayed in
Table **5**. The mean ratings for all videos per category and the mean scores across all videos in the category are represented in Arousal-Valence (AV) Cartesian Space in Figure 27 and Figure 28.

Table 5. Mean AV scores per videoclip used in the survey.

ID	Title of clip	Mean Valence	(±SD)	Mean Arousal	(±SD)
PL1	Babybitesbrosfingers •	7.4	1.59	4.3	2.13
PL2	Babydoesn'tlovehis daddy	6.85	1.36	3.32	1.99
PL3	Babyshiccupandjauph	7.27	1.38	4.03	2.19
PL4	Cookiebaby	6.92	1.30	3.78	1.81
PL5	Smartbabywithpacifier •	6.64	1.47	2.95	1.84
PL6	Excalatorspinning	6.47	1.51	4.48	2.06
PL7	Beatboxbabydance	7.13	1.24	4.08	2.09
PL8	Catsucklesair •	7.21	1.38	4.38	2.04
PH1	Babydancebeyonce	7.49	1.44	4.43	2.34
PH2	Babyfailshulahoop •	7.62	1.34	4.51	2.34
PH3	Babycontrolscheers •	7.39	1.46	4.3	2.28
PH4	Bridelauphingduringvows	7.46	1.18	4.7	2.11
PH5	Girlthrownintobasketballhoop •	6	1.50	5.76	2.08
PH6	Pandasneezealot	7.17	1.25	4.37	2.34
PH7	Singingdog	7.17	1.95	4.6	2.25
PH8	Weedingphtographerfail	6.8	1.42	4.88	1.53
NL1	Armbentfromskateboard	3.29	1.93	5.47	2.27
NL2	Bmxfaceplant	3.86	1.57	5.28	2.02
NL3	Boyfaceplants •	3.77	1.56	5.42	2.15
NL4	Bullwrongtarget •	3.1	1.51	5.77	2.28
NL5	Snowboardercrashes •	4.08	1.45	4.85	2.07
NL6	Tablebackflip •	4.34	1.88	4.97	2.17
NL7	Skaterfallsbreakwrist	2.93	1.62	5.65	2.03
NL8	Bikeintowall	3.92	1.59	5.24	2.15

NH1	Bikefalloffcliff -intense	3.19	1.79	6.32	2.03
NH2	Bullthrownandtrample- intense	2.96	1.92	6.13	1.80
NH3	Carhitsskater -intense	3.15	1.59	6.24	2.19
NH4	Breakdancerkickskid-intense	4.1	1.94	6.06	2.26
NH5	Crocbitesman - intense	3.01	1.82	6.79	2.10
NH6	Fatboyrollercoaster -intense •	5.68	1.35	6.26	1.75
NH7	Horribleskiaccident	3.43	2.44	5.71	1.90
NH8	Motorcyclejumperfruiflesshort •	3.54	1.16	5.56	1.87
NEU1	Bart	4.94	1.20	2.51	1.56
NEU2	Boydrinkingtea	4.75	1.71	1.92	1.40
NEU3	Assembly	4.92	1.59	2.33	1.80
NEU4	Airport2 •	4.8	1.28	2.5	2.00
NEU5	Cablecar	5.03	1.37	2.37	1.46
NEU6	Café	5.1	1.37	2.22	1.40
NEU7	Cityinthenight	5.2	1.24	2.56	1.63
NEU8	Denvertrain •	4.86	1.46	2.65	2.10
NEU9	Eatingpizza	5.01	1.74	2.41	1.52
NEU10	Eatingwithchopstics	4.91	1.51	2.33	1.77
NEU11	Gilsbruthingtheirteeth	4.84	1.84	2.41	1.55
NEU12	Hairwashing •	5.24	1.41	2.11	1.56
NEU13	Hikinginthewood	4.74	1.69	2.19	1.60
NEU14	Museum	4.85	1.55	2.15	1.43
NEU15	Nystreet •	4.92	1.40	2.96	1.71
NEU16	Ridingthetube1	4.88	1.62	2.2	1.44
NEU17	Ridingthetube2	4.92	1.68	2.18	1.65
NEU18	Sanfran	4.85	1.36	2.35	1.60
NEU19	Sittingonthesofa	5.17	1.43	2.18	1.51
NEU20	Snow	5.18	1.39	2.68	1.29
NEU21	Swimlaps •	5.12	1.45	2.41	1.70
NEU22	Tea	4.95	2.04	2.09	2.00
NEU23	Treadmill	4.9	1.39	2.22	1.29
NEU24	Vangoghmuseum	5.08	2.12	2.3	2.05
NEU25	Pillow	5.06	1.50	2.13	1.48
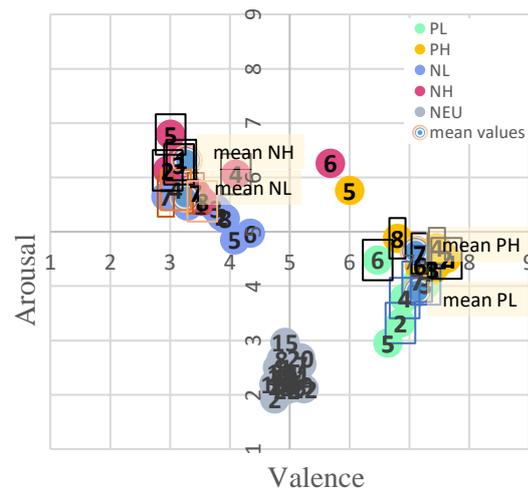
Figure 27. Mean valence and arousal scores per video are represented in the AV space, for the 5 affective categories PL, PH, NL, NH, NEU. The mean values per affective category (four in total) are represented as 'mean values'. The five videos per affective category that were selected for the studies are indicated by a square surrounding the video-points.

Video clip selection for future studies: Five clips were selected per affective category (PL, PH, NL, NH) and 20 videos for the neutral category based the selection criteria set out in the methods section (section 4.2.2). The selected videos are displayed as framed with squares in Figure 15. The excluded videos are marked with a blank dot (•) in Table 5. More specifically, mean ratings were calculated for each video, and outliers per affective dimension, such as videos PH5 and NH6, were removed to create more homogeneous groups. Based on the feedback of our survey participants, video NH8 was also excluded because it was perceived as being extremely negative. The remaining videos were included or excluded based on visual inspection with respect to the clear representation of the five categories on the AV space (see Figure 15) and the selection criteria (see criteria 2-10 section 4.2.2). The mean valence and arousal ratings for these selected videos are displayed in Table 6.

Table 6. Mean valence and arousal scores for each affective category after video selection for future studies.

Categories	Mean Valence (SD)	Mean Arousal (SD)
PL	7.11 (±0.23)	3.90 (±0.37)
PH	7.10 (±0.47)	4.63 (±0.16)
NL	3.26 (±0.25)	5.63 (±0.12)
NH	3.28 (±0.47)	6.30 (±0.29)
NEU	4.97 (±0.15)	2.33 (±0.23)
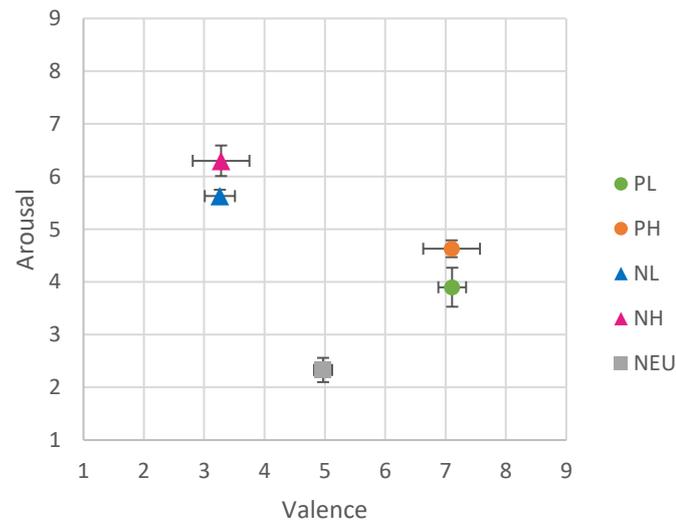
Figure 28. The mean arousal and valence scores across all selected videos per category. The vertical and horizontal lines indicate the standard deviation for each affective dimension.

Afterwards, one-way repeated measures ANOVAs were performed to test whether the valence and arousal ratings significantly differed between the five categories.

Valence ratings. The results showed that the five categories were significantly different in their valence ratings ($F(4,324) = 325.402 = p < .001$, eta.$= .801$). Post-hoc tests with Bonferroni correction showed significant differences in valence between the videos from the four affective categories against the neutral videos ($t(7) = 10.34$, $p < .001$), and significant difference between the positive and the negative categories of the same arousal (PH-NH: $t(7)=9.936$, $p<001$, PL-NL: $t(7) = 17.79$, $p<.001$). The valence differences between PH-PL and NH-NL were found not significant ($p=.426$ and $p=.930$) which showed that they elicited the same level of valence regardless of their level of arousal. This effect was expected per the initial selection of the videos (criteria no. 7 and 8).

Arousal ratings. There were also significant differences between the categories for the arousal ratings ($F(4,324) = 129.161$, $p<.001$, eta.$=.615$). Post-hoc tests showed significant differences in arousal between all conditions (PL-PH: $t(7)=-2.39$, $p=.48$), NL-NH: $t(7) = -3.74$, $p=.007$, PH-NH: $t(7)= -8.66$, $p<.001$, PL-NL: $t(7) = -7.30$, $p<.001$, PL-NEU: $t(7)=9.02$, $p<.001$, NL-NEU: $t(7)=26.37$, $p<.001$). These results showed that affective videos were significantly more arousing than non-affective (neutral) videos and that significant difference were found between the valenced

groups. The negative videos per category were found significantly more arousing than the positive ones, supporting our initial expectations.

4.2.4. Discussion

This survey was designed to validate the selection of videos for five affective/neutral categories (from the video database [397]), [397]), that were chosen based on the quadrant arousal-valence model (high-arousing positive, low-arousing positive, high-arousing negative, low-arousing negative, neutral). The results from the online survey presented, showed that the selected video clips were able to induce three distinctive valence states, positive, neutral and negative, and three distinctive levels of arousal, a high and low for the affective conditions (positive and negative) and one for the non-affective non-arousing category (neutral). This design allowed us to investigate the effect of valence and the effect of arousal independently of one another (as per the circumflex model of affect, see Section 2.3.1, p.44).

The videos were selected based on the category creation criteria specified in section 4.2.2. The first criterion was the familiarity with the videos was less than 30%. The response from the survey showed that the majority of the videos selected scored very low on familiarity scores, which showed that our sample target age group had not been previously exposed to them via other studies or types of media. This effect was expected since the specific video database was recently published and included video clips from online streaming platforms. The familiarity percentage of the videos selected from this survey was less than 25% with the majority of videos scoring 0% for all participants (signifying no familiarity). The mean familiarity score across all videos was 4.92% which indicated that the videos have not been watched by the targeted population before.

The main analysis was related to the mean valence and arousal ratings for each of the five video categories, namely PL, PH, NL, NH and NEU. The results from the survey showed that the videos per affective category achieved the expected valence scores which satisfied criteria 2 to 6 for valence. For arousal, the mean scores reported per category also satisfied the specified criteria, although negative categories exceeded the expected arousal ranges, compared to the scores presented by [397]. The neutral videos scored very low on arousal as expected, which satisfied criterion 6. In order to investigate arousal and valence in future studies independently, the two positive and two negative video categories were designed to induce the same valence ratings (criteria 7 and 8). This was indeed the case. The two

positive categories, PL and PH, achieved the same mean valence ratings. The same was true for the two negative categories, NL and NH.

In terms of arousal, based on previous research, it was expected that affective videos would always have higher arousal ratings relative to neutral videos (criterion 9). In addition, the two positive categories (PL-PH) were designed to have the same arousal difference as the two negative categories (PL-PH) (criterion 10). The findings were clearly in line with criterion 9, i.e., arousal ratings were higher for affective compared to neutral videos. However, the negative videos induced a stronger effect along the arousal dimension than the positive videos. The difference between NL and NH was found significantly different, however the difference between PL and PH was not found significant. As a result, criterion 10 was not satisfied. This result was somewhat expected, since negative stimuli have been reported in the past to elicit a larger effect of on arousal than positive stimuli ([403]–[405], due to negativity bias [400] and attentional bias [406].

Overall, the survey study confirmed the creation of video database for the induction of three discrete levels valence, negative, neutral and positive, and three levels for arousal, low, medium and high. This database enabled us to investigate the effect of valence and arousal on the physiological signals in the next studies collected via the sensors embedded in the hardware prototype. The next sections describe these studies. As we explored the feasibility of detecting valence and arousal from the novel set-up using facial sensors, we called the next studies ‘*Feasibility studies on emotion detection*’ (i.e., FEDEM). The next section will describe the study that focused on valence detection and the following one will describe the study that investigated arousal detection.

4.3. Towards valence detection from EMG for Virtual Reality applications (FEDEM 1)

In 2016, we proposed a novel hardware solution, 'Faceteq', for facial muscle activation monitoring [329] in VR, consisting of eight electromyography (EMG) sensors. We hypothesise that such an interface can track the valence information needed for continuous emotion assessment in VR. Our team designed this pilot feasibility study to explore the effect of spontaneous facial expressions on the EMG sensors incorporated on the Faceteq interface. To our current knowledge, this is the first study where integrated surface facial EMG sensors have been used for spontaneous valence detection in VR. We investigated the feasibility of this approach in controlled conditions, using audio-visual stimuli on a monitor. The video clips were selected to induce five affective states, each per quadrant of the affective space and one for neutral. In this study, participants watched a randomised sequence of the selected video stimuli while self-rating their emotional state continuously. For this study, we only analysed the valence ratings of those videos After a specifically designed signal pre-processing, we aimed to classify the responses into three classes (negative, neutral, positive). Our prediction was that the EMG sensors on the location predefined (see 3.2) would be able to read facial muscle activations responsible for facial expressions of positive and negative affective states (see also 2.3.2).

4.3.1. Methods

4.3.1.1. Participants

For this study, 35 participants were recruited (20 females, 15 males). We excluded the data from one participant due to synchronisation issues, bringing the total of participants used for data analysis to 34. The ages of the participants ranged from 18 to 40 years old (M_{age}: 22.8 (±5.2) years). Each participant was compensated by a £5 voucher for their time.

4.3.1.2. Materials

The physical space –

The study took place in a room (3m long x 3.5m wide) with two computer desk spaces divided by a wall partitions, see Figure 35. An infrared enabled camera (sample rate: 60Hz) was installed on one of the desks, which allowed camera recordings with poor lightening conditions. The camera was directed at the participant face during the recording session. This way the experimenter was able to control and view the video presentation and the participant's reactions while being in a separated area of the room. The participant's desk was supplied with a mouse, a 42" monitor, and a foam surface for the comfort of the participant's hands. The presentation of the stimuli on the participant's screen was managed by proprietary application that was developed.

Stimulus presentation environment – For the study, an application with three environments was developed. The first one was a self-rating training environment where participants were introduced to the terms of arousal and valence. During this training period (duration: 15-30 min) we asked the participants to get acquainted with rating these two dimensions using their mouse's pointer on our Continuous Affect Self-rating (CASR) interface (**Error! Reference source not found.** & **Error! Reference source not found.**). The second one was a grey (relaxation) scene where the participants were asked to relax while neutral base-line data were recorded, and the third environment was a semi-dark cinema environment where the videos were presented next to the CASR interface. The participants were asked to minimise their head movements during the recording, to avoid motion-related signal artifacts.

Stimuli sequence – The selected videos (see Section 4.2 of this chapter) were counterbalanced across participants and presented as follows: Five videos from one category, followed by four neutral videos, then followed by five videos from another category and so on, until all videos have been played. The selected original videos had approximately the same duration, however some of them were slightly longer or shorter. The videos were therefore carefully modified to a fixed duration of 27 seconds based on the timing of the affective events happening in each video. Shorter videos were played in a loop. Grey images lasting 8 seconds were added as 'breaks' after every video. The overall duration of the video presentation was 22 minute long. The videos sequence within a category was randomised for each participant.

Figure 29. Video presentation environment (participant's view).
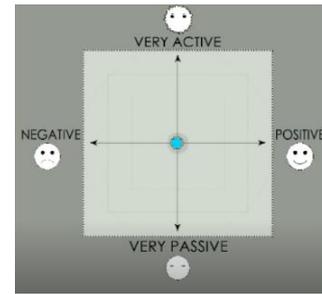
Figure 30. The CASR interface. The blue dot is controlled by a mouse.

Monitoring Equipment and sensors – The software solutions and hardware devices developed for this study are described below. The Faceteq insert prototype (Figure 33) was equipped with eight surface dry electromyography (f-EMG) sensors on the right and the left side of the face (Channels; 1 & 2 on Zygomaticus major, 3 & 4 on Frontalis, 5 & 6 on Orbicularis oculi, and 7 & 8 Corrugator muscles) using an adapted protocol described by [118]. The participants were wearing the insert throughout the study (example of user in Figure 34). The full sensor set-up and the area where the study took place are shown in Figure 35.

Figure 31. The Faceteq prototype used for the study.

Figure 32. Example of participant wearing the EMG sensors, headphones while viewing videos on a screen.

Figure 33. Photographs of the apparatus (right) and physical area of the study (left).

4.3.1.3. **Experiment procedure**

The study took approximately 50-60 minutes per participant at our laboratory at Innovation Centre at Sussex University. Following an introduction and explanation to the protocol, the participants were asked to sign the informed consent form and get acquainted with the set-up. To prevent knowledge of the recording leading to conscious or unconscious changes in facial expression, participants were told that we were monitoring the electrical conductivity of their skin. Questionnaires verified that the participants were not suffering from anxiety, depression or any disorder that can affect their facial movements at the time that the study took place. Prior to the video presentation, participants were instructed how to rate their felt emotions in terms of arousal and valence, through a training session using the CASR rating system controlled by a wired mouse. Example videos were used for this session which were not included in the main study. Once the participants were feeling comfortable and confident using the system, we started the main study. During the study, each participant watched a randomised sequence of the video clips for 22

mins. During this time, they were asked to rate their emotions in terms of arousal and valence using the CASR interface. They were advised to start rating as soon as a video commenced. After each video and during the grey images or 'breaks', participants were instructed to return their rating pointer to the centre of the CASR interface (the neutral area). During the video presentation, video capture of the participant's face and physiological responses were recorded. All sensor data streams were synchronised with the video presentation via the Faceteq API.

4.3.1.4. **Data processing**

The EMG signals from 8 channels were recorded at 1000Hz sampling rate. Firstly, a baseline correction function was applied, by subtracting the mean EMG values. We then removed 50Hz and their harmonics up to 350Hz using Notch filters. The signals were band-pass filtered from 30 to 450Hz. Extreme outliers caused by motion artefacts were removed using a Hampel filter. The clean signal from the 8 channels was then divided in epochs corresponding to the stimuli durations, minus 3 seconds from the beginning of each video.

Next, the Root-Mean Square (RMS) value per 512 samples window was calculated. As EMG are highly variable between wearers, and since we are interested in detecting valence states (negative, neutral, positive) we applied a Maximum-Minimum normalization function [407]. The RMS data were used as input features to train a C-Support Vector Machine (SVM), using the libSVM [408]. For each video and for each participant from the data set, the ground truth was defined by the corresponding participant's CASR valence scores. The data and labels were sent into an SVM (RBF kernel) for classification, using 10-fold cross validation for each participant separately. The two free parameters of the method (C and γ) were optimized for the first participant and fixed for the rest. The low computational cost of the implementation enabled the approach to provide a cross-validated readout in less than 0.5 seconds per participant.

4.3.2. Results

Overall, we tested the feasibility of our prototype for valence detection in VR. The C-SVM enabled us to map the levels of activation of EMG channels with the valenced affective ratings recorded during the video categories for each participant.

A model was created per user, for a total of 34 participants. Each model achieved a category cross-validated classification accuracy ranging from 62.8% to 96.9%, with an average accuracy across the group of 82.5% (Std: 8.2) (Figure 36).
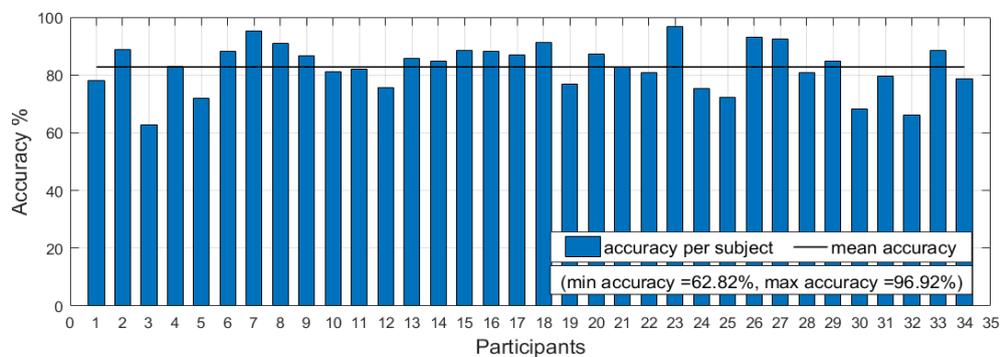
Figure 34. Classification accuracies for all 34 participants

4.3.3.   Discussion

The classification tests showed overall high accuracy in correctly matching the recorded EMG output measures into the three valence categories (negative, neutral and positive) on an individual level using person-specific self-ratings. A C-SVM classifier was cross-validated for each participant, reaching an out-of-sample average accuracy of 82.5%. As these signals were collected using the novel Faceteq hardware prototype, the results of this initial study confirmed the feasibility of our technological set-up to detect three levels of valence. These high accuracy results were in accordance with our expectations as the negative and positive affective states are physically expressed on the face via distinctively different muscles (corrugator, zygomaticus; see section 2.3.2).

Saying that, one of the advancements of this feasibility study is that in this study the EMG signals were collected during more modern video presentations, thus acquiring spontaneous facial muscle activation which is more representative to the affective responses observed in real-world conditions. Other studies often posed expressions which have been predominantly used for discrimination of emotional expressions using EMG signals [50], [67], [309]. Additionally, the self-ratings of each participant per video were used as 'ground truth' for the labelling of the categories for each model (instead of using the same labels for all users), which are found to facilitate increased accuracy rates, as they are suggested to be better suited for ground-truth measurements [409][309].

A second advancement of the study was the introduction of continuous affect ratings. Moreover, the prolonged CASR training session allowed for users to rate their emotions comfortably and effortlessly without looking at the mouse. As result, no major motion artefacts were observed in the data, partly due to the participants limited head movement but also to the careful fit of the sensor mask interface on the face of each user by making slight modifications were made for each wearer. This results in reduced signal-noise ratios during data acquisition. This process allowed us to observe face shape and size differences between participants, which were listed and explored to inform the design of the next version of the prototype. As stated, the prototype sensor mask was tested in highly controlled and motion-constraining conditions to confirm the feasibility of valence detection using neutral and intense positive and negative videos. In real-world conditions however, the VR-user is expected to look-around and explore the environment around her in 360 degrees. The effect of movement on the data was not accounted for in this initial study but was explored in the study reported in Chapter 6.

The next section will report the arousal detection findings from sensors embedded on the same prototype sensor mask. The study followed the same experimental protocol but the data analysis was concentrating on the heart rate responses recorded during the stimuli presentation deriving from an ECG and a PPG sensor. The results from both studies provided further insights for the development of the system prototype, as well as assisted on the refinement of both valence and arousal detection in designed VR applications

4.4. Towards an Effective Arousal Detection System for Virtual Reality (FEDEM 2)

This section describes a feasibility study to explore the effect of affective video content on heart-rate recordings for Virtual Reality applications. In this work, we proposed a system for the detection of high and low arousal via capturing heart-rate responses from the face of the user using a PPG sensor. An ECG belt was used to compare the PPG signal. Continuous self-ratings on arousal from participants were used for the classification of heart-rate responses to low and high arousal states. The low-cost reflected-mode PPG sensor and the ECG chest-belt sensor were attached and synchronised via the Faceteq wearable interface, which was specially adjusted to the needs of this study.

As part of the Faceteq, the PPG sensor was positioned within the frame of the mask. However, since previous research on facial locations for PPG signal acquisition indicated high-susceptibility to head movement [410], the positioning of our sensor and overall fit of mask's set-up was also evaluated through this study. Also, based on our work on EMG detection, we expected that in settings where emotional stimuli are presented, certain muscles of the face would activate, as for example the frontalis muscles during brow raise. This expressive behaviour of the face could result to skin movement which would impose additional noise artifacts on the PPG signals. Although the majority of research on heart-rate detection from PPG on the forehead focused on recording data from non-emotional settings, our team investigated the feasibility of HR detection from PPG on the face during stimuli-generated elicitation of affect and spontaneous facial expressions. The same experimental procedure was followed for this study as described in Section 4.3. Videos with affective content were used to induce five affective variations of arousal and valence, namely high-arousing positive, low-arousing positive, high-arousing negative, low-arousing negative and neutral states (see section 4.2).

Although ECG recordings have been primarily used for HRV monitoring due to its distinct profile of R peaks (Figure 10), there might be many advantages of measuring Pulse Rate Variability (PRV) from the PPG. The PPG is easy to use, non-invasive, cost-effective, it involves less sensors and it enables continuous, long-term recordings [411], [412]. However, PPG signals are easily susceptible to movement artefacts and the detection of R-R intervals from arterial pulses from distant sources (e.g. fingertips or legs) could potentially be erroneous [413]. Researchers have

explored the improvement of pulse rate estimation from reflective PPG sensors, by utilising accelerations data (three-axis) to remove motion artifacts [414]. PPG sensors have been utilised in a large variety of experimental studies and on numerous body locations, including fingers, hands, forearms, earlobes, wrists, auditory canal, legs, but-tocks, and the back [411]. Researchers have also recorded reflected PPG signals from the forehead using an elastic band for heart-rate and HRV analysis for the monitoring of soldiers in the battlefield [415] and neonatal patients [416]. PPG forehead placement showed advantages over other peripheral body location because it offered greater sensitivity to pulse changes during low blood flow [417], and because it was less susceptible to motion artefacts during certain body movements [418].

PPG sensors have also been utilised in VR research. Besides placing the sensors on common body-locations e.g. fingers [419], several attempts have been made towards facial placement and HMD incorporation, e.g. middle of the forehead using a headband [420], and directly placed on the face plate of the HMD [421]. However, the quality of the signal was not evaluated for these approaches. As PPG measurements could be susceptible to changes in light perfusions and movement artefacts, we envisaged that by incorporating PPGs on an interface between the HMD and the user's skin, we could obtain a clear pulsative reading for reliable arousal detection in VR. Hence, we took this PPG placement approach for this study.

The objective of this study was to test the feasibility of arousal detection with a PPG sensor placed on the superficial temporal vein. We also wanted to explore its performance efficiency when compared to an ECG (conventional method) and to the combination of both modalities, in affective video watching settings. Our prediction was that PPG reading when adjusted on a wearable face frame superimposing the superficial temporal vein, will be less susceptible to movement generated from facial expressions, and thus provide useful insights on the level of physiological arousal of the wearer, comparable to the ones provide from ECG readings.

4.4.1. Methods

4.4.1.1. **Participants**

For this study, the same participant pool was used as per the previous study. Due to technical issues occurred during data collection, the heart-rate data from both PPG and ECG sensors were only collected from the last 15 participants. Data from further four participants were excluded from the analysis after data quality assessment. We observed noise artefacts on the PPG steams related to wrong sensor placement and potentially facial movement during the data collection. For two participants the ECG belt could not be well fixed to the participants' chest size. As a result, data from 11 participants were used (referred to as P1-P11; 5 female and 6 male), with a mean age of 21.5 years (±2.6). Participants did not suffer from anxiety, depression or any disorders of cardiovascular nature which could affect their heart-rate metrics at the time of study Each participant was compensated by a £5 voucher for their time. The study was reviewed and approved by the Research Ethics Panel of Bournemouth University (reference ID: 12025)

4.4.1.2. **Materials**

The software solutions and hardware devices developed for this study are described below.

The stimuli presentation – The stimulus presentation environment was the same as the one use in the previously reported FEDEM 1 study (section 4.3.1.2).

Monitoring equipment and sensors – The interface prototype was equipped with a custom-made PPG sensor (reflection mode) on the upper left side of the mask (see Figure 37), corresponding to the area over the superficial temporal vein and artery. Additionally, a custom-made ECG chest-belt was developed comprising two ECG sensors, which were connected to the Faceteq™ insert. Both data streams were recorded simultaneously.
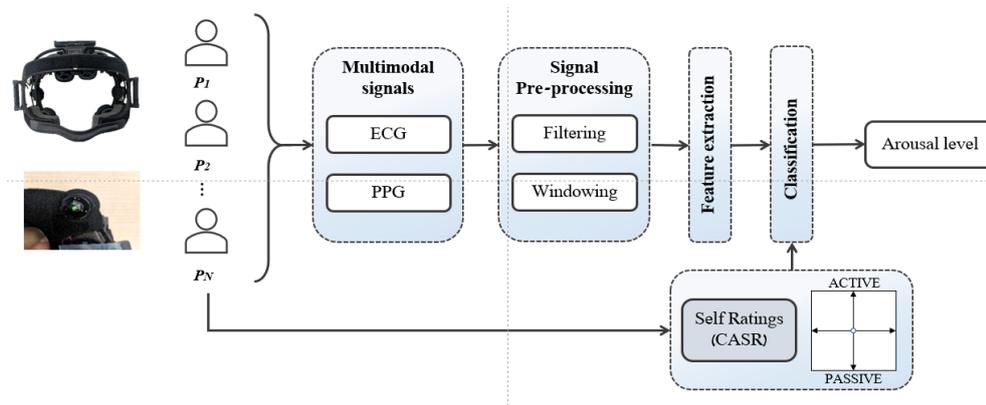
Figure 35. High-level overview of the proposed system. The Faceteq prototype is depicted on the upper left side.

4.4.1.3. **Experiment procedure**

The experimental procedure was the same as the one use in the previously reported FEDEM 1 study, explained in section 4.3.1.3. Specifically for this study, participants were asked to avoid caffeinated drinks on the study day, as caffeine has shown to trigger increases in blood pressure and heart-rate [422], and also decrease the eye-hand coordination [423]. The latter which could affect the continuous self-ratings was an important element of the study, described earlier in sections 3.3 and 4.3.1.3).

4.4.1.4. **Data processing**

Signal pre-processing – PPG and ECG recordings from all participants were recorded using Faceteq API (sampling rate: 1000Hz). The analysis steps were as follows: First, the recorded raw data were filtered (Notch filter: 50Hz; band-pass Butterworth filter: 0.5Hz and 6 Hz for the PPG, 5Hz and 25Hz for the ECG; order: 2). Subsequently, the filtered recordings were divided in 25 seconds long time-window epochs corresponding to each video stimulus, excluding the first 2 seconds. Each epoch was further subdivided into 4.5 seconds over-lap 5 second windows. Next, a peak detection method was applied on the PPG and ECG epochs, to identify the R-peaks. The distances between peaks were calculated for all the detected peaks within each time-window.

Feature extraction – The mean peak distance (IBImean) and the Root-Mean Square (RMS) of successive R-R interval distance (RMSIBI) per epoch were calculated. The whole feature vector was transformed based on the Minimum-Maximum normalisation [407] for each participant. The total number of processed samples per video-length was 48 per metric, resulting for 20 videos to a total number of 960 samples per participant per metric.

Arousal Classification tests – As the automatic state recognition can be constrained by individual user differences, two scenarios were explored: a user-dependent and a user-independent approach. The following tests were performed using: (1) the PPG derived metrics, (2) the ECG derived metrics and (3) the combination of both PPG and ECG derived metrics. The two outputs (IBImean and RMSIBI) per modality were used as input to train a C-Support Vector Machine (SVM) using a gaussian kernel. The open-source libSVM framework [408] was adopted to train the binary C-SVM.

Table 7. Agreement scores across users per video (mean values, standard deviation and Coefficient of Variation).

Video	M (Std), CV	Video	M (Std), CV	Video	M (Std), CV
1	0.7 (.73), 0.9	8	1.0 (.49), 0.5	15	1.1 (.88), 0.8
2	0.4 (1.00), **2.5**	9	0.6 (.74), **1.2**	16	1.4 (.45), 0.3
3	1.2 (1.24), **1.0**	10	0.5 (.94), **1.8**	17	1.5 (.85), 0.5
4	0.2 (.65), **1.0**	11	1.6 (.81), 0.5	18	1.3 (.92), 0.7
5	0.7 (.86), **1.2**	12	2.0 (.90), 0.4	19	1.7 (.86), 0.5
6	0.8 (.62), 0.7	13	1.2 (.73), 0.6	20	1.4 (.84), 0.6
7	0.7 (1.0), **1.4**	14	0.8 (.66), 0.8		

In the *user-dependent* classification scenario, a 10-fold cross-validation was applied for each participant separately. In the *user-independent* scenario, we applied leave-one-participant-out cross validation by pooling all 11 datasets and predicting the rating of each participant in turn based on the remaining 10. The two free parameters of the method (the regularization penalty C and the standard deviation of the kernel function) were optimized exclusively on the training data for both dependent and user-independent scenarios.

The corresponding arousal CASR ratings per participant were used as the ground truth definition. Self-reported scores across users in terms of arousal for each video typically showed low dispersion (coefficient of variation, CV, **Table 7**) indicating high rating agreement per video across participants (the lower the CV

score, the higher the between-users rating agreement). Videos which presented rating disagreement were part of the positive emotion-inducing categories (videos 1-10), indicating higher rating variability during positive content.

The mean CASR value per participant for the user-dependent, and the mean value across all participants' ratings for the user-independent scenario were utilised as the division point for the of high and low arousal classes. The total number of samples per participant was 960 (total: 10560s). The mean number of samples for high arousal levels for the user-dependent scenario is 532 ± 20 (mean $\pm$ SEM; ranging from 368 to 602 samples); and 5852 for all participants (P1-P11). Thus, the distributions of the high and low arousal classes were largely balanced, for both user-dependent and user-independent scenarios.

4.4.2. Results

We tested the feasibility of arousal detection via PPG sensor from the superficial temporal vein in VR. The C-SVM enabled us to map the level of arousal with the metrics calculated from the PPG and ECG recordings (IBImean and RMSIBI) during the presentation of four audio-visual stimuli categories. The results from the classification tests performed for the user-dependent and user-independent approach are presented below.

User-dependent scenario – In Figure 38, the receiver operating characteristic (ROC) curves per experiment performed are illustrated. The ROC curve is largely used as a graphical method to show the diagnostic ability of binary classifiers, and it is created by plotting the true positive rate (the correctly predicted labels for one class) against the false positive rate (the incorrectly predicted labels for the same class) of a classifier model. Generally, classifiers that give curves closer to the top-left corner indicate a better performance. In the ROC figures presented below, each line represents a participant. The areas under curve (AUC) are included on the bottom right corner of each plot. The AUCs here are used to summarise the performance of the each classifier, as a general measure of predictive accuracy considering all decision thresholds [424], since in the goal of this section is to establish a broad comparison between both recording modalities.. Despite the variations in performance between participants, the system's capability observed for

detecting changes in arousal is higher for the fusion versus unimodal approaches for most of the participants (Figure 38c).

(a)PPG   (b)ECG   (c)Fusion
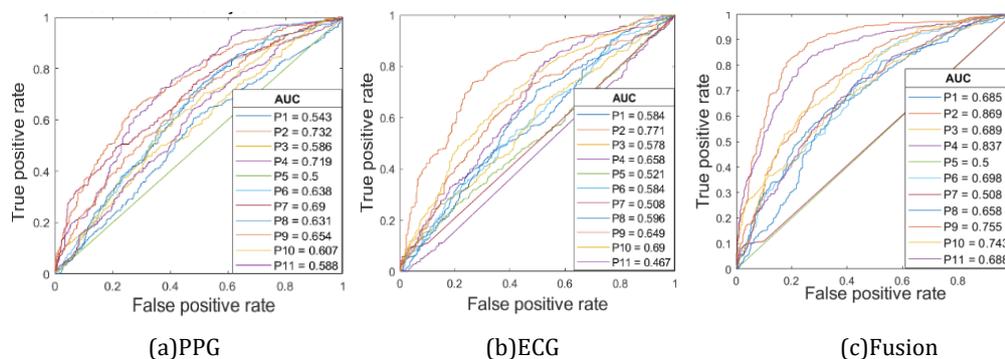
Figure 36. ROC curves per participant per experiment performed (user-dependent).

To evaluate the performance of each classification experiment, we compared the AUC per metric in recording modality pairs (e.g. PPG against ECG) using the two sided Bradley's test at the significance level $\alpha = 0.05$ [424], [425]. The results from this test are reported on Table 8. We denote in bold when their AUC means are significantly different (see details in [424]). Detection performances between PPG and ECG modalities were significantly different for six out of 11 participants (see participants 2, 4, 6, 7, 10 and 11, $p<.05$). Moreover, the combined PPG-EEG metric (termed here the fusion approach), outperforms the PPG and ECG modalities individually for 9 out of 11 subjects (Table 8).

Table 8. Bradley scores between experiments per participant

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
PPG-ECG	-1.5	**-1.8**	0.3	**2.5**	-0.7	**2.1**	**7.2**	1.3	0.2	**-3.3**	**4.6**
PPG-Fusion	**-5.6**	**-7.0**	**-3.9**	**-5.7**	0.0	**-2.4**	**7.2**	-1.1	**-4.3**	**-5.7**	**-4**
ECG-Fusion	**-4.0**	**-5.3**	**-4.2**	**-8.3**	0.7	**-4.6**	0.0	**-2.4**	**-4.5**	**-2.3**	**-8.8**

Significance level at 5% | *significant values are indicated in bold

User-independent scenario. The system showed a similar ability to discriminate high and low arousal levels from ECG, PPG and from the fusion of the two modalities (Figure 39). The AUC values for each experiment are included on the bottom right corner of the plot. The performance of the system when using only the PPG metrics was significantly lower than ECG and fusion. Finally, the best overall performance is achieved via the fusion approach (significance level at 5%), Table 9.

Table 9. Two-sided Bradley scores between experiments performed.

PPG-ECG	**-4.3**	PPG-Fusion	**-9.1**	ECG-Fusion	**-4.**

significant values are indicated in bold
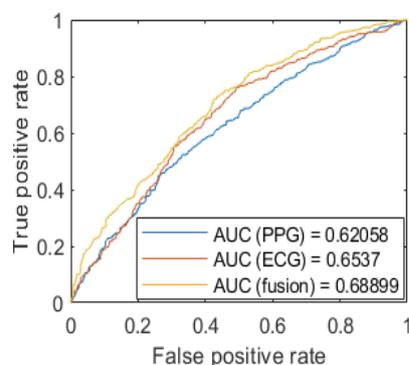
Figure 37. ROC curves per experiment (user-independent).

4.4.3. Discussion

We proposed a system for arousal detection in VR settings, by designing a novel interface which incorporates PPG and ECG sensors. During the study, participants facially expressed their emotions in response to video stimuli that resulted into limited head, and therefore sensor movements. Nonetheless, using the PPG-derived metrics, our system yielded a similar detection performance to the ECG-derived metrics for 5 out of 11 participants in the user-dependent scenario. This result supports our assumption regarding the system's capability to detect arousal via PPG recordings from the superficial temporal vessels (on the face), subject to individuals' variability. Moreover, the fusion of both methods provides an enhanced performance overall, which was expected as the fusion of physiological sensors is highly considered for the reliable detection of heart-rate, increased accuracy and reduction of incorrectly detected classes cause by noise or motion artefacts [426], [427].

The arousal detection issues that occurred in all three classification experiments for participant 5 suggest that changes in heart rate during audio-visual stimulation are induced and elicited in different intensities among individuals. Thus, the detection capacity of the system was less reliable. Additionally, detection issues could have resulted from bad sensor placement, sensor's quality, intense movements (like during laughter) which could also reposition the sensors, or to skin sweatiness.

Overall, the system's detection using PPG in the user-independent scenario performed slightly worse than using ECG. This result could be attributed to the observed signal quality for this participant sample, and to the ability of PPG sensor

to detect heart-rate changes during emotional expression. In terms of signal quality, the signal from the PPG sensor for some participants was clear, demonstrating prominent peaks throughout the recording, while for others the peaks observed were weak and often mixed with noise related artifacts. Therefore, the position proposed in this study for the PPG sensor could be selected with caution, following careful adjustment of the mask and sensor per user before data acquisition, while also taking into consideration the participants' face shape, and the movement of the face during emotional expression. Alternative facial positions for the PPG sensors were explored and used in the next studies. Although PPG-derived metrics achieved lower accuracy than ECG, PPG sensors are affordable and easy to use, making them strong candidates for wearable integration in practice. Likewise, although ECG sensors are difficult to integrate at the moment, we envisage that improved ECG sensors will be readily available for integration with wearable devices and clothes in future. Thus, given the enhanced performance for the fusion set-up demonstrated in this feasibility study, the combination of both sensors for arousal detection seems a robust approach for multiple applications incorporating immersive technologies.

4.5. **Chapter Discussion and Conclusions**

In this Chapter, the first studies to evaluate the feasibility of the system designed (including affect induction paradigm, the novel sensor set-up and the physiological data processing flow) to detect valence and arousal as parameters of affect are presented.

Prior to the data collection, an affect induction paradigm was designed. A list of audio-visual stimuli materials from a recently published database (at the time of the study) were carefully selected following a validation survey with 82 participants. The video clips were rated in terms of arousal and valence and were distributed into five discrete affective areas based on the linear combination of the valence and arousal polarities including neutral (i.e., four affective quadrants and one neutral area). This design allowed to study the effects of high and low arousal on our sensors, together with the effects of negative, neutral and positive valence. As part of the selection process, the familiarity of the target age group with the videos was also assessed, and videos which were widely known or viewed before were excluded. This re-validation survey on the selected videos per category permitted us to confirm the ability of certain videos of the positive category to induce the expected levels of positive responses to our target age group, due to the depicted context (e.g. babies), and to identify and exclude extreme negative videos. The video library created as a result from this survey was used for the subsequent feasibility studies.

The two studies using EMG for valence detection and PPG-ECG for arousal detection allowed us to evaluate positively the feasibility of our affect detection approach in controlled laboratory conditions. A video stimuli presentation experimental process was adopted, in order to induce spontaneous affective responses to participants. In the first study, the detection of positive, neutral and negative valence was achieved with a mean out-of-sample accuracy of 82.5% across participants, following a user-dependent analysis and classification approach. In this approach, a classification model was created per user, and the self-ratings on valence per user for each video were utilised as ground-truth labelling of the classes.

The same classification approach together with a user-independent one (where all users' data were pooled together) was followed in the second study, in which we explored the ability of the system to detect high and low arousal from a PPG sensor placed on the mask which covered the superficial temporal artery and vein, as part of the novel mask prototype. An ECG sensor was also simultaneously

used during the study, which is considered as the 'golden standard' sensor for heart-rate detection. Although the sample size of this study was small (11 participants), the results for the classification tests showed that the overall ability of PPG to detect arousal is in most cases slightly worse than the ECG. However, the fusion of both modalities for arousal detection proved to be advantageous against the unimodal approaches, consistent with previous relevant studies [367].

The results of the data analysis for valence and arousal detection, coupled with the observations and experience gathered by conducting these studies with multiple participants offered insights to potential limitations related to sensor placement (esp. for the PPG sensor), the materials, and the characteristics of the proposed system. This knowledge was used to improve the equipment, experimental protocol and analysis pipeline of the third study described in Chapter 6. Specifically in these studies, the application of the sensor set-up on individuals outside the research team showed some minor problems with sensor placement due to variant head and face sizes, skin folds and face curvatures [387]. The empirical observations combined with the need for careful signal quality assessment for each user informed the development of the next generation of the Faceteq mask interface, and the 'EmteqVR_app' software for real-time signal monitoring. We envisage that a larger sample size and the enhanced technical abilities of the new prototype would enable the system to achieve higher accuracy on valence and arousal detection, which would allow for the development of a more robust affect detection model. More specifically, we aim to enhance the utilisation of the detection common facial patterns in clusters of participants for a more accurate valence detection, and a better PPG placement for more accurate arousal detection.

In both aforementioned studies, the controlled nature of the experimental procedure allowed for the reduction of expected noise-artifacts derived from the head movement of the users, but we did not account for facial muscle movement as the result of emotional expression induced by affective stimuli. The effect of emotional stimulation on facial muscles was expected and measured using the f-EMG sensors. The main objective of building the prototype mask was to combine multimodal measures for detecting both dimensions of affect. Thus, the simultaneous detection of heart-rate and facial movement (for arousal and valence) was paramount, and therefore the noise on the PPG signals caused by facial movement was anticipated.

The highly controlled experimental paradigm adopted for these studies allowed us to explore the sensors sensitivity to physiological changes caused by

affect-related variations, while maintained a clear connection between the stimulus (as the origin point of affect stimulation) and the resulted impact on the signals recorded. This clear link between cause and effect is of great importance in order to understand not only the physiological outcome (the 'symptoms' of an affective state), but understand *what* caused it, and perhaps even combine them to predict further changes in physiology. This context-related approach could provide a path towards the application of componential appraisal models for emotion detection [67] and assist future automatic state recognition (see section 2.3.1). Ensuring the direct connection between stimulus and effect in VR was further explored for the development of the virtual stimuli scenarios utilised in the next study (Chapter 5).

As part of the next step in feasibility for affect detection in VR settings, the same affect detection approach was applied in active VR settings, using a commercial HMD. Naturally, the motionless passive nature of the first feasibility experiments, does not represent the every-day usage requirements of the highly interactive nature of immersive VR. Indeed, in VR, the user is usually free to experience the virtual space in three dimensions and look around in 360º (depending on the capacities of the immersive technology used). In the next study our stimuli consisted of virtual 3D-spatial environments, specifically designed for creating a fully immersive set-up. This setting could be experienced using motion capture and head tracking, enabling us to explore further the capabilities of our improved system to detect affect in VR settings. We wanted to record naturally occurring affective states within fully immersive VR scenarios, similar to contemporary VR experiences that are available on the market. Inspired by research on the feeling of presence and immersion (see section 2.3.3), we decided to develop a VR experience that allows the user to explore the room naturally. That is by simply walking around, without the need of a controller or a mouse for directing their movement. The virtual environments were custom-designed and programmed to track and control the affective interactions with virtual stimuli and their parameters within VR (i.e., to know the origin point of the stimulation, what the participant is focusing on). Chapter 5 introduces next the design, development, and validation of these virtual environments.

Summary of key findings for Chapter 4

- Online survey (Section 4.2): Validation of a selection of 40 affective videos, i.e. 20 positive and 20 negative videos, as well as 20 neutral videos. These can be used for further studies.

- Feasibility of negative, neutral and positive valence detection (Section 4.3): Valence detection was feasible as evidenced by the user-dependent (using a 10-Kfold CV) classification accuracy of 82.5% (SD: 8.2).

- Feasibility of low vs high arousal detection with PPG vs ECG sensors (Section 4.4): There was a high inter-subject variability and the signal to noise ratio was low (5/11 subjects PPG comparable to ECG). This finding led to mask design.

Chapter 5

Development and validation of stimulus material for an affective VR study

5.1. Designing affective interactive VR environments

This chapter is describing the development and validation of the VR environments used for affect induction in the VR study. This VR study used the EmteqVR equipment to detect arousal and valence and it was recorded in the Science Museum in London. More details on this study, and specifically on the data analysis using classifier are reported in Chapter 6. The introduction will explain the rationale behind the choices made for the VR environment development, describes how we marked specific events within the VR scene, and practical solutions for the experimental protocol. Afterwards, findings from an online survey conducted on 67 participants will be reported. As an overview, the survey was developed to evaluate videos and images of the virtual environments as well as specific events that were programmed

for the VR scenes. Participants were instructed to rate their perceived level of valence and arousal, the memorability of each stimulus, and their levels of presence for each VE video using a Presence questionnaire. Additional questionnaires about personality and alexithymia were given to the participants at the end of the survey. The user ratings validated the stimulus material selected for each VE and assisted a better understanding on the affective impact of each individual stimulus/event on participants. As stated above, the findings from the online survey (see section 5.4) were used to inform the experimental setup of the study reported in Chapter 6.

As a recap, virtual reality offers a flexible tool for the construction of affect evoking situations and environments, similar to those depicted in popular libraries of affective images or videos (e.g. IAPS [405]). Actually, this tool might be more powerful than the usually used conventional stimuli as images and videos. In fact, a progression from using still images to videos has been observed in past research. Gross and Levenson [428] said that "Films also have a relatively high degree of ecological validity, in so far as emotions are often evoked by dynamic visual and auditory stimuli that are external to the individual" (p. 88). Properties such as the screen size, colour, motion, audio incorporation, [355], [429]–[432] found in videos were suggested to influence the potency of the experience thus increasing the intensity of responses compared to still images. Today, immersive VR technologies offer high-resolution display properties, audio incorporation, high framerate for movement synchronisation and the animation of features in the virtual space, while also allowing the design of interactive content. If videos can therefore outperform images in emotion elicitation, can immersive interactive VR environments outperform videos and which features do they need to achieve this?

Researchers argue that real-like experiences can evoke naturalistic emotional responses, and so, an immersive simulation can provide the host environment for real-like experiences [433]. One of the main differences between non-immersive and 'fully immersive'[3] (using VR) experiences is the level of presence they can elicit to the user, which is related to the type of interaction that the user has with the content and the coherence of body actions with spatial VE structure [332]. Other factors explained by Singer and Witmer [331] can be categorised as control and sensory based, such as the immediacy of control, mode and degrees of

---

3 Term used according to Kalawsky [561], who specified as 'fully immersive' a VR system presenting a 360o information space on a display.

control, action responsiveness and environment modifiability. By enriching these 'interactivity' factors in a VE, we can increase immersion and in theory increase presence (see sections 2.3.3 & 2.4).

In other words, in conservative affect elicitation mediums, the user is asked to passively observe a stimulus / scene or listen to a sound in a controlled linear fashion, whereas in the immersive medium scenario the user 'steps' inside the space of the image or video, surrounded by it. Within this intricate, sensitive system, a VR user can experience presence [434], allowing themselves to get absorbed by the virtual world, and respond to its content possibly in a naturalistic manner. Perhaps, the effect of immersive stimuli derive power from not just the function of the enabling technology (immersion), but the fidelity of the simulation to the extent that it covers the subjective expectations of the users within it; the link between proprioception and sensory data on a physical (e.g., synchronisation of motion tracking and display) but also a conceptual level. For example, if a user is holding a cup full of coffee and she lets it fall on the floor, coffee should be spilt, or if a fire starts, we expect to see smoke and feel its warmth.

Immersive technologies can provide the tool for the creation of real-like or ecologically valid, affect-inducing experiences which can be implemented in controlled laboratory conditions. Indeed, nowadays more and more researchers agree on the effectiveness of Virtual Environment for emotion induction [171], [199], [273], [282], [433] . However, would it be misleading to assume that not all immersive experiences have this potential and that certain parameters need to be met to succeed? Perhaps the factors for successful immersion, presence and emotional elicitation are linked. The level of interactivity and 'physical interaction' in the VE, can enhance the immersivity of a simulation and create the powerful sensation of physical Presence [183]. In turn it has been suggested the level of immersivity induce significantly higher self-reported arousal responses [171] and that the level of presence is linked with intense emotions [218], [280], [435].

Unfortunately, the existing research on emotion elicitation using interactive immersive content is scarce. Additionally, immersive content (i.e., 360 videos, also see section 2.4.2), is rarely actively experienced in research studies, to the full potential offered by the current commercial VR technologies. To be able to progress in this area of research and development, our team was faced with two main challenges. The first challenge was the lack of existing libraries with pre-validated affective stimuli in highly interactive and immersive VR. Secondly, the existing 3rd

party applications and VR experiences were not designed to allow for the controlled study of affective responses elicited in VR, lacked tools and solutions to successfully evaluate the link between the cause (stimulus) and the effect (emotional impact) required for the nature of this research. Consequently, our team decided to design and develop the virtual experiences required for the adaptation of our affect detection approach in VR from scratch, giving us full creative and experimental control over the properties of the virtual environments (VEs). The ability of those environments (and their elements) to induce the targeted affective impact was validated using an online survey which is described in section 5.3.

As a result, four affective VEs were designed; one intended to relax the user (used for baseline recordings before entering the affective scenarios), and three indoor VEs for the induction of neutral, positive, and negative valence. The three indoor VEs shared the same room structure as the virtual replicas of a real office room. Existing literature on the effect of low-level audio-visual features (e.g., brightness, colour hue, sound manipulations) informed the design of the VEs. The environments were populated with 3D objects (as stimuli), which were designed to enhance the overall targeted affective impact for each VE. Those objects (here are refereed as 'events') had various sizes and attributes. They were placed within the virtual rooms, in various locations, taking advantage of the overall virtual space and structure of the room replicas, thus allowing the user to experience them by freely exploring the rooms. Some of those stimuli were static objects while others were animated. The activation of animated objects including sounds (the so-called 'interactive *events*') was programmed to occur based on certain predefined criteria (i.e., based on the user's interaction and time spent in the scene).

An event-marker system was developed to annotate in real-time the times when each event was actively triggered and made visible to the VR user. This system was used to synchronise the timings of activated events with the data recorded from the physiological sensors. Another key feature of the VR simulation was the development of a gaze-based interaction system which was applied to track the stimuli viewed in real-time by the users. This system was completely invisible to the user, thus allowing them to explore the virtual rooms freely, at their own pace and volition. This custom-made gaze-based interaction system combined with the event-marker system provided the tools controller non-linear stimuli presentation in VR.

The description of the environments developed together with the stimuli tracking and event-marker solutions are presented in the next section. The ability of

those VEs (and their elements) to induce the targeted affective impact on our target age group was validated using an online survey, which is described in section 5.3.

5.2.   **Description of the Virtual Environment**

In this section, the design of four 3-D VEs is described. The environments were named 'Baseline-relaxation scene', 'Neutral scene', Positive scene' and Negative scene'. The software used for the development were Unity3D game engine [352] to design the VE scenarios, and Autodesk Maya 3D [436] for the design of the 3D objects.

The first VE was a **baseline-relaxation scene** using a 360° underwater environment and a smooth water audio-track (Figure 40). Participants of the main study would be entering this scene before each VE, for at least 2 minutes per visit. This environment was designed to allow the participant to relax before entering one of the main affective scenes, thus decreasing the physiological arousal which could be elevated due to the novelty of the media experienced [396]. The idea for the introduction of the water element of this VE was inspired by research on the restorative effects of aquatic environments [437], [438]. We expected this environment to evoke low arousal and neutral/slightly positive valence levels. The addition of aquatic elements, and marine biota was avoided in order to control for high positive valence ratings [439].

Figure 38. Screenshot from within the VE used for relaxation and baseline recording in pre-study survey.

The **remaining three VEs** were based on an existing office space (see photos of details of the actual room used in Appendix C Appendix*: Study Materials'). The VEs were mapped according to that space and populated with virtual counterparts of the physical objects. The virtual office room was 2.3m width x 2.90m length x 2.20m

height with an allocated walking area of 1.6m x 2m which was consistent between all VEs. The dimensions and the basic synthesis of the space in those VEs was kept identical. In all settings, the room contained a bookcase, two office desks with chairs, a window, lights, two PCs with monitors, a small cupboard with a printer, a garbage bin, a mirror and two paper notebooks. The individual configurations for each VE were adjusted to evoke specific affective responses. In the positive VE and the negative VE, several parameters were altered to evoke either positive or negative affective responses, based on the low-level visual modifications and the integration of static and interactive objects (see Figure 41). The neutral VE was created to evoke a neutral mood with low arousal levels. These configurations are discussed per VE next.
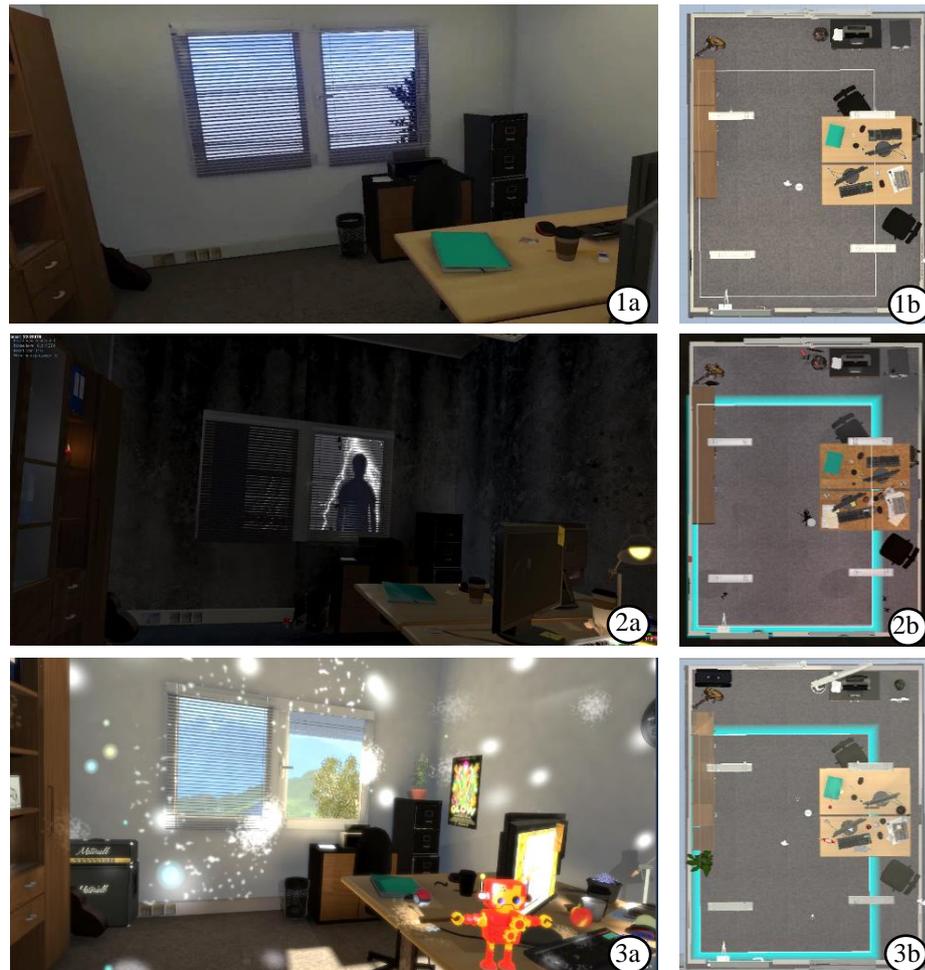
Figure 39. Screenshots taken of the three affective VES from the user's view (a, left side) and top view (b, right side). The first one (1a & 1b, top) is from the neutral environment, the second (2a & 2b, middle) is from the negative one and the third (3a &3b, bottom) is the positive VE. The square areas outlined in the top views (1b,2b,3b) show the perimeter of the walking area of the user.

Neutral VE. The neutral environment contained all static basic objects of the office synthesis without the elements that were designed for negative or positive valence elicitation. The colour palette and temperature was kept in grey and cold faded tones, with low contrast to reduce the possibility for increased physiological arousal [294], [440]. No audio samples or interactive events were planned for this VE. The lighting conditions were soft and dimmed with smooth shadows. This way the room was not strongly lighted or bright but completely visible for exploration by the user. A roman blind was designed in front of the window, to reduce the incoming light from the virtual sun embedded in the scene. The space outside the room, visible through the window was set to a grey, cloudy view with faded colour detail.

Negative VE. The colour palette for this room was set to intense, dramatic contrasts. The overall atmosphere was inspired from two scary/horror games in VR: 'Resident Evil 7 VR' and 'Affected: the Manor' [441], [442]. The main lights were direct, switching on and off as if they were faulty using a custom-built switch with short, randomised time intervals. The walls and floor were covered with an additional material resembling of dirty, unpainted concrete. The synthesis of lights and textures was based on the design experience of the research team. Multiple additional objects (event description in section 5.2.1) were placed inside the room, including interactive and animated spiders, stressful notes, a ghost-like figure in the mirror, an animated shadow-figure appearing outside the window, an animated rat, a candle, litter placed around the bin, a fire trigger and an alarm. The majority of those events were triggered by the gaze of the user (see description of the gaze-based interaction in Section 5.2.3), e.g., a shadow outside the window or a ghost face in the mirror. Others like the 'spider attack' event was be triggered once, attempting to jump-scare the user. The rest of the spiders were activated throughout the whole VE experience by slowly follow the users gaze in the room, climbing on walls and main virtual furniture. Once the user had spent 65 seconds in the scene, a fire was triggered, and 10 seconds later the fire alarm went off requesting the user to head towards the exit and leave the room. The fire alarm included a bright, red light circling around the room in quick intervals and a loud siren. These events were designed to increase the physiological arousal towards the end of the experience and intensify the negativity of valence. Audio were incorporated in all interactive objects, including light bulbs, the ghost event, the shadow (lightning), the rat, the fire alarm, and the spiders. Some of audio were downloaded from the free sound audio library [443] .

Positive VE. Similarly, all audio-visual parameters were set to provoke pleasant feelings with variations of arousal for the positive environment. Multicolour synthesis was selected, including bright tones with intense colour hues and saturation. All lights were set brighter across the whole room and an intense sun light was designed to enter through the window. The roman blinds were not covering the view anymore allowing for the user to look outside the window. Along with static objects including posters, post-its, an apple, and pictures, multiple interactive events (described in Section 5.2.1) were programmed to be activated in different time ranges based on the gaze of the user. These interactive events were: a flock of

butterflies entering the room through the window, birds flying outside the window, fairy lights or star dust inside the room, a robot dancing on the table, webcam feed of the user on the mirror, flowers growing, a plant moving and an interactive videos of a goat on a picture screen. Again, audio was incorporated in all interactive events and animated objects. Approximately after 75 seconds in the room, and if the user was not engaged with another event/object, the windows would open allowing laughter sounds to enter the room.

5.2.1. Affective stimuli: Interactive and static events

Table 10 shows the list of events and objects per scene used in the main study. Screenshots of all objects and events are available in Appendix C. Videos of the rooms and videos/pictures of the events were sent for evaluation via an online survey to participants (section 5.3). In this section, the overall interaction design and how the events can be triggered within the VEs will be described.

Table 10. List of objects and events for each virtual environment.

	Negative	Neutral	Positive
1	Fire Alarm	Bookcase	Green plant
2	Documents	Clock	Baby poster
3	Window - Lightening/silhouette	Green Notebook	Light explosion
4	Glitch in viewpoint	Grey Notebook	Reflection in mirror
5	Fire	Guitar	Dog poster
6	Overflowing bin	Window	Butterflies
7	Flickering light (bulb fusing)	News board	Robot
8	Spooky mirror	Calendar/Cup	Monitor message
9	Spiders in room	Computer Mouse	Stardust (Light particles)
10	Light (bulb exploding)	Desks	Guitar
11	Spilt drink (cup)	Bin	Flower
12	Rat	Mirror	Birds
13	Spider attack	Carpet Floor	Amplifier
14	Spooky music	Monitor	Beach ball
15	Candle/skull*		Goat picture*
16	Office room*		Backpack*
17			Pokemon ball*
18			Office room*
19			Window (mountains)*

Items marked with (*) were added later on and thus were not included in the survey

Every detail in the 3D VE room was designed and positioned manually Hence, all objects and their features (such as animations) can be controlled by the programmer.

Current available VR experiences with tasks like games or creative applications like painting in 3D, allow the user to interact with 3D elements and control them in real-time. Similarly, in this study, we designed the VEs to allow free-walk exploration in 3D rooms, as a more natural, innate way of environment exploration rather than using controllers and joysticks. This 'naturalistic' way of navigation the virtual space could promote and enhance increased feelings of immersion and presence (see section 2.3.3). As the affective stimuli were positioned in context at different locations of the room, and since the users' gaze is completely dependent on their own body movements, stimuli were activated when visible to the user or when the user is directly looking at them (using a gaze-based technique explained in section 5.2.3).

For this reason, all objects and features (together we call them 'events') had an 'interactive marker' which was activated when once of the following conditions are met: (a) the gaze of the user is directed towards the object, and the fixation duration is larger than 2 seconds, or (b) the time passed since the start of the VE experience (based on a predefined sequence of event activation per VE). Once the pre-defined conditions for each stimulus were satisfied, an interactive marker was added with an 'ID', a number corresponding to the activated event. Up to three events could be active at the same time, including two events activated separately by the user's gaze and one event activated by time passed. The event markers allowed us to also track the interaction periods on the data stream, by exporting the timestamps of the activated events together with the rest of the recorded data (physiological signals and movement data).

5.2.2. The task and area of interaction

Figure 42 shows an aerial view of the virtual office. The starting position of user was the same across all VEs (point A) looking towards the other end of the room (point B.) The area was designed to allow the user to approach certain areas and interactive objects while also avoid some others. For example, starting point A, the user could look to the right and approach the chair and the mirror on the wall, and would avoid walking onto the virtual tables as they would have perceived as physical obstacles. The desks were replicas of the actual desk of the experimenter, which also placed in the same area in the physical world. The idea behind it was to make the overall experience more believable by giving some of the expected proprietary haptic

feedback. For example, when the user could touch the virtual desk, she could actually feel the real physical desk.
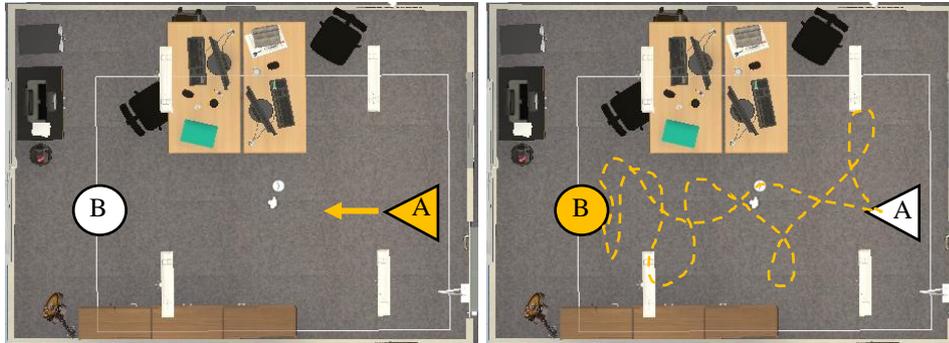
Figure 40. Top view of the scene used for the VEs. The write rectangle shows the perimeter of the user's walking area. The user would start from point A (left figure) then to point B while exploring the room, and back to point A (right figure) to exit the scene.

5.2.3.   Gaze ray-casting for event marking

In order to find the objects that are in the user's point of view in real time, an invisible ray was casted from the middle of the user's point of view, facing forwards. The function '*Raycast*' from the '*Physics*' library in Unity Engine [444] was used. Since the room was overall small and the field of view set at 60º wide, there were only up to 4 different interactive events visible at the same time. We prioritised the objects/events that were in closer proximity to the user's position and the ones that were actually visible to the user, i.e. not hidden behind another object such as a monitor hiding the poster on the wall or the flowerpot on the desk (see example Figure 43). Figure 44 shows the ray-casting in real time within the positive VE. In this figure, the camera icon is the viewpoint of the user looking towards the virtual robot. In this particular area of the scene the affective, interactive events were the robot and the monitor. All other objects used as decorative contextual props were categorised as static objects, e.g., apple, mouse, pencil, coffee-cup. As seen in Figure 44 and Figure 45 all these objects and events were currently in the user's field of view. For demonstration, ray-casting lines are displayed in yellow for neutral objects and in red for interactive.
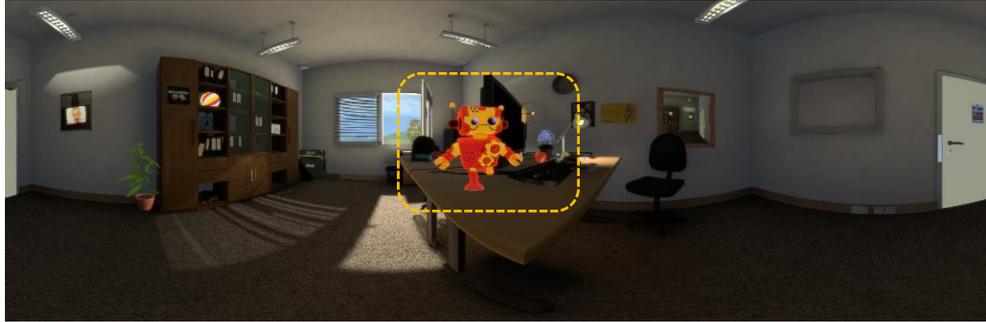
Figure 41. 360 view of the VE. The yellow box (middle) shows the area that was visible to the user.
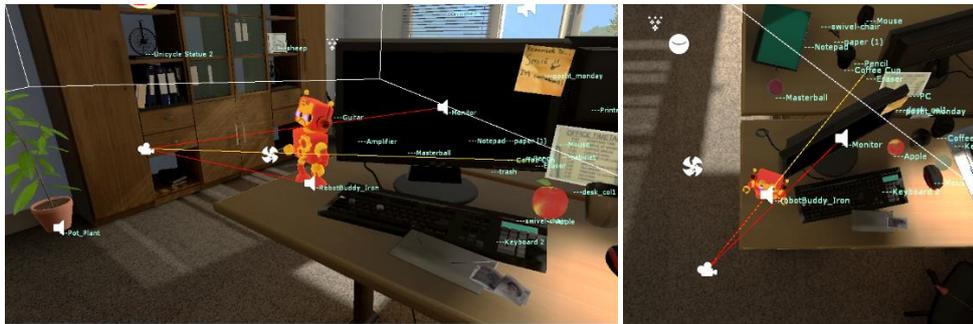
Figure 42. Side and top view of the ray-casting from the viewpoint of the user towards objects in the VE. The rays are displayed in yellow for static neutral objects and in red for interactive events.

Figure 43. Left side: Experimenter view of the scene. Right side: User view of the scene.

Once the algorithm detected one or more event-markers, it activated the corresponding event(s). For example, looking at the robot activated a specific audio clip and the robot started dancing based on a predefined animation. The animation and the audio clip only stayed activated if the user's gaze continued to look at the robot. Otherwise, it would automatically freeze. In addition, all surrounding visible 'static' objects were registered as 'visible objects' in the system and were also exported with timestamps in separate list (ASCII format).

Unlike the static objects, all events were designed to be visible to each participant for at least one time throughout the VR experience. For this purpose, certain conditions were designed, which were applied according to the nature of the stimuli-events. Short events (e.g., email on the monitor, in Positive VE) were able to replicate their animation up to 5 times once the gaze-ray was directed at them for longer than 2 seconds. While longer events (such as butterflies or fire) would only happen once, following a pre-defined activation time (i.e., >60 seconds in the VE). All interactive objects and events were enriched with audio clips which especially in the case of the longer events, attracted the attention and the gaze of the user towards the direction of the event or object. In the case where two interactive events were active at the same time, both their audio clips would be active at the same time. The volume of the audio clips was programmed to change in relation to the user's gaze and position, relative to the wave source. Audio clips from interactive objects would fade out and stop when the user would stop interacting with the corresponding object-event. Some more elaborate events, like fire in the negative VE, triggered a sequence of other abstract stimuli (non-interactive) such as the fire sound, smoke in the room, and the fire alarm with the red light. These events were all timed to be activated within 15 seconds from the activation of the event 'fire'. Of those only the 'fire' event and the 'fire alarm' had individual event markers.

5.2.4. Exiting the scene

Exiting a VE could be executed either manually by the experimenter or by the participant, via the participant's direct interaction with the virtual door in the office-based VEs. Specifically, the participant could go towards the virtual door and 'touch' it with the controller, see Figure 46. The door would only open when the participant had spent at least 70 seconds in the VE. Once the participant exited a room, they would return to the 'baseline VE'.

Figure 44. 360º view of an office based VE. The virtual door (indicated by the orange outline, left) was interactive and could assist participants exiting a VE by 'touching' it with the use of the hand controller.

5.3. Evaluation of elements of the affective VR environments: An online survey study

The four VEs were designed to elicit different ranges of valence and arousal in VR (see section 5.2). The effects of the design of those environments were validated using an online survey. Audio-visual materials were extracted from each VE and their corresponding events. For each VE, a walk-through video was embedded in the survey (each video showing a VE lasted 76 seconds), which was recorded from the point-of-view of a VR-user. After each VE video, individual short videos and images of each event were added (14 events in total per VE). For each VE or event, participant was asked to rate their perceived valence and arousal using SAM rating scales embedded into the survey. This way, we were able to obtain a subjective overall valence and arousal rating (per affective dimension) for each VE scene (see section 1), but also a rating each stimulus-event which gave us insight into the affective impact differences in between events across multiple viewers (see section 5.3.2.1), and inform the design of the VR study presented in chapter 6.

Additionally, to validate the affective impact of the event stimuli, a memory question was added (see section 2.3.3). Participants were asked to report whether they remembered seeing each event in the VE walk-through video. We hypothesised that the memory accuracy will be enhanced for more affective stimuli than non-affective ones. Previous work with conventional (i.e. 2D) stimuli showing that emotional arousing stimuli are most likely to be stored in memory and recalled against neutral stimuli [445]–[448]. In section 5.3.2.3, the memory accuracy scores

were calculated for each event and their relationship to the affective scores was analysed.

For each VE scene, a presence questionnaire was added in the survey, resulting to three total presence scores. The IGroup Presence (IPG) questionnaire [176], [449] was added to measure the expected presence scores. Since the VE scenes were viewed via videos on a screen on this survey, the level of presence they could induce was expected to be low. By comparison, the immersive viewing of those VEs was expected to induce high levels of presence in the main VR study (which is described in Chapter 6). The analysis of the presence scores is presented in section 5.3.2.5.

Participants were asked to fill a basic demographic questionnaire, a personality questionnaire (Big-5, 44 items) [450] and the Toronto Alexithymia Scale TAS20 (20 items) [339]. High alexithymia is a trait related to the deficiency of an individual to identify and describe emotions of their own and others, with various effects on self-ratings and differences in the expected physiological responses induced by affective stimuli (also see section 2.3.3, p. 68). Therefore, for this survey, we explored the effect of high and low alexithymia on the perceived valence and arousal ratings. Additionally, we analysed the effects of personality traits on affect self-ratings. The results of the analysis are provided on section 5.3.2.5.

In the next sections, the affective impact of each VE and their events will be evaluated against the intended valence and arousal scores (neutral-low arousing, negative-high arousing and positive-high arousing VEs). The conclusions of the analysis are presented in section 5.4.

5.3.1. Methods

5.3.1.1. Participants

The responses from 67 participants (out of 91) who completed the survey were used for the following analysis. As part of the initial steps of the survey, the participants were screened for severe psychological, mental disorders and/or encephalopathy. Participants with severe phobias of spiders, fires, enclosed spaces, and of the dark were not allowed to continue with the survey, registering an incomplete session. We excluded in total 24 participants, 22 due to incomplete responses and 2 due to invariant responses. 92% of the selected group of participants were female and 7.6%

male. The mean age was 19.78 (±2.51). In terms of experience with games, 7.5% had no experience, 58.5% had little experience (novice), 30.3 % had average experience with games and 13.6% considered themselves as experts. However, participants had little experience with virtual reality overall, with 36.3% having no experience at all, 43.9% having little experience, and 19.7% having average experience. None of the participants considered themselves experts with VR. The participants were recruited from the Sussex Innovation Centre staff and Bournemouth University student populations via online advertisement. Interested participants were asked to read the information sheet which they received via email before participating. The study was reviewed and approved by the Research Ethics Panel of Bournemouth University (reference ID: 18848). The participants were compensated for their time via Sona credits or £5 Amazon vouchers.

5.3.1.2. **Materials & Procedure**

The survey consisted of three main parts and was programmed in Qualtrics Software [451]. The first part included a description of the study, the methods used, and the participation consent form. This part also included a short demographic questionnaire, on age, gender, fluency in the English language, and level of education. The experience of participants with VR and games was assessed with two questions which were ranged from 'novice' to 'expert' with four overall levels. Questions related to the exclusion criteria of the study were added, such as questions whether the participants were suffering from anxiety and depression, encephalopathy, and fatigue syndrome. Additionally, to protect participants from the exposure to stressful stimuli, participants were screened for extreme phobias to spiders (arachnophobia), fires (pyrophobia), enclosed spaces (claustrophobia) and fear of the dark (nyctophobia), all of which were related to the stimuli presented. Participants who responded positive to the exclusion related questions and phobias were excluded from the study.

The second part was a practical instruction to ensure that participants were familiar with the format of the survey and the meaning of the valence and arousal rating scales. Examples were given using videos which were not the VE stimuli. The participants were reminded that the rating of each video or picture should reflect their immediate personal experience, and no more. Then, a brief sound check was performed using a short bell sound. This ensured that all users could hear the auditory

information of the videos. The participants were encouraged to relax before starting the survey and were reminded to take short breaks of up to 5 minutes throughout the survey.

The preparation step was followed by the evaluation of the VEs. The evaluation consisted of four main sections, one per VE i.e., the baseline, neutral, positive and negative VE. The sections were randomised for each participant. Each for the VE sections consisted of a video which showed the VE from a first-person perspective within VR (which was always shown first). With the exclusion of the baseline VE, each video was followed by the presentation of fourteen events whose sequence was randomised. Each VE video lasted approximately 75 seconds to reflect the duration of the experience designed for the main VR study. The participants were instructed to imagine their experience as if they were within the presented environments. The arousal and valence rating scales were added after each VE video. Time-based control was added for each VE video to ensure the participants had watched the whole video before rating it.

Fourteen events per VE scenario (neutral, positive, negative) were extracted as images and videos giving a total of 42 object-items to be rated; videos were preferred for interactive or animated objects/events involving sounds, and still images were preferred for static objects. Per stimuli, we asked the participants to rate valence and arousal, and answer a memory question (2 levels) about whether they remembered this event/object from the video they watched earlier. We hypothesised that three conditions were equally memorable, however the most memorable events within each condition would also be the ones eliciting high affective rating (reported via SAM) along the two affective dimensions. The scale used for valence ranged from 1-very negative (unpleasant) to 9-very positive (pleasant), and for arousal from 1 low (sleepy) to 9 high (active, excited). The completion of all questions was requested by each participant.

Once the VE video and events were rated, participants were asked to rate their perceived feelings of presence within the respected VE. The option to view the video of the VE again was given to re-jog their memory. The IGroup Presence (IPG) questionnaire [176], [449] was used, consisting of 14 items assessing sense of presence in virtual environments (e.g., "I was completely captivated by the virtual world"). The IPQ is a popular questionnaire in the area of VR ([452],[453]) as it has exhibited good psychometric properties across multiple participants [454]. Items from other presence questionnaires, namely Slater-Usoh [455], Witmer and Singer

[181], Hendrix [456], Carlin, and Hoffman, & Weghorst [457], were also used to construct of the IPQ. Responses are provided using a 7-point Likert scale [63]; rated from -3 (not at all) to 3 (very much). These responses were then mapped to 1 to 7 for analysis purposes.

At the end of the stimuli assessment (third part), participants were asked to fill a personality questionnaire (Big-5, 44 items)[450] and the Toronto Alexithymia Scale TAS20 (20 items) [339]. An optional comment box for participants was added for recording the participants' feedback.

5.3.2.   Results

The results section is divided into five sections: (a) results for each VE condition (positive, neutral, negative) when analysing the 'VE video' (section 1), (b) results for each VE condition when analysing event-related measures relative to the occurrence of static and interactive objects (section , (c) memory accuracies for each VE environment and event-stimuli, d) presence ratings for each VE environment, and d) individual differences in alexithymia, personality traits and their relation to arousal and valence ratings in each VE environment.

2.3.2.1. Data analysis of the entire VE video

The affective impact in terms of valence and arousal of each environment as a whole was assessed first. The average arousal and valence scores were calculated for three main VEs (office-based) and the relaxation-baseline VE (underwater-themed) across participants. As expected, the neutral VE was rated as neutral in valence (Mean: 4.81, SD: ±1.48) and a low in arousal (2.55 ±1.68). The ratings for the negative VE were low in valence s (3.12 ±1.66), meaning they were perceived as negative, and high in arousal levels (6.13 ±1.83). The ratings for the positive VE were high in valence (6.18 ±1.49), meaning they were perceived as positive, and moderately high in arousal levels (4.55 ±2.22). The baseline VE generated for valence ratings close to neutral (5.54 ±1.83) and low arousal ratings (3.72 ±2.24). The means (M) described along with the medians (Md.) and standard deviations (SD) for the VEs are presented in **Error! Reference source not found.** The mean scores per VE are also presented on the affective space diagram in Figure 47.

Table 11. Mean arousal and valence ratings for each VE as recorded in the online survey

VE	Arousal scores			Valence Scores		
	Mean Arousal	Md.	(SD)	Mean Valence	Md.	(SD)
Neutral	2.55	2	(± 1.67)	4.81	5	(± 1.47)
Negative	6.13	6	(± 1.81)	3.12	3	(± 1.64)
Positive	4.55	5	(± 2.21)	6.18	6	(± 1.49)
Baseline	3.72	3	(± 2.24)	5.54	5	(± 1.83)
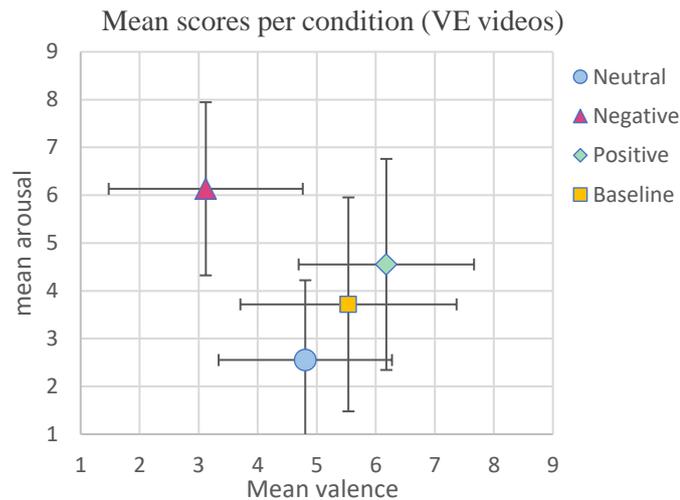
Figure 45. Mean valence and Arousal ratings for each VE condition presented from the first-person point of view of the user within VR.

Inter-Rater Agreement for VE videos. The coefficient of variation (CV=SD/mean) was used as a measure of calculation of the dispersion of the participant's ratings for each video. Two CVs were calculated as percentages for each VE, one per affective dimension (all results in Table 12). Low CV shows low dispersion and therefore high agreement between raters.

Overall, the videos showing the walk through in the neutral and positive VEs designed, were rated in high agreement between participants ($CV_{val\_Neutral} = 30.5\%$, $CV_{val\_Pos} = 24\%$) for their perceived valence (CV < 50%, with criteria of 0.50%). The negative scene generated a higher variation in the valence scores across participants resulting to a lower agreement ($CV_{val\_Neg} = 52.7\%$). The arousal scores of the negative scene showed high agreement between raters scoring $CV_{ar\_Neg} = 29.5\%$, and good agreement for the positive scene $CV_{ar\_Pos} = 48.5\%$. The arousal ratings for the neutral scene varied across participants, giving a high CV (low agreement) of 65.4%. For the baseline VE, the agreement score for valence were

high (CV$_{Val\_Baseline}$ =33.07%) while the agreement scores on arousal were low (CV$_{ar\_Base}$ = 60.22%).

Table 12. Coefficient of Variation (CV) scores per VE video across survey rates.

VE	CV Valence scores (%)	CV Arousal scores (%)
Neutral	30.56	65.38
Negative	52.70	29.53
Positive	24.04	48.50
Baseline	33.07	60.22

Valence ratings of VE-videos. As stated above the valence ratings in the negative scene were the lowest, and ratings in the positive scene were the highest with the other ratings in between. A Shapiro-Wilk test of normality was performed and indicated that ratings for both dimensions were not normally distributed ($p<.05$). Therefore, a non-parametric Friedman's test was conducted on the valence ratings for the four VE conditions. The test showed that the ratings were significantly different between VE conditions, $\chi^2(3)$ =91.02, $p<.001$. Consequently, Bonferroni corrected post-hoc Wilcoxon tests were used to compare each of the conditions. These tests showed significant differences between most conditions. More specifically, the valence ratings in the negative VE were significantly lower than in the neutral VE ($z = 5.39$, $p<.001$), the positive VE ($z = 6.77$, $p<.001$), and the baseline VE ($z = 6.27$, $p<.001$). This is shown by also the significantly higher ratings for the positive VE than for the neutral VE ($z = 4.706$, $p<.001$), but not compared to the baseline VE. The baseline VE had also significantly higher valence ratings than the neutral VE ($z = 2.62$, $p = .036$). These findings show that the three office-based VEs achieved expected valence ratings. However, the baseline VE room induced more positive valence ratings than initially intended.

Arousal ratings of VE-videos. Like for the Valence ratings, a Shapiro-Wilk test of normality indicated that ratings for both dimensions were not normally distributed, and thus a non-parametric Friedman's test was conducted on the arousal ratings for the four VE conditions. The test showed significant differences between the four VEs, $\chi^2(3)$ =79.86, $p<.001$. Bonferroni corrected pairwise post-hoc Wilcoxon tests showed significant differences between arousal scores reported. More specifically, the arousal scores of the Negative VE were significantly higher than the neutral VE

($z =6.78$, $p<.001$), the baseline VE ($z = 5.59$, $p<.001$), and the positive VE ($z = 4.81$, $p<.001$), rendering the negative VE as the most arousing condition. The positive VE had significantly higher arousal ratings compared to the neutral VE ($z =5.42$, $p<.001$), and the baseline VE ($z =2.31$, $p = .021$). Unexpectedly, the baseline VE also achieved a significantly higher arousal ratings than the neutral VE scene ($z = 3.91$, $p<.001$). In summary, arousal ratings were as expected for the VEs, except for the significant difference between the neutral and the baseline conditions.

The results from the statistical analysis showed that the valence and arousal ratings of the videos depicting the experiences within the VR scenarios were significantly different, with the neutral scene being rated as neutral in valence with low arousal levels, while the affective scenes (positive and negative) were able to elicit stronger affective states in terms of high arousal and strongly antithetical valence levels in the Arousal-Valence (AV) space. However, the aquatic baseline scene did not portrait our initial design expectations, scoring higher in valence and in arousal ratings than the neutral scene. For this reason, the environment was replaced by a duplicate of the neutral VE with only the basic architecture structure of the room and no stimuli (the final version on the scene depicted is presented in Chapter 6).

5.3.2.2. **Event-related data analysis**

Apart for the VE videos, participants rated all 14 stimuli (objects/events) per scene (neutral, positive, negative) using the same valence and arousal SAM scales (1-9). The objects/events that were rated are in the Table 55, Appendix C. Participants were also asked whether they remembered the objects/events. The findings for the valence and arousal ratings are presented here. The findings for the memory accuracy are shown in the subsection 5.3.2.3.

Mean valence & arousal ratings. In Table 13 the mean valence and arousal scores across stimuli were calculated per VE. For the stimuli events in the neutral VE, the mean valence ratings (across all events) were 4.63 (±0.15) and the mean arousal ratings were 2.61 (±0.35). For the objects / events in the positive VE, the mean valence ratings were 5.72 (±0.56) and the mean average rating for the arousal ratings was 4.06 (±0.68). For the negative scene, the valence ratings were 3.45 (±0.45) and the arousal ratings were 5.18 (±1.04) on arousal. Ratings for each object/ event are

depicted in Figure 48 based on their average valence and arousal scores averaged across participants. As expected, the events and objects within each condition elicited different levels along the valence and arousal dimension, as characterised within the AV space. The depicted 'V' shape is similar to shapes valence and arousal scores reported in several studies [458].

Table 13. Mean valence and arousal scores across stimuli per VE

Stimuli/VE	Mean Valence scores	Mean Arousal scores
Neutral	4.63 (±0.15)	2.61 (±0.35)
Negative	3.45 (±0.45)	5.18 (±1.04)
Positive	5.72 (±0.56)	4.06 (±0.68)
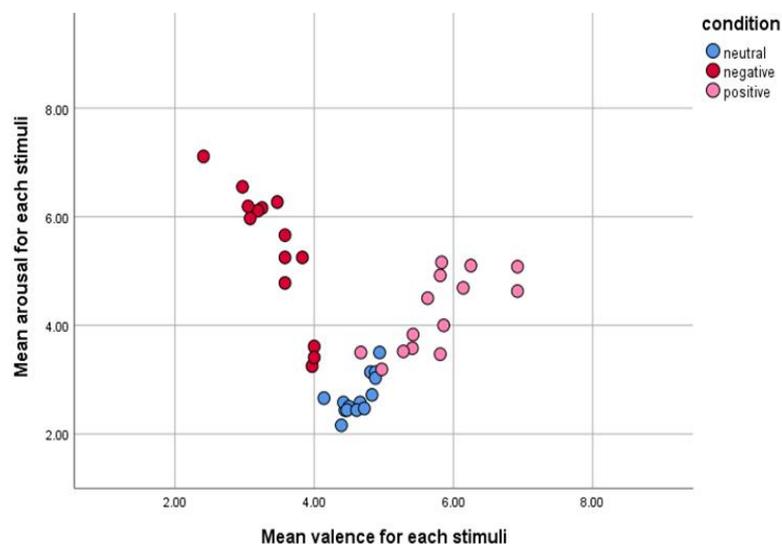
Figure 46. Mean valence and arousal ratings for all events in the three VE conditions. The x-axis (bottom) displays valence ratings and the y-axis (left) displays arousal ratings.

The event-based analysis was made to inform the design for the VR study described in Chapter 6. The results allowed us to find objects/event that did not induce the required affective range, which was predesigned for each VE, or objects that were not consistently rated, exhibiting high levels of variability. In general, the categories were found to be well separated. No stimuli exhibiting negative valence was found in the positive category, and similarly no positive valenced stimuli were found in the negative category.

From the inspection of the events whose ratings approached the neutral, low arousing area of affective space, for the positive and negative VE, we observed that they belonged mainly in the group of 'static' objects. The valence and arousal scores

were recalculated to observe the effect of those objects in the average valence and arousal scores of each affective VE. It was found that for example after excluding the static objects of 'guitar' and the 'amplifier' from the positive VE condition, the mean arousal ratings increased to 4.20 (±0.69) compared to the previous ratings of 4.06 (±0.68), bringing the mean arousal score for the events of this VE category closer to the arousal scores of the negative VE ratings. The inclusion of the static objects populating each VE allowed us to observe the average affective impact of each VE from an all-stimuli inclusive point of view. As the VR experience was meant to be explored in an interactive manner, the affective impact of all static context related objects per VE would be dependent on the duration and subjective interaction per user.

Note, the most arousing stimulus in the positive VE condition was the 'light explosion' (mean arousal= 5.03). However, the 'light explosion' was found 'confusing' and generated some ambiguous comments by our survey participants. Therefore, this event was excluded from the main VR study. Figure 49 shows the scores per event per VE condition in the affective space, based on their mean valence and arousal ratings (also see Table 55 of the Appendix C).
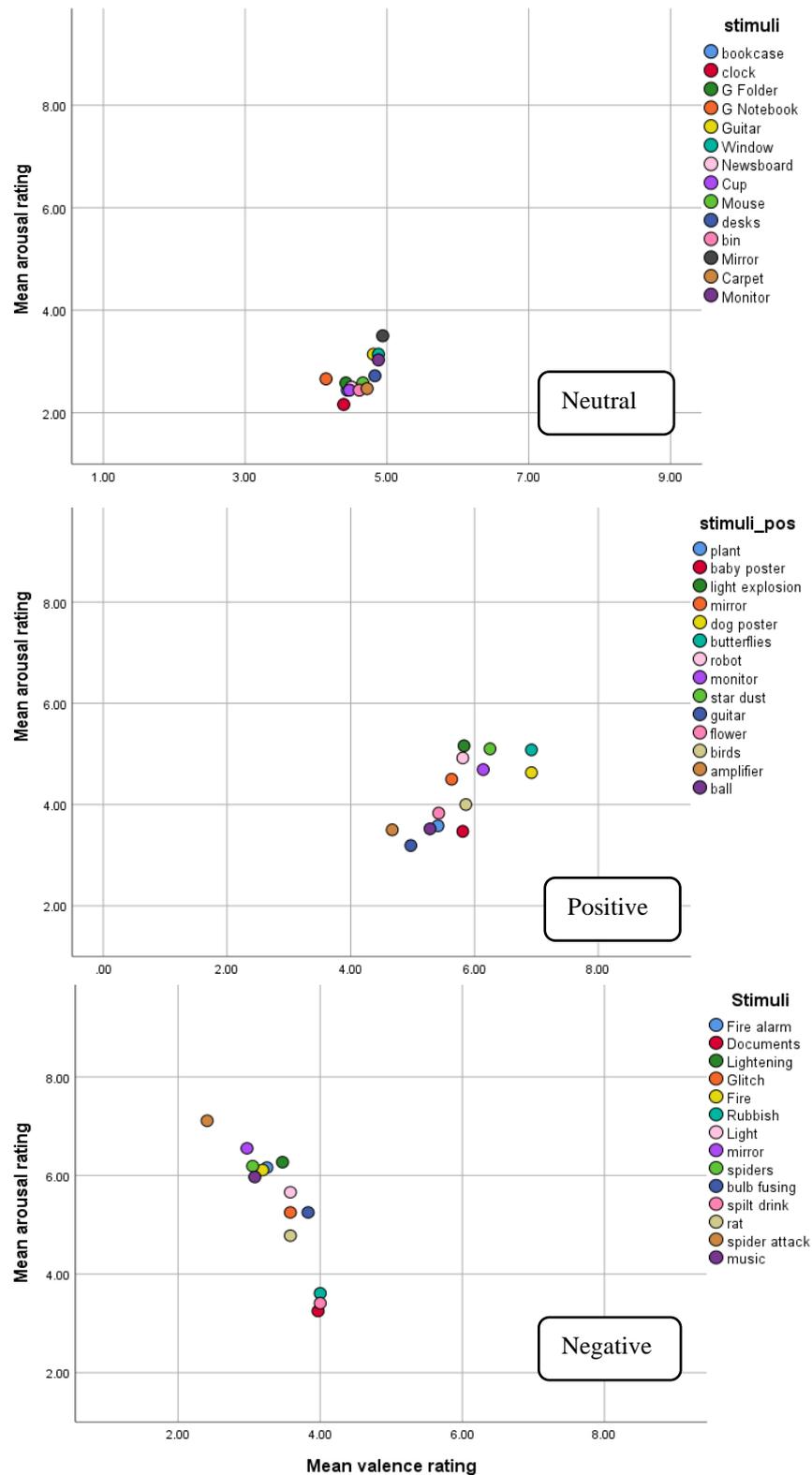
Figure 47. Mean valence and arousal ratings for each stimulus in each of the VE conditions. A list of all stimuli including static objects and animated events is displayed on the right side of each figure. A. neutral VE condition. B. positive VE condition. C. negative VE condition.

Inter-rater Agreement. The coefficient of Variation (CV) was computed for each event stimulus from the SAM scores from all participants. Overall, low inter participant variation (high agreement) was found for valence ratings in all three conditions (mean_CV_{val} < 50%). This was also the case for arousal ratings (mean_CV_{ar} < 55%). Saying that, the participant agreement was a bit lower for arousal ratings compared to the valence ratings, as reflected in higher CV scores. More specifically, high agreement was recorded for valence ratings in the neutral condition (32.36±3.02%) and the positive condition (30.92±2.97%). The variation was slightly higher for the valence ratings in the negative condition (49.99±10.10%). For the arousal ratings, the highest agreement was present in the negative condition (41.88±8.54%), followed by the positive condition (53.4±5.56), and the neutral condition (63.4±3.06%) see Table 14. Due to the size of the table, the CV results per event are attached in Table 56, in Appendix C.

The mean CV scores per dimension across all the events per VE followed the same pattern as the CV scores per VE video ratings (see previous section). Overall, for both scores, VE video and mean across events per VE, high inter-rater agreement was observed for positive, neutral valence but not for negative valence ratings. Instead, higher-interrater agreement was found in the negative arousal ratings compared to the other two conditions. This pattern observed in the CV scores per VE reinforces the need for both dimensions self-rating measures in order to distinguish between conditions, as otherwise, one rating measure would not be sufficient.

Table 14. Mean Coefficient of Variation (CV) scores across events per VE condition.

VE	CV Valence scores (%)	CV Arousal scores (%)
Neutral	32.36	63.40
Negative	49.99	41.88
Positive	30.92	53.40

5.3.2.3. **Memory accuracy of events within each VE environment**

A quick memory question was added at the end of each stimulus "Do you remember this object/event?". For each stimulus participants rated if they remembered it or not (1 or 0). All participants answered the memory questions for all events in the survey.

The memory accuracy levels of an event would allow us to deeply evaluate the affective impact of our stimuli and identify events which elicited stronger responses thus leaving a stronger trace in the memory of most participants. Thus, the memory accuracy scores for the objects/events were used an additional way to evaluate our study design and to differentiate between affect inducting events and non-affective ones.
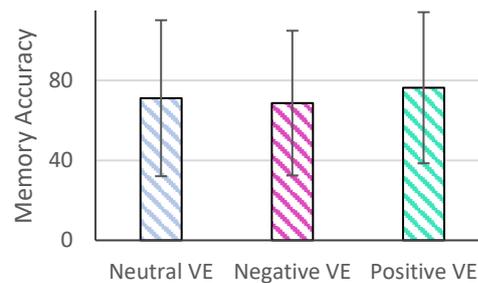
Figure 48. Mean memory accuracy scores across events per VE condition.

As a first step, the memory accuracy scores for each stimulus were calculated (average score across participants*100) for each object/event per condition (see Figure 50). For those, we also calculated the mean memory accuracy averaged across all events per VE condition. The results showed similar mean scores between conditions. The mean accuracy across events for the neutral condition was 71.12 ± 39.02, for the negative condition was 68.66 ± 36.20, and for positive condition was 76.33 ± 37.80. The memory accuracy scores were tested for normality and homogeneity of variance. The non-parametric related samples Friedman's test on the average memory accuracy scores across events per VE showed no significant differences between the VE conditions, showing that all three conditions were equally memorable. Note, the variability of the answers was high in this online survey. These results showed an overall high average level of memory accuracy for all three scenes.

Next, the mean memory accuracies per event across participants were examined in reference to their corresponding valence and arousal rating scores. The memory accuracy scores for each event are presented on Table 15. The events per VE condition that scored the lowest on memory accuracy (lower 25%) and the events that scored the highest scores (upper 25%) are highlighted with pink and green colour respectfully.

From the stimuli used in the negative condition, the 'spider attack' event was the most memorable, which was also the most arousing, event ($V_{score}=3.76$,

A$_{score}$=6.66), followed by the 'mirror' (ghost figure) event (V$_{score}$=3.52, A$_{score}$=6.43) and 'spiders' across the room (V$_{score}$=2.76, A$_{score}$=5.90). Similarly, the most memorable stimuli for the positive scene were the 'butterflies', the 'robot' and the 'dog poster', whereas, the least memorable stimuli were 'guitar,' 'flower' and 'ball'. Not surprisingly, the least memorable objects/events across all three VE conditions were the static objects, such as the 'cup', 'guitar', 'documents', 'grey notebook', and 'rubbish bin' (mean memorability <50%). Although they existed in all scenes, they were not expected to be memorable nor arousal eliciting. The remaining objects and events rendered a high memorability (between 55 - 95%) with the exception of the 'rat' event in the negative scenario which scored lower than expected; the low memory accuracy score could be caused by the reduced visibility of the event in the video used in the survey (located on the lower bottom of the screen).

Table 15. Mean memory accuracy in percentage for each stimulus in each condition.

Neutral VE			Negative VE			Positive VE		
Stim.ID	M	Std.	Stim.ID	M	Std.	Stim.ID	M	Std.
Bookcase	91.05	28.55	Fire Alarm	86.57	34.10	Green Plant	89.55	30.59
Clock	61.19	48.73	Documents	22.39	41.68	Baby Poster	70.15	45.76
Green Folder	55.22	49.73	Lightening	91.05	28.55	Light Explosion	88.06	32.43
Grey Notebook	41.79	49.32	Glitch-View	55.22	49.73	Mirror Reflection	85.08	35.63
Guitar	50.75	49.99	Fire	89.55	30.59	Dog Poster	88.06	32.43
Window	89.55	30.59	Rubbish Bin	26.87	44.33	Butterflies	95.52	20.68
News-board	71.64	45.07	Flickering Light	91.05	28.55	Robot	95.52	20.68
Cup	35.82	47.95	Mirror	94.03	23.69	Monitor Message	88.06	32.43
Mouse (PC)	76.12	42.64	Spiders	88.06	32.43	Star Dust	83.58	37.04
Desks	97.02	17.02	Light Bulb	74.63	43.52	Guitar	44.78	49.73
Bin	55.22	49.73	Cup (Spilt Drink)	23.88	42.64	Flower	52.24	49.95
Mirror	83.58	37.04	Rat	41.79	49.32	Birds	76.12	42.64

Carpet	92.54	26.28	Spider Attack	97.02	17.02	Amplifier	55.22	49.73
Monitor	94.03	23.69	Spooky Music	79.10	40.66	Ball	56.72	49.55

TOP 25% ▇   BOTTOM 25% ▇

Based on the hypothesis that the memory accuracy scores are enhanced for more affective stimuli, Spearman correlations between the memory accuracy scores (following tests for normality, see section 3.6.4) and the valence and arousal ratings were analysed, separately for the positive and negative VE condition. The findings (Table 16) show that more positive or negative the valence ratings (deviation from neutral) and higher arousal ratings are strongly correlated with higher memory accuracy. Please note, valence and arousal ratings were strongly intercorrelated ($r(12) = .68$, $p = .007$ for the positive VE, $r(12) =$ -.94, $p<.001$ for the Negative VE, and $r(12) = .49$, $p = .018$ for the neutral VE) which makes it more difficult to interpret this finding (see discussion).

Table 16. Correlations between memory accuracy and the valence and arousal ratings for each affective condition.

Memory scores *	Corr. Coef. (r)	P (Sig)
Valence (Negative VE)	-86*	<.001
Arousal (Negative VE)	.92*	<.001
Valence (Positive VE)	.58*	.032
Arousal (Positive VE)	.69*	.006
*Significance at a 0.01 level		

Two simple linear regressions were calculated to predict memory accuracy based on the a) valence distance from mean and b) on arousal ratings per stimulus. Valence distance was calculated as:

Valence Distance = abs (valence –mean), *where* mean = 4.6.

Valence distance scores significantly predicted memory accuracy scores, $\beta = 16.40$, $t(1) = 3.23$, $p = .003$. Valence distance scores also explained a significant proportion of variance in memory scores, R^2 of 0.216, $F(1,40) = 10.47$, $\underline{p = }.003$. Similar results were observed using the arousal scores $\beta = 7.11$, $t(1) = 4.43$, $p < .001$, with R^2 of 0.172, $F(1,40) = 8.31$, $p = .006$. This finding is also displayed in Figure 51. In line with the previous correlational analyses, the findings from the regression analysis

show that memory accuracy can be partially predicted by the affective value of an object / event, as measured by the valence and arousal ratings.
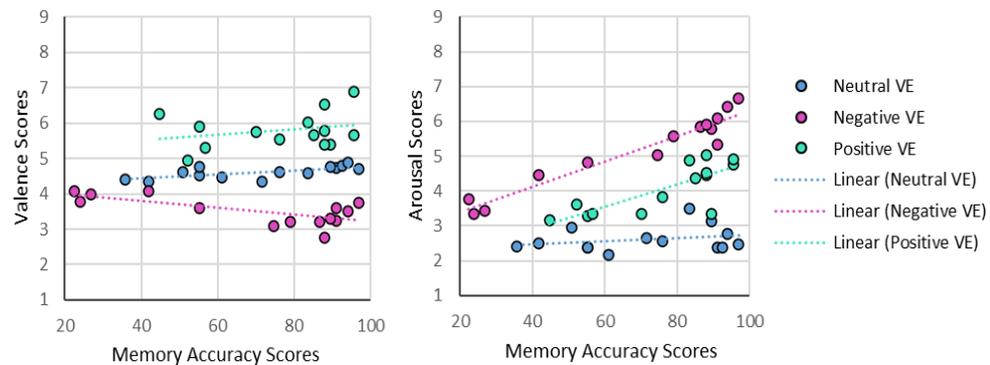
Figure 49. Left figure - relationship between valence ratings (y-axis) and memory accuracy for each stimulus. Right figure - relationship between arousal rating (y-axis) and memory accuracy (x-axis) for each stimulus. Red dots are representing stimuli presented in the negative VE condition, green dots represent stimuli from the positive VE condition, and blue dots represent stimuli from the neutral VE condition.

5.3.2.4. **Presence Ratings for each VE environment.**

The IGroup Presence questionnaire was answered after each of the three affective VE video clips (positive, neutral, negative) in the online survey. The questionnaire is consisted of three independent subscales and a general item not belonging to a subscale. The subscales are namely 'spatial presence' (the sense of physically being present in the VE), 'involvement' (the level of devoted attention and involvement experience in the VE) and 'experience realism' (measuring the subjective experience of realism in the VE) [449]. The additional item assesses the general feeling of presence as the 'sense of being' in the VE. A score for all three subscales and one for the general presence were calculated across participants for each VE video. Table 17 shows the result per VE condition.

Table 17. IGroup presence mean results per subscale and condition

	Neutral VE	Positive VE	Negative VE
General Presence	1.75(±1.72)	1.94(±1.62)	3.06(±1.74)
Spatial Presence	1.87(±1.24)	1.96(±1.33)	2.80(±1.43)
Involvement	1.64(±1.21)	2.27(±1.25)	2.78(±1.43)
Experience realism	1.80(±1.11)	1.38(±1.09)	1.84(±1.20)

As expected, all three conditions scored low on presence because VE experiences were presented as videos through non-immersive interfaces in an online survey. The mean presence score for the neutral condition was 1.77±1.02, for the positive condition 1.89±1.13, and for the negative condition 2.62±0.55 (see scores for all four factors per condition in Figure 52). In Figure 53 the presence scores as the expression of the three main subscales (involvement, reality and spatial) per condition are shown.
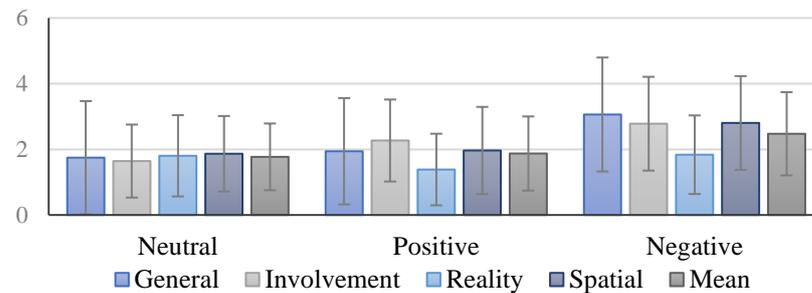
Figure 50. Plot showing average presence scores per condition.
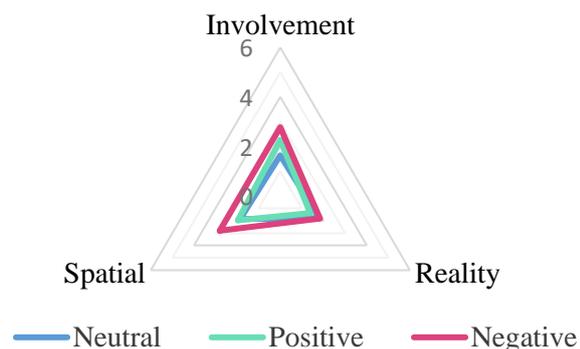
Figure 51. Radar plot showing average scores for each subscale: Involvement, spatial presence ('Spatial') and experience realism ('Reality') per VE. General presence is excluded from this radar plot.

Kolmogorov-Smirnov and Shapiro-Wilk tests of normality were carried over the average presence scores for each dimension which indicated that the presence scores for the neutral and positive scenarios were not normally distributed. The difference between the three conditions in terms of presence was statistically significant as shown by a Friedman's related-samples ANOVA by ranks ($\chi^2(2) = 34.17$, $p < .001$). Wilcoxon tests with Bonferroni corrections were conducted to evaluate whether the participants' scored higher average presence in one of the conditions against the others. The results indicated significant differences between the neutral and the negative condition ($z = 5.55$, $p < .001$, where mean ranks in favour of the negative

conditions was 36.57 while the mean of ranks in favour of neutral was 19.67), and the positive against the negative scores ($t = -5.12$, $p < .001$, where mean ranks in favour of the negative conditions was 38.13 while the mean of ranks in favour of positive was 19.03). However, the difference between the presence scores of the positive and the neutral condition was not significant. As a result, it was concluded that the negative scenario drew significantly higher feeling of presence than the other two scenarios, even though they were all watched as a pre-recorded video in this survey.

Since the negative scenario elicited also high arousal ratings (see section 1), we investigated the relationship of presence scores with the arousal scores for each condition. Interestingly, we found a significant positive correlation between the arousal ratings and the presence ratings for the negative scenario ($r = 0.362$, p = .003). The correlation between the arousal ratings for the positive scenario and the presence scores was also tested, and was found significantly positive (r = 0.257, p = .036). These results show that mean presence scores in our sample increased with higher arousal, regardless of the polarity of the affective context, whether positive or negative. As a next step, the correlation between valence scores and presence scores showed a negative significant correlation for the negative condition (r = -.263, $p = .031$) but not for the positive scenario (r = 0.24, $p = .054$). Once more, the results confirm the relationship between highly affective content and presence ratings as discussed in section 2.3.3.

5.3.2.5. Individual differences in alexithymia and personality and their relation to arousal and valence ratings in each VE environment.

Alexithymia. The Alexithymia scores were calculated per person and a simple, binary categorisation of the score was made based on a division point (division point = 51) as used in previous research [339]. With this categorisation, alexithymia scores higher than 51 were labelled as 'high' and the rest as 'low'. Out of 67 participants, 17 were categorised as having 'high' levels and 50 as having low levels of alexithymia.
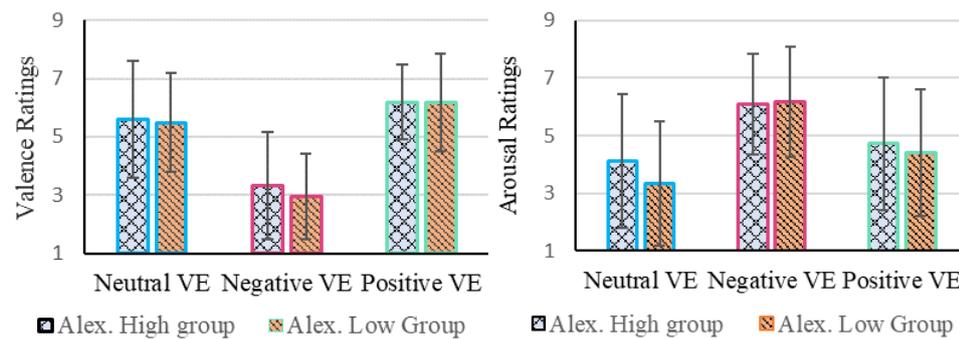
Figure 52. Mean valence and arousal ratings per alexithymia group. Left figure: Mean valence ratings for each condition. Right figure: Mean arousal ratings for each condition.

The left side of Figure 54 shows that valence scores for people with high alexithymia were slightly higher for the negative and neutral conditions. In other words, the valence perceived by participants with high alexithymia was not affected by the stimuli manipulation (affective VEs) as much as the participants with low alexithymia. The effect of alexithymia on valence scores was explored with a 3 x 2 mixed ANOVA with the within-participant factor the three VE conditions and the between-participant factor the alexithymia group (low vs. high alexithymia levels). The dependent variables used was the absolute valence distances between negative and neutral, and between positive and neutral (as in section 5.3.2.3). The result of the test showed that there was no significant main effect of the alexithymia group on the valence ratings ($F(2,130) = 3.43$, $p = 0.68$).

In the right side of shows the mean arousal rating of each condition for the alexithymia groups. A higher mean arousal rating was observed on the participants of the high group in the cases of neutral and positive condition, and slightly lower means arousal rating for the negative condition. A 3x2 mixed ANOVA was conducted to compare the effect of alexithymia group on the arousal ratings in the three VE conditions. There was not a significant effect of alexithymia groups on the arousal ratings, $F(2,130) = 1.17$, $p = .314$. The group division was unbalanced, since only about 25% per cent of our participants could be categorised as 'high' alexithymics. Overall, the effects of alexithymia of A/V scores were not found to have a significant impact on valence and arousal scores.

Personality Traits. In this section, findings regarding the relationship of each trait (extraversion, neuroticism, openness, agreeableness, and conscientiousness) with the arousal and valence ratings from the survey are presented. Due to the sample size

and the absence of any control metrics or supervision in place for the completion of the personality questionnaire, we present the following findings as preliminary results.

To explore the effects of those five traits on the valence and arousal ratings reported, 3 x 2 mixed ANOVAs with the factors |VE condition (positive, neutral, negative) and Group (low vs high personality trait). These ANOVAs were conducted for each personality trait (5) and for each rating type (arousal vs. valence), separately. The groups were created by using a median split to define a high and a low category. Table 18 shows the number of participants for each group. As in previous sections the results for valence ratings are described first.

Table 18. Number of participants per trait category (high, low).

	High category (N)	Low category (N)
Extraversion	31	36
Agreeableness	33	34
Conscientiousness	32	35
Neuroticism	33	34
Openness	32	35

The findings from the ANOVAs performed on the valence ratings are presented in Table 19 and Figure 53.

Table 19. Analysis results from effect of personality traits on valence ratings.

Personality Trait	Main Effect Group	Main Effect VE Condition	Interaction Group x VE Condition
Extraversion	-	-	-
Agreeableness	-	-	-
Conscientiousness	-	-	-
Neuroticism	sign	-	sign
Openness	-	-	-

For the neuroticism trait, the analysis revealed a significant main effect of Group ($F(2,78) = 3.95$, $p = .023$) The mean scores are shown in Figure 53. Across all conditions, people in the high neuroticism group had higher valence ratings (5.1) compared to people in the low neuroticism group (4.3). However, this effect was further modulated by VE condition type. More specifically, Bonferroni corrected independent t-tests did not show significant group differences for the neutral ($t(65) = 1.40$, $p=0.5$) and for the positive VE conditions ($t(65) = .34$, $p=2.2$), but for the negative VE condition ($t(65) = 4.33$, $p<.001$). Here, higher valence ratings ($N_{high} =$

3.9) were given by the high neuroticism group compared to the low neuroticism group (N_{low} = 2.4). Note, for both neuroticism groups, valence scores were highest for the positive VE condition, medium for the neutral VE condition and lowest for the negative VE condition. However, the high neuroticism group rated the negative VE as less negative than low neuroticism group.
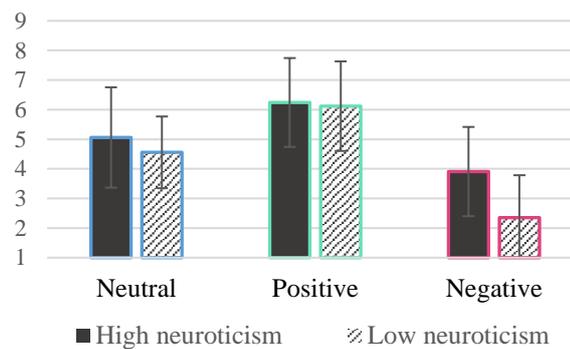
Figure 53. Mean valence ratings for high and low neuroticism groups.

Following the same process, the effects of the five personality traits on arousal ratings where analysed. The findings are presented in Table 20 and mean arousal ratings per group and VE condition are shown in Figure 54

Table 20. Analysis results from effect of personality traits on arousal ratings.

Personality Trait	Main Effect Group	Main Effect Condition	Interaction Group x Condition
Extraversion	-	-	-
Agreeableness	-	-	sign
Conscientiousness	-	-	-
Neuroticism	-	-	-
Openness	-	-	-

Significant effects were found for the agreeableness trait where a significant interaction between group and VE environment can be reported ($F(2,78)$ =6.33, p = .003). This interaction was further analysed using Bonferroni corrected post-hoc t-tests. First, VE condition differences were analysed for each group separately using paired t-tests. It was found that for the low agreeableness group, arousal ratings were lowest for the neutral condition ($A_{low-NEU}$ = 2.2), at a medium level for the positive condition ($A_{low-POS}$ = 4.1and at highest for the negative condition ($A_{low-NEG}$ = 6.4). All ratings were significantly different from each other (Neutral-Positive: $t(33)$=-5.29, $p<.001$, Neutral-Negative: $t(33)$ = -11,65, $p<.001$, Positive-Negative: $t(33)$=-5.95, $p<.001$). However, this was not the case for the group with high agreeableness scores ($A_{high-NEU}$ = 2.9, $A_{high-POS}$ = 5.0, $A_{high-NEG}$ = 5.8). Here, significant arousal rating difference were found between the positive and neutral conditions ($t(32)$ = 4.51, $p<.001$), and the negative and neutral conditions ($t(32)$=7.04, $p<.001$),

but not between the positive and negative conditions ($t(32) = 2.2$, $p=.10$).
The latter result showed that this group rated both negative and positive VE conditions as high arousing. Note, there were no significant group differences between the low and high agreeableness groups when each of the VE conditions were analysed separately with independent t-tests (Neutral arousal ratings: $t(65) = 1.58$, $p=.33$, Positive arousal ratings: $t(65) = 1.53$, $p=.39$, Negative arousal ratings: $t(65) = 1.13$, $p = .78$).
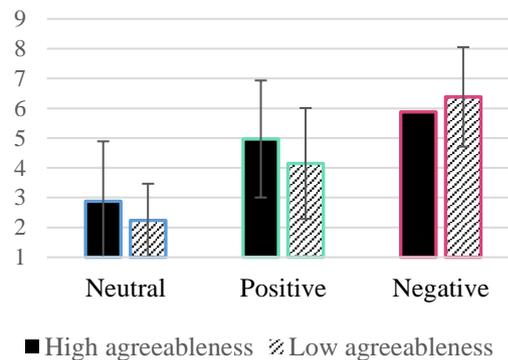
Figure 54. Mean arousal ratings for high and low agreeableness groups

In summary, two of the personality traits (extraversion, introversion, agreeableness, openness, neuroticism, and conscientiousness) were found to have affect the valence and arousal ratings for each VE. High agreeableness was found to produce similar high arousal ratings for both positive and negative VE scenarios. High agreeable people are expected to collaborate [459] and perhaps be more succumbed to experiment bias, thus rating the VEs in a manner that would benefit the researcher. The trait of extraversion was not found to significantly affect with valence distance and arousal ratings, although extraversion has been suggested to be linked to increased experiences of positive emotions in the past [460], [461]. The tests conducted showed a significant difference on valence scores between people with high and low neuroticism, which was significantly prominent in the negative condition. The mean valence scores indicated relatively more positive ratings for the members of the high neurotic group compared to the low one. We would generally expect highly neurotic individuals to have a tendency towards experiencing negative emotions more intensely [462], and therefore rate the negative videos more negatively. These studies used different types of stimuli including realistic photographs, human faces and intense negative features e.g., blood etc. which have

an active reference to real-life. In our case we used virtual spatial stimuli of non-human objects, which were created to represent existing objects. Thus, our stimuli potentially do not reflect the same affective impact as the stimuli used in other studies. Perhaps this effect could be investigated in the future studies. However, this effect was not visible in our valence ratings, which indicate that the video of the negative scenario did not have the expected negative effect on people with high neuroticism. It is though worth noting that the sample size was limited and the responses from both groups exhibited high standard deviations in their valence ratings. Additionally, as the online survey required approximately 50 minutes for completion per participant, it is possible that these effects on the ratings could be influenced by fatigue and/or boredom.

5.4. **Chapter Discussion and Conclusions**

In this chapter the development and the validation of the VR environments designed to induce different variations valence and arousal were described. In section 5.2, the process of designing the environments as a replica of an existing office room was explained, as well as the spatial elements including a balanced number of audio-visual stimuli per VE (which were divided into 'interactive events' and 'static objects'). A proprietary event-system was developed to detect and track the apparition of stimuli and the user's engagement with the virtual stimuli. This event-system permitted the tagging of events-markers and the saving of those from within the application for synchronisation with physiological signal data. Interaction mechanisms were developed and incorporated within the virtual simulation in order to a) trigger certain interactive events based on the user's movement and gaze (non-linear interaction), and to b) facilitate the user's free movement in the virtual spaces, so that they could explore and experience the contents of each room in a personalised manner. Special effort was put in the design of each environment from 3D objects, 3D-sounds, animations and textures to physics and light/shadow rendering, to create an overall realistic space for the user to feel present within. The validation of using the designed virtual environments (as well as the stimuli which populated each) as a method of affect induction, was made by creating and disseminating an online survey.

The survey (explained in section 5.3) included videos of the environments recorded from the point of view of a user in VR, as well as images from each object/event (1 video and 14 stimuli in total per VE). The conditions (VEs) were randomised between themselves, and within each, the images of the stimuli were also randomised. The participants were asked to rate their perceived affect in terms of arousal and valence using SAM scores for both videos and stimuli. The mean affective scores for the stimuli were re-validated against memory accuracies scores, which were calculated by the participants' responses to whether they remembered the object from the initial VE video. Additional questionnaires on demographics, presence and individual differences in alexithymia were also added. In total, four virtual environments were validated, a baseline, neutral, positive and a negative. With the exception of the baseline, each VE was designed to stimulate a predefined (by design) range of valence states (either neutral, negative or positive), with different variations of arousal evoked by various stimuli within each environment.

Responses from 67 participants, showed that the videos of the VE-scenario conditions achieved to induce the expected mean arousal and valence ratings. More specifically, the neutral scenario elicited neutral valence and low arousal, while the positive and negative scenario elicited highly positive and highly negative valence respectively, and above-average arousal, which was found to be increased for the negative VE (see Section 1). With the exception of the baseline VE, the conditions were found to induce significantly different valence and arousal scores from each other. The baseline environment was rated as more positive and more arousing than the neutral environment, against our initial expectations, which prompt us to redesign this environment (see Appendix C for more information). Overall, the validation of the VEs and the individual stimuli used was an important step towards studying and reliably inducing affect in VR. As shown by the example of baseline environment, it is imperative that more research groups take a similar approach in order to interpret their data more accurately before stepping directly to physiological data collection.

Apart from the ratings per VE video, the valence and arousal ratings per event (14 in total for each scenario) were also analysed. Primarily, the average ratings of all stimuli per VE were calculated and analysed, re-validating the ratings per VE-video which were earlier observed. The ratings of the stimuli for each VE allowed us to inspect the effect of the environmental properties in more detail, and to find and exclude those that did not induce the affective ranges desired. Interestingly, the stimuli which were rated as more arousing and highly valenced,

were the VE-specific interactive ones compared to VE-specific static objects and other elements which existed in all VEs. The interactive events were also the most memorable, as confirmed by the memory accuracies scores. Significant correlation was found between memory accuracies, valence and arousal scores. Linear regressions also showed that memory accuracy of a stimulus could be predicted by their affective ratings. Overall, all three VEs were equally memorable, as expressed by the mean scores across the stimuli per VE. These results agree with the relationship between affect intensity and memory (see section 2.3.3.2).

Presence scores were also reported in this survey for each VE video. We expected that the presence scores withing those VEs would be higher using VR technologies in the main study. Indeed, the level of presence was rated low, which was expected due to the nature of the experience and the type of the content-presentation tool that was used. When the presence scores between the three VEs were compared, a significant difference was found, with the negative VE provoking higher presence scores than the other scenarios. Tests indicated strong relationship with arousal ratings for both positive and negative VEs, which agrees with the theory discussed in Section 2.3.3 on the relationship between presence and arousal. This result also suggests that by utilising high arousing affective elements a VR designer may be able enhance the presence levels that users feel within.

The individual differences in alexithymia in relation to the valence and arousal ratings for each VE were also explored. High alexithymia can cause difficulty in recognising and regulating one's emotions and has been shown to have an influence on how physiological responses are elicited in an affect stimulating context (see section 2.3.3). The differences between the two groups however did not yield statistically significant results. Therefore, the effects of alexithymia on physiological responses were not further explored in the next studies.

In summary, the design of the affective, virtual experience comprising of three virtual room replicas, was validated by the survey. The validation was three-fold, using SAM ratings per environment, SAM rating per stimulus, and memory accuracy scores. This affective library specifically designed for immersive consumer-VR headset, is to our best of knowledge the first involving 3D, interactive, virtual environments and individual objects/events which are triggered by the user's actions, mapped to existing physical room to allow the user to naturally walk within it. The designed VEs were used for our next study, the Feasibility study 3 (termed

VR study), involving physiological data acquisition from participants wearing a VR headset at the Science Museum in London explained in Chapter 6.

Summary of key findings for Chapter 5

- Four interactive 3D VR scenarios for affect induction were created and validated using an online survey.

- Three of these designed scenarios (neutral, positive and negative scene) were able to induce low arousing neutral, high-arousing positive and high-arousing negative responses in participants. These were compared to the fourth baseline scenario. This validation was three-fold from post-VE ratings, ratings per event/stimulus in each scene and from memory accuracy scores.

- An event-detection system was developed based on a custom gaze tracking for dynamic data annotation in real-time. This system along with the validated scenarios were used in the next study.

Chapter 6

Affect Detection in Virtual Environments (FEDEM 3 study)

6.1. Introduction

In this chapter we will describe the third study designed to explore the feasibility of our technologies to detect affect in virtual reality settings. Until recently, affect modulation was implemented predominantly using non-immersive technologies and stimuli. Affect stimulation can be influenced by the level of involvement of the user with the content provoking realistic responses and experiences [433]. In VR this can be in turn reflected in the level of presence the users feel, which can be affected by the level of immersion of the technologies utilised as explained in section 2.3.3.

Currently, immersive technologies offer advanced, light-weight user tracking methods combined with high-resolution graphic tools. These advances in turn encourage the overcoming the traditional use of these technologies in seated experiences with distractive and cumbersome set-ups, and move towards the incorporation of unintrusive, wearable technologies for naturalistic exploration and interaction within VR. Currently, the number of studies that utilise virtual reality as an emotion induction tool are scarce. However, in the majority of those, e.g. [294], the experience of the user with the content is often spatially and physically constrained. Nowadays, multiple affective databases of images, videos, and even 360° immersive videos exist. However, all of these require passive observer-like participation of the user, not taking advantage of the 3-dimensonality offered by this constantly-advancing VR medium. The utilisation of 3D VEs have started to emerge in emotion research (such as [294], [463], [464]). However, as far as we are aware the effect of active exploration within immersive VR settings on affective responses (compared against passive observation) has not been sufficiently investigated. The presented study addresses this issue.

For this, a fully immersive VR experience comprising 3D affective VE scenarios was custom-built with two modes of interaction: active and passive. The aims of the study were: firstly, to validate the affect-detection system and the proposed sensor set-up (see description of the system in section 3.2) for use within virtual reality experiences, and secondly, to compare the ability of the system to detect voluntary facial expressions and spontaneous affective responses recorded within directly (inter)active and vicariously passive VR conditions. As affect detection studies in VR were scarce at the time of the study, and existing libraries on affective stimuli were limited to 2D content, custom-made human-scale 3-D virtual environments (VEs) were designed for the purpose of this study.

In short, the VEs developed were replicas on an existing physical office room. Each replica was populated with 14 audio-visual stimuli (referred to as 'events') and decorative elements corresponding to the arousal and valence levels that each was indented to evoke (see section 5.2. for more details). To allow for naturalistic interaction and exploration in VR, the VEs were designed to be explored using a commercial head-mounted display (referred later as VR headset), which allowed the user to look and move freely in various directions, by leveraging high resolution motion orientation (via IMU sensors), and mapping that position and orientation to the view of the 3D VE inside the VR headset (see section 2.3.3).

In order to fully utilise the free-walking capabilities of the VR technologies, a user-gaze based interaction tracking system and an event-marker system were developed. These systems enabled us to track the elements of the space that the user was looking at, and thus provide the link between affective responses and the stimuli in non-linear, immersive VR experiences. These systems allowed the analysis of physiological changes in relation to the context of the user's virtual interactions. The importance of context for the interpretation of affective responses was explained in section 2.3.1.The study paradigm designed for this study could assist in future studies utilising immersive VR technologies.

The study presented in this chapter was to our best of knowledge the first large-scale affect-detection study where commercial VR technologies were utilised (with free-walking capabilities) in conjunction to physiological data acquisition. For this study, the affect detection system (with novel sensor set-up) was used as described in chapter 3. The sensor set-up was designed for unobtrusive integration with existing commercial VR devices (i.e., HMDs) leveraging the potential of fully immersive free-walking VR experiences; a medium that has the capacity to become

the ideal experimental tool for behavioural and affective sciences (section 2.3.3). The proposed sensor set-up was mounted onto an existing high-resolution commercial VR headset (which is fitted on the face and head, thus providing constant contact with the skin of the wearer). Moreover, it was designed to transmit continuous readings from the sensors wirelessly, via Bluetooth connection.

For both interaction modes, active and passive, the commercial Vive VR headset by HTC was used[465]. The *active mode* was built in the way that allows for users to explore and interact with the 3D-spaces. The active users were able to walk, explore and even make actions to affect the outcome of an event. In the active VR scenario, interactive stimuli were embedded within the 3-D environments and the participant had the complete control over her movement and duration of interaction with the surrounding stimuli. To control the interactivity[4] levels between the two groups, but also ensure that both are exposed to exactly same scenarios and events for compatibility, pre-recorded videos of real active users' experiences in the VEs were used in the *passive* mode. This way the event triggers were not influenced by the user's actions and the duration of the videos was predefined. The users were merely observers in a vicarious, seated experience. The stimuli were recorded through the point of view of participants within the active scenario and presented to the participants of the passive scenario on 2-D screen within a VR cinema environment. Importantly, one of the main differences between the two scenarios was the level of interactive control that the user had, e.g., being inside a 3-D space versus watching it through a 2-D video (see factors of presence described in section 2.3.3).

For the purposes of this study, three additional VEs were designed apart from the four affective VEs, bringing the total to seven VEs. More specifically, we designed (1) a 'training CASR' (explained in section 3.3), (2) a 'VR adaptation VE', (3) a 'home cinema', (4) a 'Baseline VE', (5) a 'Neutral VE', (6) a 'Positive VE', (7) and 'Negative VE'). The first two VEs were used for training and testing purposes. The training CASR scene was used to train participants how to use the CASR system to rate their felt affect along the x-axis of the controller corresponding to valence and the y-axis corresponding to arousal. The VR adaptation scene allowed

---

[4] Interactivity here is used as the level of enriched interaction the user can have with the virtual world, involving control and sensory factors explained by Witmer and Singer [181] as the degree of control, immediacy, environment modifiability, multimodal presentation, degree of movement perception and active search.

users to perform a similar task to the main study in a dedicated virtual environment that was not used in the main study. This environment allowed them to learn the limits of the walking area and view events happening all round (360°) in the 3-D virtual space. The third was the main VE for the passive group, and the remaining four environments were designed for the Active group (the differences between the two groups are explained in section 6.2.4). The baseline VE was designed for baseline recordings and to allow participants to relax between conditions. The three remaining VEs were designed to provoke negative, neutral and positive affective states. These three environments sharing common architecture, were populated with events designed to stimulate various combinations of valence-arousal. The selected VEs and events were pre-evaluated in an online survey using videos and pictures (see section 5.3.2). The gaze-based interaction algorithm implemented for the active mode of the study, recorded and marked events throughout the participant's experience with the immersive content. These events were then used as markers for the processing and analysis of the streamed data from both the active and passive mode.

The main study involved the execution of both experimental protocols (active and passive modes) at the Science Museum in London for six weeks, outside laboratory conditions. The participants were divided into two independent groups, active and passive group. The aim of this study was to investigate the feasibility of the sensor set-up to detect changes in affect within VR and to investigate the potential benefit from using a highly immersive and interactive set-up. The following related hypotheses (H) were formed:

Behavioural Rating Measures

H1: The VEs evoke the predesigned neutral, positive and negative affective states in both groups as measured by continuous valence and arousal ratings (manipulation check).

H2: Participants in the active groups show stronger affective responses compared participants in the passive group.

H3: Affective positive and negative VEs are more arousing and memorable than the neutral VE.

H4: Participant in the active group feel higher levels of presence compared to the passive group.

H5: Affective positive and negative VE induce a higher level of presence than the neutral VE.

Physiological Measures

H6: Our EMG sensors are sensitive enough to detect posed facial expressions.

H7: Our EMG sensors can reliably detect spontaneous valence changes in passive and active VR settings.

H8: Our PPG sensor and the combination of EMG and PPG sensors can reliably detect arousal changes in passive and active VR settings.

H9: Despite higher noise levels on physiological sensors, stronger affective responses (and thus VEs should be more distinguishable) in the active group than the passive group because of higher presence and interactivity levels.

Computational Modelling Outcomes

H10: Both valence and arousal classifiers perform well when data from participants are used for both training and testing sets (for mixed-subjects user-independent approach against separated-subjects).

H11: Training a predictive model per user (user-dependent) could provide higher performance in mean classification accuracy than generalised models (user-independent ones).

This chapter's outline will describe the experimental design and the analysis that was conducted to investigate the described hypotheses. The VR study (FEDEM3) methods section is described in section 6.2, including information on the participants, the materials and methods, and the experimental procedures. The signal pre-processing steps for data analysis used for the recorded data in the main study for both groups are described in Section 6.2.5. The results from the analysis of the valence and arousal self-ratings, the memory accuracy scores, presence scores, the expression imitation exercises (voluntary), the physiological signals based on event-

markers (event-based analysis), and the classifications findings of spontaneous responses for each group are described in Section 6.3.

6.2. Methods

6.2.1. Participants
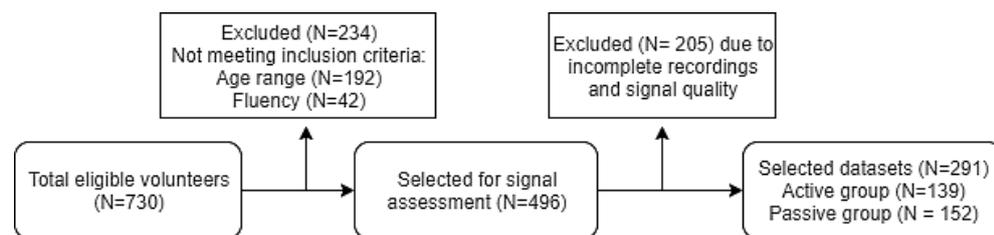
Figure 55. Dataset selection flowchart

A group of 730 volunteers between ages of 14 to 45 years (M = 26.25 ± 10.40) participated in the study. From those, 496 participants were initially selected because they were within the age range of 18-35 years and fluent English language speakers. The selected group suffered from neither cardiovascular, medical, or psychological conditions. After reading the participant information sheet, written informed consent was obtained at the beginning of the study while being introduced to the experiment protocols. The study was approved by the Bournemouth University Research Governance and Ethics Committee (ID: 18848).

To assess if the participants had trouble understanding the task, we screened them by running a short verbal rating session. In practice, we asked them to think of four different scenarios and rate their valence and arousal for each one. Data from participants with no fluency in English and the ones unable to follow the rating test were not recorded. Sessions with participants of low English fluency were also not completed and the data from those participants (N=42) were not included in the analysis. Additionally, we excluded datasets from participants with incomplete recordings and with noisy data (e.g., intense noise artefacts on the signals or low signal-to-noise ratio, e.g., speaking during the stimuli elicitations and chewing gum). Data were also excluded when technical problems occurred, such as sensor positioning (e.g., because of narrow faces or small heads), and unstable sensor signals (see Figure 57).

The remaining number of participants was n=291 with a mean age of 24.76 (± 4.61), of which 163 were female (56%) and 126 were male (43.3%). The selected participants had little to no experience with VR (81.44%). From those n_a=139 participants were in the 'Active' group (76 females (54.7%) and 63 males (45.3%), with a mean age$_{Active}$= 25.22 (± 4.69)), and $n_{p=}$152 in 'Passive' group (87 females (46.1%), 64 males (42.1%), and 1 transgender/non-binary individual (0.7%), with a mean age$_{Passive}$ = 24.35 (±4.51)). The data recorded from these participants formed the database which was used for the analysis that will be presented in the following sections.

6.2.2.   Materials and Methods

Virtual Environments (VE) – Seven virtual environments were designed to study affective responses: ((1) 'training CASR' (explained in section 3.3), (2) 'VR adaptation VE', (3) 'home cinema', (4) 'Baseline VE', (5) 'Neutral VE', (6) 'Positive VE', (7) 'Negative VE').

The *training CASR* was a simple VE with a screen where instructions were given on how to self-rate using the CASR tool, upon a wireless hand controller (Figure 58).
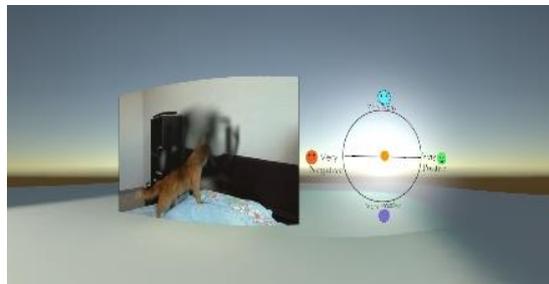
Figure 56. 360º view of the CASR-training VE. The screen in front of the user played four short videos. The AV space on the right of the screen gave visual feedback of the user's rating.

The *VR adaptation VE* one VE designed to assist users familiarise themselves with the VR technologies used, the movement boundaries and the ability to explore spaces by looking at 360 degrees. This was an outdoor VE, consisting of trees and moving elements which the user could practice exploring while self-rating. The VE featured a 3D path that has the exact same size as the VEs office-replicas.

The dark *home-cinema VE*, was used mainly for the passive mode. This VE was also used for the recording of posed facial expressions. It included a screen for video presentation visible to the participant, and a user interface button to select videos which was only visible from the experimenter's desktop view.

The remaining four VEs used for active mode data collection (see Figure 59). The four affective scenarios were alterations of one 3D office room replica, consisted of: a) the *baseline*, b) the *neutral*, c) the *positive* and d) the *negative* version of the room. In all the virtual office-room versions, the dimensions of the room, together with the point of entrance of the user were kept identical. Variations were made in their content to allow for affective manipulations within VR. Section 5.4 contains detailed descriptions about these VEs, their variations and the specific events/objects.
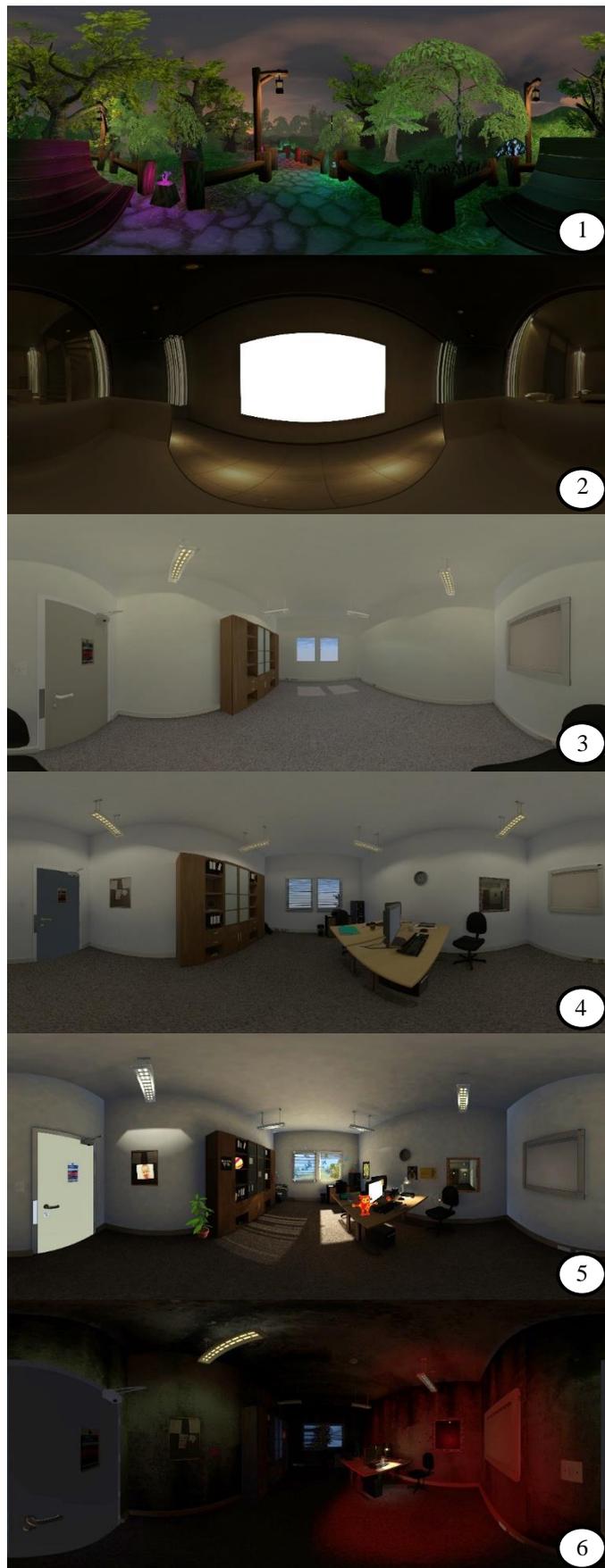
Figure 57. 360o screenshots of the VEs: (1) VR adaptation, (2) VR cinema, (3) Baseline VE, (4) Neutral VE, (5) Positive VE, and (6) Negative VE.

Figure 58. Sketch of floor plan and photo of entrance area of the 'Who Am I ?' gallery at the Science Museum London. The study was conducted here. The space was designed in a way that two sessions of active VR and two sessions of passive VR could run simultaneously.

Physical Space – The study was conducted at the 'Who am I?' gallery space within the Science Museum in London. The area was divided into two main spaces, the main gallery, consisting of six desks with computers and the storage room, which was emptied for the purposes of the study. Two sessions of active VR and two sessions of passive VR could run simultaneously as shown in the floor representation (left side) in Figure 60. The red dots correspond to the locations of the participants in the VR sessions. The room for movement is outlined by walking-space (blue rectangles) for active group setups, and by sitting-down (blue circles) for the passive group setups. The remaining desks and desktops, and four tablets were used for the information of users, the completion of the questionnaires and additional activities for the children of volunteers, including colouring exercises using pictorial representations of facial expressions.

The dimension of each active VR walkable area was approximately 2.5m x 3m, including the desk where the experimenter was sitting. Those desks were successfully masked (and thus avoided by participants wearing the headset during the study) by positioning virtual desks and objects on the same place the physical ones were, following the idea from [466].

6.2.3. Apparatus and instrumentation

Four HTC Vive HMDs were used, together with 6 EmteqVR interfaces. Additionally, to ensure participants wearing the HMDs would not accidentally touch or hit a physical object outside the interaction area (the '*play-area*'), two external webcams were positioned in the corners of room to monitor the participants' movement. Two additional web-cameras were positioned on the wall, in the physical location of where the virtual mirror would be in the VEs. The feed from the camera was used as the reflection for the virtual mirror. This effect was only visible in the positive scenario, and when the users' gaze was hovering over the mirror. That way the participant looking at the virtual mirror would see their real self-mirrored back. The area per active VR system was calibrated using two HTC Vive base stations at an approximate height of 2.5 m (2 base stations per headset). The passive VR headsets were calibrated using the same base stations.

Six desktop computers (OS: Windows 10) were used overall, i.e., two for the passive VR setup, two for the Active VR setup, and two for the completion of forms and questionnaires. Additionally, four tablets (OS: Android) were used for the completion of forms and questionnaires. A pair of headphones was connected with each HMD using the VIVE deluxe audio strap including easily adjustable headphones [467]. The view from each participants headset was streamed on the experimenter's screen connected to each computer. The OBS software was used to combine and record simultaneously the view of what was displayed in VR together with the corresponding camera's feed of the user. The collected videos were used for synchronisation and detection of movement artefacts.

6.2.4. Experimental Procedure

The experimental procedure of this study entailed two main phases, i.e., the VR experiment phase and the questionnaire phase. The order of these parts was counterbalanced across participants (random allocation). The entire study consisted of seven experimental protocol steps which are outlined in Figure 62. These steps are described in more detail below. The entire study had a duration of approximately 40 minutes.

Step 1 - Introduction: Once a participant volunteered to participate, they were given a physical or digital version of the participant information sheet and consent form.

An experimenter was available to answer any questions and explain the phases of the study. Participants with medical, psychological (e.g., anxiety, depression), cardiological (e.g., arrythmia, pacemaker usage), facial related conditions (e.g., facial palsy, stroke) and intense phobias were discouraged from participating. Participant IDs were allocated to the volunteers using randomised 6-digit numbers to ensure anonymity of the data.

At this stage, participants were divided into the two interaction mode groups. The remaining steps were followed by both groups, except for Step 3.

Step 2 –Set-up & CASR rating: The participants wore the VR headset together with the physiological sensors (EmteqVR). Whilst seated on a standard office chair (see Figure 61) they held the controller and rated four different affective videos using the CASR tool while following the instructions of the experimenter who was in close physical proximity. This step was introduced to assess the sound quality and to train the participant in the continuous rating one's own affect in term of valence and arousal. This step had no predefined duration.

Figure 59. The participant is wearing the HTC Vive headset + EmteqVR and headphones, holding a controller. On the left side, one of base-station tripods is visible and a web-camera fixed on the wall. With the exception of EmteqVR, controllers and base-stations, all devices were connected via cable to the PC.

Step 3 –Active Group only: Once participants felt comfortable with rating their affective state, they were asked to enter the 'VR adaptation scene'. The translocation from one scene to the other was controlled externally by the experimenter. Once inside the adaptation scene, participants were asked to stand up and try walking around while exploring every part of the scene within their reach. For the participants

safety the outlining of the walkable area (else called 'play-area' grid provided by the SteamVR application for the Vive headset) was activated when they stepped too close to obstacles or walls in the physical room. After the first exploration, participants were asked to start rating how they feel while exploring the objects around them. This step allowed them to familiarise with the length and width of the walking area and with the concept of continuous affect rating while exploring a VE at the same time. This step had a duration of approximately 3 minutes.

Step 4 - Baseline: Participants were asked to relax before proceeding to the affective scenes while sitting on a chair. Following the data quality control and a recording of one minute, the experimenter would verbally ask a set of questions. Firstly, participants were asked whether they felt dizziness, nausea or any discomfort which originated from the VR simulation experience. Next, they were asked how well they remembered specific events that occurred in the previous VE. Then they were asked to rate their experience in terms of arousal (ranging from 1 (very low) to 9 (very high)) and valence (ranging from 1 (very negative) to 9 (very positive), end of experience rating). Using the same rating scales, participants were also asked to rate how much they enjoyed the VR experience and how present they felt inside the VE [468] (1 (very low) to 9 (very high)). The baseline scene was presented before each affective VE (total 3 times) as seen in Figure 62.

Step 5 - VEs: The participants experienced all three affective scenes (i.e., neutral, positive, negative) in a counterbalanced order following a Latin square. The participants were not aware of the nature of the VEs before they entered them nor of the order of presentation. They were instructed to relax, stand up when ready, and to explore the VE (active group) or watch the videos (passive group) while rating how they felt throughout the entire experience using the CASR and the hand-controller. Once the minimum required duration in VR had passed (75 seconds), participants could exit the VE by touching the virtual handle of door using their controller or by expressing it verbally. Any verbal communication with the experimenters throughout the VE exploration was strongly discouraged, except in the case of sickness and discomfort.

Step 6 - Expressions: After the VR experience, participants were asked to mimic facial expressions following video displayed on a virtual monitor whilst still wearing

the setup. Dynamic expressions compared to static ones were suggested to enhance facial mimicry response [469]. Thus, a video of a person performing facial expressions with corresponding audio instructions was used to direct the users. The video included a short relaxation exercise to ensure they would start the exercise with a neutral, relaxed face. Participants performed the required facial expressions while seated on a chair. In the video, a woman performed the following set of facial expressions while sitting in front of a white background: 1) a closed mouth smile displaying a happy expression, 2) an intense frown displaying an angry expression, and 3) a forehead wrinkle and eyebrow raise as a surprised expression. Each expression had three repetitions which were performed for 3 seconds each, followed by short breaks of 3 seconds between repetitions, and 7 seconds between different expressions. These facial expressions were chosen to record the EMG activation of the facial muscles upon which the EMG sensors were placed (i.e., frontalis, zygomaticus major, orbicularis oculi and corrugator, see 3.2).

Step 7 – Q2: Participants were asked to complete the second set of questionnaires ('Q2') consisting of the TAS-20 on Alexithymia, Expressivity Questionnaire, and Personality Questionnaire (as used in online study survey (section 5.3). See also sections 3.3.1 and 3.6). All questionnaires were designed using the Qualtrics Software [451] (the surveys used are available in Appendix C. The scores from each questionnaire were calculated immediately afterwards, and they were given to the participants at the end of the session, together with a brief description of the questionnaires.
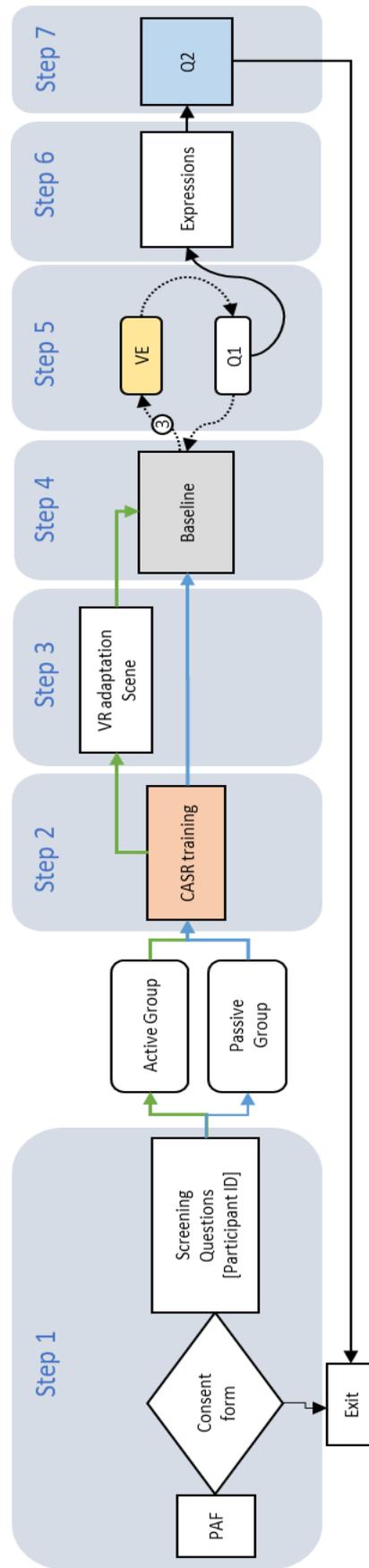
Figure 60. Outline of the experimental protocol used for the VR study divided in steps (1-7). Step 1 included the introduction to the study, the Participant Information Form (PAF), the consent form, a demographic questionnaire with screening questions and the allocations of participant IDs. Participants were then divided into an active and a passive group. They were trained on how to use the CASR interface using the Vive Controller (Step2). Active group users were introduced to room-scale VR using the VR adaptation scene (Step 3). Next, both groups experienced the affective scenarios preceded by a baseline recording session and followed by a short recording of facial expressions of three emotions. In the end, participants were asked to complete questionnaire (Q2) which included questions about their personality, as well as alexithymia and expressivity scales.

6.2.5. Signal Pre-Processing for Data Analysis

Electromyographic signals (EMG)

EMG data were processed with the *Signal Processing* toolbox in MATLAB using a similar approach as the one described for the study in Chapter 4 (section 4.3). For this study, the low cut of the bandpass filter was adjusted to 50Hz (previously set to 30Hz). The processing chart of EMG signals is shown in Figure 63.
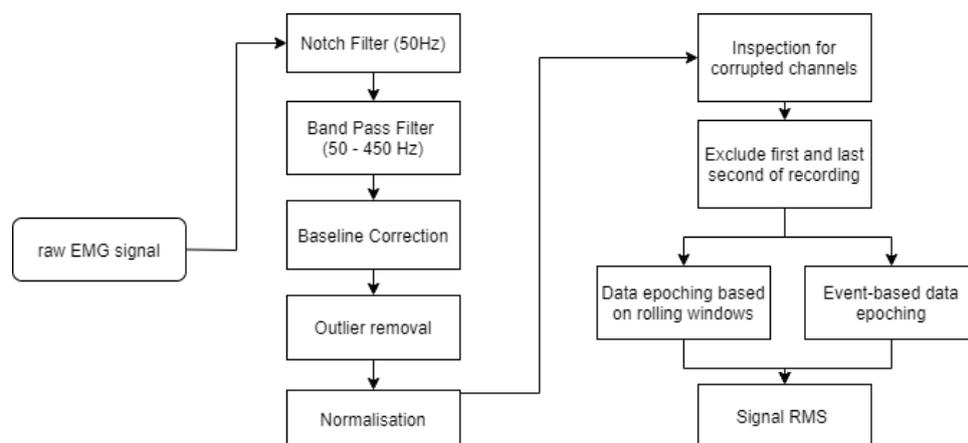
Figure 61. Flow chart of the EMG signal processing steps.

As a reminder, first, data streams from each electrode (Sampling rate: 1000Hz) were filtered and visually inspected for malfunctions and low signal-to-noise ratio which in our case could be cause by interference with movement artefacts, bad electrode placement, or faulty fitting of the sensors. A notch filter on 50Hz and harmonics (from 100 to 450Hz) was applied on all signals prior to other pre-processing steps. Next, a Butterworth bandpass filter at 50-450Hz (6th order) was applied. The low-cut filter was adjusted relative to the Fedem2 processing protocol, to reduce interference caused by movement. Baseline correction was applied by subtracting the mean value of the signal per channel. Extreme outliers were removed using a Hampel filter at a 600 samples (600ms) window. The first and last 1000 samples (1 second) from each recording were excluded from the processed signal, before segmenting data into epochs (details below).

Next, the signals were normalised using the min-max normalisation method (**1**), where the normalised signal was calculated by subtracting the minimum value across the four recordings per user (x_{1-4}: 3VEs and voluntary expressions) and

dividing by the difference of the maximum value across recordings minus the minimum value.

Equation 1. Minimum-Maximum normalisation.

$$x_{norm} = \frac{x - \min(x_1, x_2, x_3, x_4)}{(\max(x_1, x_2, x_3, x_4) - \min(x_1, x_2, x_3, x_4))} \tag{1}$$

Two approaches were used to epoch the data: an event-data based approach and a standard rolling-window approach. In the event-based approach, data epochs were selected based on the event markers. The epoch length has a minimum length of 10 seconds, reaching up to 20 seconds, from the beginning of an event marker (duration of the events was relative to user's interaction). The average range of event-markers per recording was 10-14 per participant. Afterwards, the root mean square (RMS) value of the signal of the epoch (Equation 2) was computed.

Equation 2. Root mean square of EMG signal.

$$RMS = \sqrt{\frac{1}{N} \sum_{n-1}^{N} x_n^2} \tag{2}$$

where x_n is the signals from the EMG channels, and N is the length of x_n.

In the standard-rolling window approach, rolling window epochs with the size of 5000 samples (5 seconds) with 2500 samples (2.5 seconds) overlap between windows were analysed. Since the duration of the recording varied between participants, the number of epochs was relative to the overall duration of the recording. The following descriptive statistics of the RMS signal were calculated per epoch: median, minimum, maximum, and standard deviation.

Photoplethysmographic signal (PPG)

The PPG data from the EmteqVR interface (sampling rate: 1000Hz) was processed to reduce noise artifacts and extract information on the heart-rate peaks (R-R wave). The PPG data was filtered with a Butterworth bandpass filter (0.5 – 4Hz). A Hampel filter was applied (300 samples (300ms) window) to remove any

outliers. The data were synchronised with the EMG data and epochs were created for the EMG and PPG data in exactly the same way to create comparable data.

The R-peaks were detected using the '*findpeaks*' [470] MATLAB's function using custom values to constrain the detection. More specifically, the function was applied twice. In the first run the mean peak amplitude for the whole recording was assessed, excluding the first and last second of the recording. The second time, we run the peak detection function again on the whole signal, by setting the minimum peak prominence to be equal to the previously calculated mean peak amplitude divided by 5 and minimum peak distance set to 300 samples.

The number of beats-per-minute (BPM) and the mean inter-beat-interval (I-B-I) were calculated from the R-R wave. In addition, the pulse-rate variability (PRV) was measured by calculating root mean square of the successive differences (RMSSD), and the standard deviation of the NN (R-R) intervals (SDNN) using the HRV tool designed by Marcus Vollmer [471]. This tool has been used in similar studies for HRV related biometric measurements in the past [472]–[474].

Self-rating values from CASR

The data from the CASR tool, were synchronised with the physiological signals using the system time stamps. The data were epoched along with the EMG and PPG signals while considering the average human response-delay of 200ms [475]–[477]. Median valence and arousal values were calculated for each epoch.

6.2.6.   Machine learning strategy & approaches.

A rolling window-based dataset was constructed by combining the data corresponding to each epoch from all modalities used, i.e., EMG, PPG, CASR. Two different approaches were adopted (see also section 3.6),

  a) a user-independent approach by building a classification model based on the combination of data across participants, and then data is randomly split to cross-validate the classifiers,

  b) a second user-independent approach in which data from users were kept separately for training and testing of the models

  c) A user-dependent "causal" approach, by building a personalised model per user. We used data obtained from the start of the recording session

for training the classifiers, and then validated them with the subsequent data of the user's session (see details below).

For both approaches we used three classifier methods, i.e., a C-Support Vector Machine (SVM), a Naive-Bayes (NB), and a k-nearest neighbour (KNN) classifier method. These classifiers have been suggested and used in affect recognition studies before and they showed promising results [327]–[328]. More specifically, KNN was used for f-EMG and heart-rate signals [376], [377], SVM for the discrimination of facial muscular activations [369], [378]–[381], and NB for affect detection from physiological signals [310], [382], [383]. Thus, they were our choice in this study. Automatic hyperparameter tuning was applied to optimise the penalty parameter σ and the kernel function parameter (γ) for the radial basis function (RBF) of the SVM[408], the distance metric and the k variable of the KNN[478], and the 'width' or Kernel smoothing window width parameter of the NB [479]. For binary svm classification the '*fitsvm*' function in Matlab® was used, and for multiclass classification we used the '*fitcecoc*' function [480]. Similarly, the MATLAB® functions '*fitknn*'[481] and '*fitcnb*'[479] were used for the optimisation of the parameters of the KNN and NB classifier.

For the user-independent approach, 30% of the participants were randomly selected for the hyperparameter optimisation, and the remaining participants (70%) were used for the cross-validation. This was done for each classifier. For the user-dependent approach, the first 70% of the data of each participant was used as training set, and the remaining 30% as the testing. In other words, the temporal sequence of recordings was preserved for each user for causality purposes. The sets used for the user-independent approach and the user-dependent's testing sets were normalised by subtracting the median (Md) values and dividing by the median absolute deviation of the training se*t(*s) (Equation 3).

Equation 3. Normalisation formula used on datasets prior to classification.

$$x_{norm} = \frac{(x - Md)}{Md(|SD(x)|)} \qquad (3)$$

The models were evaluated by calculating the accuracy level and F-scores. All the algorithms were developed in Matlab© R2019b.

6.3. **Results**

This section includes a description of the findings, divided into four main sections related to the research hypotheses listed in section 6.1.

1. Section 6.3.1 contain the analysis of subjective self-reported behavioural rating data (hypotheses 1- 5).
2. Section 6.3.2 presents physiological data analysis and classification data analysis of posed facial expressions (hypothesis 6).
3. Section 6.3.3 contain the physiological data analysis from the event-markers in the three VEs for both groups (hypotheses 7-9).
4. Section 6.3.4 contain the classification of valence and arousal changes from rolling time-windows of physiological data of the three VEs for both groups. This section is further divided into three classification approaches: user-independent with and without pooling data from different participants, and a user-dependent prediction approach (hypotheses 10-11).

The post-VE arousal and valence ratings per scene and the continuous self-ratings via the CASR tool are analysed in section 6.3.1.; enabling us to investigate the affective ranges evoked by the VEs, for the active and passive groups (**Hypotheses 1 and 2**). Additionally, the memory accuracy scores per event and averaged for each VE are analysed across participants to allow for the comparison of affective against neutral VE scenarios (**Hypothesis 3**). Next, the presence scores as reported in the VR experience are compared between the two interaction modes (groups) and between the affective and the neutral VE conditions (**Hypothesis 4 and 5**).

In Section 6.3.2 the analysis of the three voluntary expressions (smile, frown, surprise) the participants were instructed to perform after the VR experience are analysed. This step was performed to test the sensitivity of the EMG sensors to detect these posed facial expressions (**Hypothesis 6**) Additionally, three competitive classifiers were used to detect these expressions.

In Section 6.3.3 an event-based analysis of spontaneous responses per VEs presented using data recorded from the EMG and PPG sensors throughout the VR experience. This analysis allowed the view of the physiological changes caused by event-stimuli,

and the strength of the sensor measures on discriminating between the affective VEs (**Hypothesis 7,8**). The differences of the measures to discriminate the VE conditions are examined per group (**Hypothesis 9**).

The final section, 6.3.4 includes on the classifications tests we performed on the spontaneous responses collected via physiological data during the VR experience. EMG data were used for the training of classification models for valence detection, and PPG data were used for the arousal detection. The continuous self-ratings were used to evaluate the performance of the classifiers (additionally for Hypothesis 7 and 8). For all analyses described, the classification tests are performed on data from the total participant population, and also separately for the active and passive groups. The user-independent classification tests were performed twice (per approach), in order to explore the relationship between classicisation performance and the ability of the trained model to detect affect reliably using new users' data (**Hypothesis 10**). These were: a user-independent approach by pooling users' data together, and a user-independent by separating data from participants between training and testing sets. Additionally, the superiority of user-dependent (one model per user) classification against use-independent models was explored in the last subsection (**Hypothesis 11**).

6.3.1. Participants' self-assessment: Valence and arousal ratings

In this section, the participants subjective valence and arousal ratings are analysed. This analysis relates to the hypotheses 1, 2, 3, 4 and 5.

6.3.1.1. Mean valence and arousal ratings (post-VE).

Table 21 and Figure 64 display the mean valence and arousal ratings that were reported by the participants after each VE experience. The ratings are shown for each VE (neutral, positive, negative) and the passive and active groups separately. For the VE ratings, they seem follow the expected pattern of results. These data were analysed with mixed 3x2 ANOVA with the within-participant factor VE (positive, neutral, negative) and the between-participant factor Group (passive vs. active group) for the valence and arousal ratings separately.

Table 21. The mean valence and arousal ratings reported by all the participants (overall), the participants within the Active group (Active) and the Passive group (Passive). Standard deviations are presented in brackets. Valence ranged from 1 = negative, to 9 = positive, and arousal from 1 = low to 9 = high.

	Negative VE		Neutral VE		Positive VE	
	Valence	Arousal	Valence	Arousal	Valence	Arousal
Total (N=291)	3.39 (±1.79)	6.25 (±1.83)	4.62 (±1.45)	2.65 (±1.85)	6.79 (±1.61)	4.98 (±2.03)
Active (N=139)	3.88 (±2.13)	6.50 (±1.77)	4.96 (±1.52)	2.78 (±1.89)	7.43 (±1.25)	5.41 (±2.11)
Passive (N=152)	3.39 (±1.79)	5.80 (±2.03)	4.41 (±1.49)	2.52 (±1.78)	6.20 (±1.65)	4.51 (±2.09)
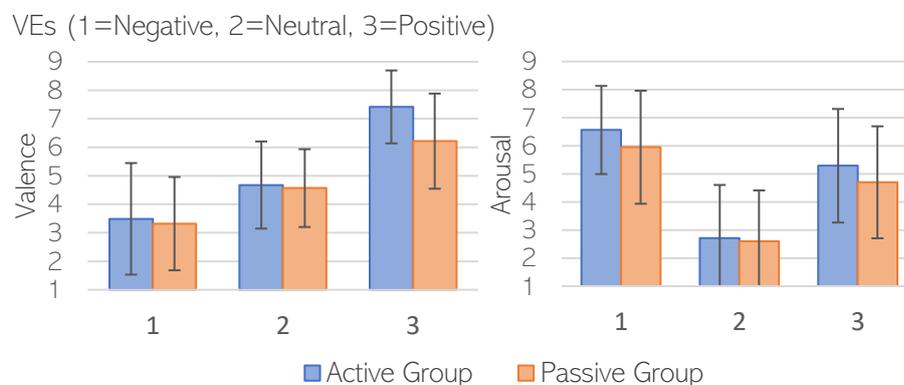
Figure 62. Mean valence ratings (left) and mean arousal ratings (right) of VEs per group (VEs: 1= Negative, 2 = Neutral, 3 = Positive).

Valence Ratings: The mixed ANOVA results revealed a main effect of VE ($F(1.877, 542.532) = 340.270$, $p < .001$). Post-hoc paired t-tests showed that valence ratings were significantly different for positive vs neutral VEs (M±SD: 6.79±1.61 vs 4.62±1.45; $t(290)=18.285$, $p<.001$), negative vs. neutral VEs (3.40±1.80 vs 4.62±1.45; $t(290)=8.674$, $p<.001$) and positive vs. negative VEs (6.79±1.61 vs 3.40±1.80; $t(290)=23.694$, $p<.001$). Secondly, the main effect of group was also significant ($F(1, 289) = 22.044$, $p < .001$), meaning that the active group had higher valence ratings across all VEs compared to the passive group (5.19±0.92 vs 4.70±0.90). More importantly, the interaction between the factors VE and group was also significant ($F(1.877, 542.532) = 10.708$, $p < .001$). This interaction was further analysed with parametric post-hoc t-tests. For both groups, significant valence rating differences were present between all three VEs, as assessed with paired t-tests (all t≥5.309, all $p<.001$). Importantly, the interaction was caused by significant group differences in their valence ratings for the positive VEs ($t(280.321)=6.943$, $p<.001$),

but there were no group differences for the neutral ($t(289)$=0.650, p=.516) and negative VEs ($t(270.001)$=0.847, p=.398; see Table 21 and Figure 64, left panel).

Arousal Ratings: The mixed ANOVA results revealed a main effect of VE ($F(1.987, 574.188) = 337.599$, $p<$ .001). Parametric paired t-tests showed that arousal ratings were significantly different for positive vs neutral VEs (M±SD: 4.98±2.03 vs 2.66±1.85; $t(290)$=17.057, p<.001), negative vs. neutral VEs (6.25±1.83 vs 2.66±1.85; $t(290)$=24.584, p<.001) and positive vs. negative VEs (4.98±2.03 vs 6.25±1.83; $t(290)$=9.071, p<.001). Secondly, the main effect of group was also significant ($F(1, 289) = 7.991$, p=.005), meaning that the active group had higher arousal ratings across all VEs compared to the passive group (4.85±1.27 vs 4.43±1.30). The interaction between the factors VE and group was not significant ($F(1.987, 574.188) = 1.954$, $p=$ .143; see Table 21 and Figure 64, right panel).

Inter-rater agreement. The agreement scores between participants were calculated using the coefficient of variation (CV) per dimension. Higher agreement (<55%) was found amongst participants ratings for the affective scenes (Negative VE: CV = 52.81% (valence), 29.25% (arousal). Positive VE: CV = 23.73% (valence), 40.63% (arousal)), with the exception of the neutral VE which scored high only on valence agreement (CV = 31.26%) but low on arousal scores (CV= 69.56%). The points of Figure 65 shows the mean arousal and valence scores per VE together with their standard deviations (error bars). The CV is represented as the mean value between the two axes per VE as the width of the circle.
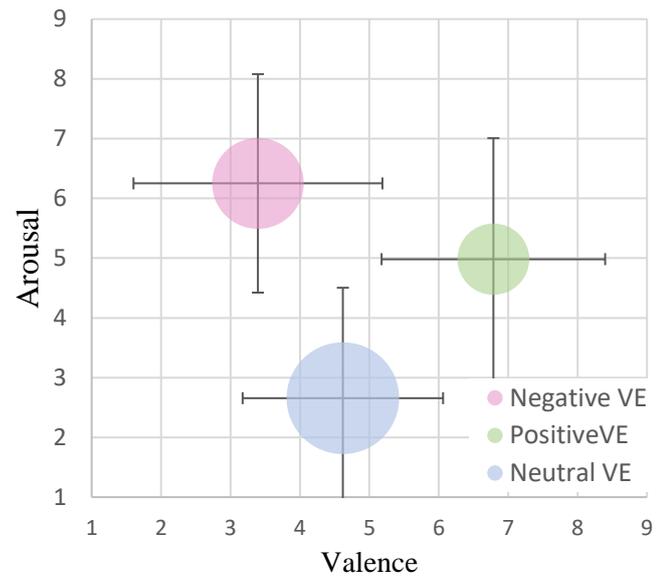
Figure 63. Mean agreement scores per VE represented by circles. The centre of circle signifies the mean value for the arousal and valence ratings per VE, while the width represents the mean CV scores for arousal and valence. The vertical and horizontal lines represent the standard deviation of ratings per axis.

6.3.1.2. **Event-related valence and arousal CASR rating scores**

Figure 66 shows the mean valence and arousal CASR ratings for each VE and group (active and passive) separately. In this initial analysis, CASR ratings were averaged for events of a specific VE.
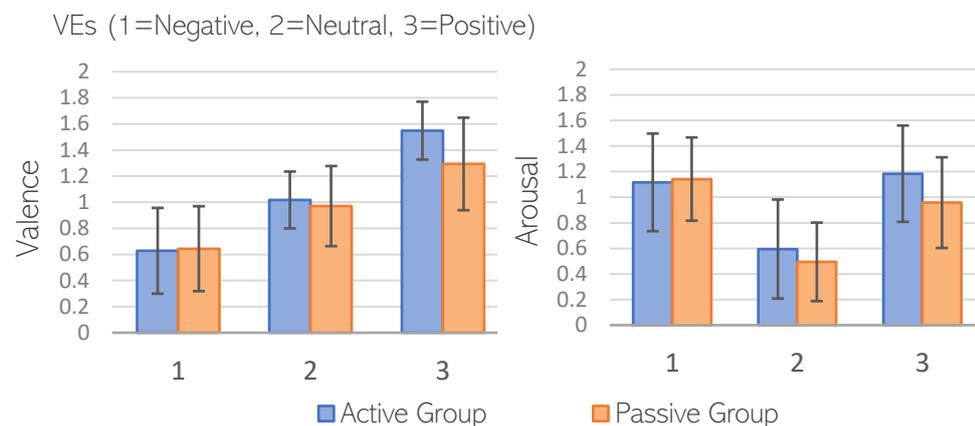
Figure 64. Mean valence CASR ratings (left) and mean arousal CASR ratings (right) of VEs per group (VEs: 1= Negative, 2 = Neutral, 3 = Positive).

CASR valence ratings – A mixed ANOVA results revealed a main effect of VE (F(1.905, 550.515) = 511.975, $p<$ .001), meaning that valence ratings were significantly different for positive vs neutral VEs (M±SD: 1.41±0.32 vs 0.99±0.27; t(289)=18.671, $p<$.001), negative vs. neutral VEs (0.63±0.33 vs 0.99±0.27; 8.007, t(289)=14.255, $p<$.001) and positive vs. negative VEs (1.41±0.32vs 0.63±0.33; Z=13.560, 28.127, $p<$.001). Secondly, the main effect of group was also significant (F(1, 289) = 204.830, $p<$ .001), meaning that the active group had higher valence ratings across all VEs compared to the passive group (1.06±0.17 vs 0.97±0.20). Most importantly, the interaction between the factors VE and group was also significant (F(1.905, 550.515) = 16.715, $p<$ .001). This interaction was further analysed with post-hoc t-tests. For both groups, significant valence rating difference were present between all three VEs, as assessed with paired t-test (all t≥8.585, all $p<$.001). However, the groups only differed in their valence ratings for the positive VEs (t(256.460)=7.423, $p<$.001), but not for the neutral (t(272.263)=1.492, $p=$.131) and negative VEs (t(289)=.426, $p=$.670), as assessed with independent t-tests.

CASR arousal ratings - The mixed ANOVA results revealed a main effect of VE (F(1.977, 571.386) = 246.697, $p<$ .001). Post-hoc t-tests showed that arousal ratings

were significantly different for positive vs neutral VEs (M±SD: 1.06±0.38 vs 0.54±0.36; $t(289)=18.670$, $p<.001$), negative vs. neutral VEs (1.13±0.37 vs 0.54±0.36; $t(289)=19.189$, $p<.001$) but not for positive vs. negative VEs (1.06±0.38 vs 1.13±0.37; $t(289)=2.163$, $p=.093$, not significant after Bonferroni correction). Secondly, the main effect of group was also significant ($F(1, 289) = 368.093$, $p<.001$), meaning that the active group had higher arousal ratings across all VEs compared to the passive group (0.96±0.26 vs 0.86±0.23). Finally, the interaction between the factors VE and group was also significant ($F(1.977, 571.386) = 9.472$, $p<.001$). This interaction was further analysed with post-hoc t-tests. For the passive group, significant arousal rating differences were present between all three VEs, as assessed with paired t-tests (all $t(138)\geq4.790$, all $p<.001$). This was not the case for the active group. Here, arousal rating differences were only present between the positive and neutral VE ($t(138) = 13.673$, $p<.001$) and the neutral and negative VE ($t(138) = 11.613$, $p<.001$). However, there was no arousal rating difference between the positive and negative VE ($t(138)=1.605$). Furthermore, independent post-hoc t-tests comparing both groups for each VE showed the following pattern. Arousal ratings differed between the active and passive group for the positive VE ($t(289)=5.261$, $p<.001$) but not for the neutral VE ($t(289)=2.394$, $p=.050$) and the negative VE ($t(289)=0.391$, $p=.554$).

After the analysis of the mean CASR ratings for each VE and group, we investigated the **valence and arousal ratings for each event within each VEs** (based on event-markers) using the CASR ratings. These event-related valence and arousal ratings are presented for each VE and for both groups in Figure 67. The mean arousal and valence ratings of the events are grouped by colour, within cartesian system following the structure of the circumflex model of affect. A so-called 'V-shape relation'[458] is observed with the in-experience continuous ratings in the AV space. The mean rating per event per group are grouped by outline colour as 'A' for Active group (blue outline) and 'P' for the passive group (orange outline). The figure clearly shows that the event ratings follow the expected distribution within the AV space, following the initial stimuli design. However, there is one exception, the passive group rated the positive events as less positive and less arousing and, surprisingly, sometimes even as negative in combination with low arousal, when compared to the active group. These group differences were not found for negative

events, but they might explain the positive valence rating differences between the groups. This will be further discussed in section 6.4)
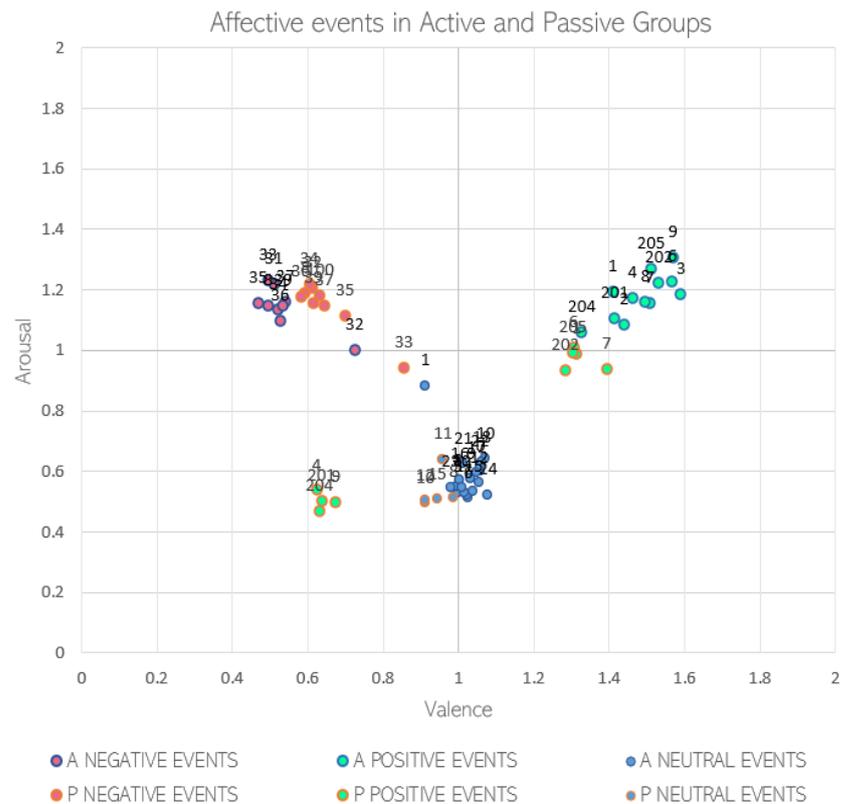
Figure 65. Valence-Arousal coordinates for each event marker, grouped by colour for each VE (negative-pink, positive-green, neutral-blue). The event markers are divided in scores derived from the Active ("A") denoted by a blue outline and the Passive group ("P") denoted by orange outline.

The findings so far have shown that arousal and valence ratings for the positive, neutral, and negative VEs were as expected (manipulation check). This was the case for both the passive and the active groups. More specifically, there was a clear difference between the valence ratings for the three VE conditions for both the end-of-scene ratings and for the CASR event-based ratings. Valence ratings were lowest for positive events / VEs, at a medium level for neutral events / VEs, and highest for negative events / VEs. Interestingly, valence ratings were more positive for the active group compared to the passive groups for the positive VE only, which can be explained when having a closer look at the event-based ratings. These ratings showed that positive events were sometimes seen as positive and sometimes as negative

(with very low arousal levels) in the passive positive condition. Concerning the arousal ratings, these were also as expected for both the end of scene and the CASR event-based ratings. Saying that the CASR arousal ratings were more sensitive to the event-based differences in the positive condition (as already discussed above). Reasons for these event-based rating differences between the passive and active groups in the positive condition will be further explored in the discussion below.

6.3.1.3. **Memory Accuracy Scores**

This section links to **hypothesis 3.** The memory accuracy scores were calculated per event across participants and the average memory accuracy scores calculated per VE. As shown in Figure 68, the mean memory 36.10 ± 12.47 for the neutral VE, 33.24 ± 15.95 for the negative VE, and 37.19 ± 15.03 for the positive VE. The memory accuracy scores are presented separately for the active and passive groups in Figure 69.
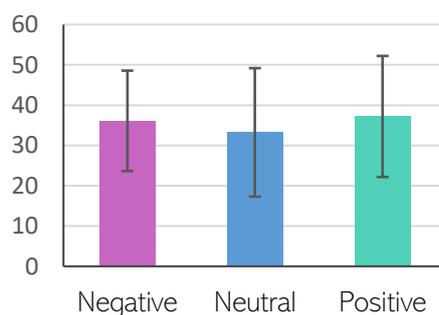
Figure 66. Plot showing the mean memory scores per condition (VE) for all participants. Error bars are standard deviations.
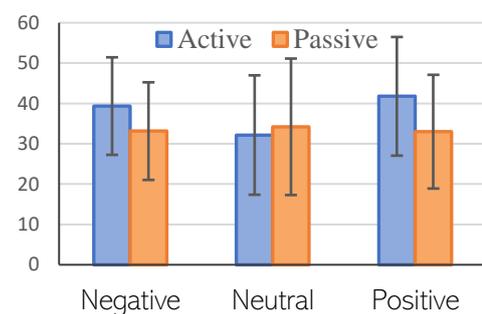
Figure 67. Plot presenting the mean memory scores and standard deviation from the Active (blue) and the Passive group (orange) per VE.

Memory accuracy scores are normally distributed, based on z-score analysis of skewness and kurtosis values. A mixed 3x2 ANOVA with the within-participant factor VE (positive, neutral, negative) and the between-participant factor Group (passive vs. active group) was conducted. The findings showed a significant effect of VE ($F(1.909, 551.595) = 38.731$, $p< .001$) showing that memory scores were significantly different for positive vs neutral VEs ($t(290) = 4.09$, p $<.001$), negative vs. neutral VEs ($t(290) = 3.08$, $p=.006$) and not for positive vs. negative VEs ($t(290)$

= 1.30, p = .57). The main effect of group was also significant ($F(1, 289) = 25.111$, $p< .001$), meaning that the active group remembered the events better than the passive group (Mas_active = 37.76% vs Mas_passive = 33.45%). Most importantly, the interaction between the factors VE and group was also significant ($F(1.909, 551.595) = 24.214$, $p< .001$). This interaction was further analysed with post-hoc t-tests (with Bonferroni Corrections).

When looking at group differences for each VE separately, independent t-tests revealed that the active group had higher memory scores for the positive VE ($t(289) = 5.19$, $p<.001$) and for the negative VE ($t(289) = 4.37$, $p<.001$) compared to the passive group. However, memory scores were similar for both groups in the neutral VE.

The following pattern can be reported, when comparing VEs for each group separately with paired t-tests. For the active group, memory scores were higher for the positive compared to the neutral VE ($t(138)=7.57, p<.001$) and for the negative compared to the neutral VE ($t(138) = 5.99$, $p<.001$). However, there was no significant difference between the positive and negative VE after Bonferroni correction ($t(138)=1.87$, $p=1.89$). For the passive group, memory scores were not significantly different between the three VEs (Positive vs neutral $t(151) = .93$, $p=1.068$, Negative vs Neutral $t(151) = .80$, $p=1.266$, Negative vs Positive $t(151) = .136$, $p=2.676$).

This analysis shows that memory accuracy was higher in the active compared to the passive group. Moreover, memory accuracy was modulated by VE type in the active group but not the passive group. In the active group. memory accuracy was enhanced for the affective conditions (positive and negative) compared to the neutral conditions. This effect was expected and it supports past research on the link between enhanced memory in conditions of affective elicitation (see section 2.3.3.2).

6.3.1.4. **Presence scores**

The analysis of the presence scores is related to **hypothesis 4 and 5** predicting that participants will feel higher levels of presence in the active compared to the passive group, and higher levels of presence in the affective VEs compared to the neutral VE. This seems to be indeed the case, as shown in Figure 70 which displays presence ratings for both groups and all three VE conditions.
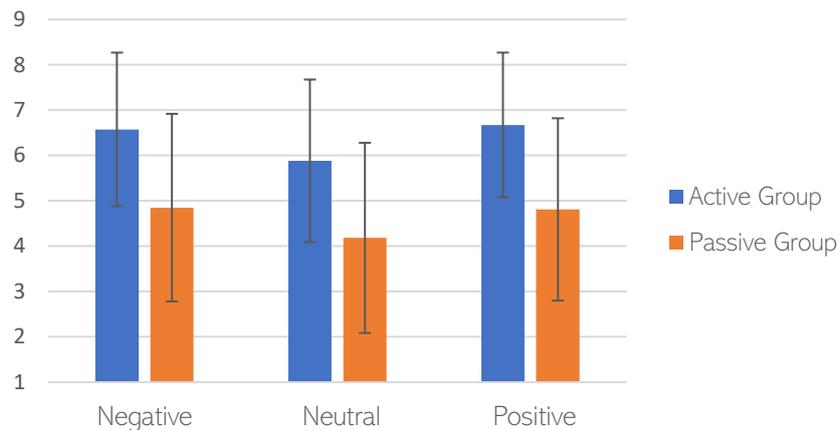
Figure 68. Presence scores of the active and passive group per each VE. Error bars display standard deviations.

A mixed ANOVA with the factors VE (positive, negative, neutral) and group (active vs passive) was conducted to analyse these data. Results showed that the main effect of group was significant ($F(1, 289) = 88.863$, $p<.001$), meaning that the active group had indeed higher presence ratings across all VEs compared to the passive group (6.47 ± 1.42 vs 4.54 ± 2.00). Secondly, the main effect of VE was also significant ($F(2, 578) = 13.724$, $p< .001$). Non-parametric Wilcoxon Rank signed tests showed that valence ratings were significantly different between the positive vs neutral VEs (M±SD: 5.65 ± 2.32 vs 5.11 ± 2.35; $Z=4.734$, $p<.001$), the negative vs. neutral VEs (5.64 ± 2.26 vs 5.11 ± 2.35; $Z=4.634$, $p<.001$) but not between the positive vs. negative VEs (5.65 ± 2.32 vs 5.64 ± 2.26; $Z=0.033$, $p=.974$). Finally, the interaction between the factors VE and group was not significant ($F(2, 577.964) = 0.972$, $p= .379$).

In summary, this finding shows that presence scores were higher in the active compared to the passive group, and higher in both affective VEs compared to the neutral VE. There was no presence difference between the positive and negative VE.

To summarize the findings from this section, the key finding is that we were able to validate the use of immersive VE as emotion induction tool. VE scenes, events and objects that were previously validated with the online survey described in Chapter 5, elicit expected valence, arousal and presence rating in this re-validation within VR settings. The three VEs were able to evoke the targeted ranges of valence and arousal

ratings across all participants. This confirmed out first hypothesis (Hypothesis 1). Note, an exception was the passive positive VE, were some events surprisingly evoked negative, low arousal ratings. Secondly, as expected arousal ratings were significantly higher for the active group compared to the passive group (Hypothesis 2). Thirdly, and in line with the previous findings, memory accuracy was higher in the active compared to the passive group. Moreover, memory accuracy was modulated by VE type in the active group but not the passive group. In the active group. memory accuracy was enhanced for the affective positive and negative conditions compared to the neutral conditions (Hypothesis 3). Finally, presence ratings were higher in the active compared to the passive group (Hypothesis 4) and higher in both affective VEs compared to the neutral VE independent of group assignment (Hypothesis 5).

In the next two sections, the analysis of the physiological measures will be presented, firstly for the voluntary facial expression (section 6.3.2) and then for the spontaneous responses within the three VEs (section 6.3.3).

6.3.2. Analysis and Classification of Voluntary Expressions

This section will report the analysis of the physiological measures from the voluntary facial expressions. This was done to validate and explore the sensitivity of the EMG sensors to detect voluntary facial expressions (hypothesis 6), the voluntary expressions recorded after the VE experience were analysed, and the between-channels comparison per expression are presented in the next section. The three chosen facial expressions are involved in the activation of the muscles underlying the EMG sensor locations [122]. As a reminder, some commonly used emotional expressions are characterised by specific facial muscle configurations [97][482]. Although, the level of activation and morphology of the expressions many vary between individuals, it is supported that in general positive and negative affect can be reliably distinguished from the zygomaticus major and corrugator muscle activity [124]. From those 'basic' or 'predominant' expressions of emotion [97], [118], in our study we chose the facial expressions of happiness (smile), anger (frown) and surprise (raising eyebrows). The EMG recordings of those three expressions, (step 6 of the study protocol, after the VE experiences) were pre-processed, analysed, and then fed into three classifiers. The expression data were extracted from the recording of N=287 participants instead of 291 (due to 2 incomplete datasets and 2 corrupted data sets). The average RMS of the activation of each EMG sensor (referred to as

'channels') was calculated per expression across three repetitions (each repetition lasted 6 seconds).

For each expression, a different pattern of channel activations was expected, see Table 22. Specifically, for smile we expected predominantly higher activation in channels 1,2 (zygomaticus) and some activation on the sensors positioned next to the eyes corresponding to channels 5 and 6 (orbicularis oculi). During frowning, higher activation in channel 7 (corrugator), some activation on channels 5 and 6 on the orbicularis oculi and some subsequent co-activation of the channels 3 and 4 caused by lowering the eyebrows (frontalis)[483] were expected. For the surprise expression, higher activation on the sensors on the frontalis caused by the elevation of the brows, reflected on channels 3 and 4 was anticipated. The EMG measures were hypothesised to be sensitive and capable at discriminating between the three voluntary expressions (Hypothesis 6), which would support the system's ability to detect facial muscle activations across individuals while wearing a VR headset.

Table 22. EMG channels to corresponding facial expressions and the generated actions based on [43], [484]

Selected Basic Expressions	Actions expected	Facial muscles involved	Corresponding EMG channels	
Smiling happily	Closing eyelids	Orbicularis oculi	5	6
	Pulling mouth corners	Zygomaticus major	1	2
Surprise	Raising eyebrows	Frontalis Levator	3	4
	Raising upper eyelid	Levator palpebrae superioris		
Anger	Lowering eyebrows	Corrugator supercilii (effect of frontalis)	7 (3,	4)
	Closing eyelids	Superioris Orbicularis oculi	5	6

Physiological Data Analysis for Voluntary Expressions

Figure 71 shows the median RMS activations and standard deviations per expression for each channel. As shown in the figure, differences in the RMS activation patterns between the expressions are observed which agree with our initial expectations. These patterns were tested with non-parametric Friedman tests because data were non-gaussian.
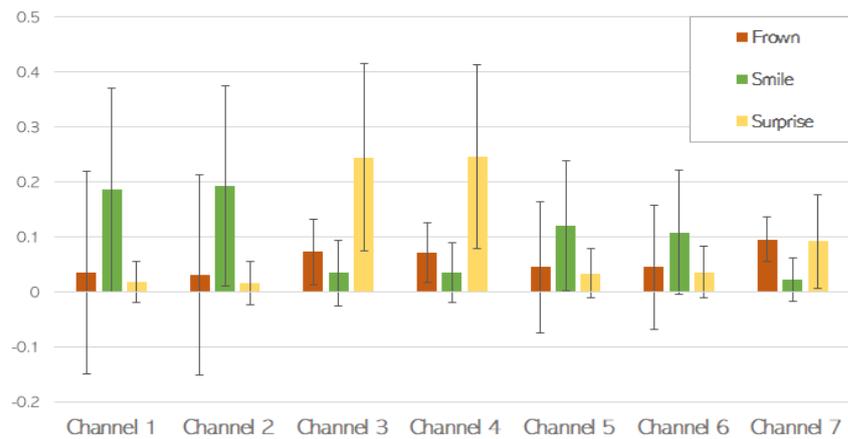
Figure 69. Bar chart showing the median RMS activations per channel for each expression (Frown, Smile, Surprise) with their standard deviations.

RMS comparisons across channels between expressions. Friedman's related samples test showed that the median (Mdn) activations of EMG channels per expression were significantly different ($\chi^2(2)$ = 41.94, $p<.001$), with the smile expression yielding the highest activation compared to the other two expressions. The activation of the smile expression (Mdn: 0.12) was significantly higher than frown (Mdn.:0.07) (Z = -7.95, $p<.001$), the activation of the surprise expression (Mdn:0.10) significantly higher than the frown one (Z= -6.08, $p<.001$) and the median activation during the surprise expression significantly lower than the smile expression (Z = -3.57, $p<.001$).

Comparison between channels per expression. To determine the differences between expressions based on the individual EMG channel activation, related-samples Wilcoxon tests were conducted to compare the median RMS values of the three expressions separately for each channel. The results from the tests are reported for channel pairs (e.g., channel 1 and channel 2, referred to as 'CH.1-2') as these sensors have been placed in mirrored locations, i.e., on the left and right facial muscle of the same type (e.g., Zygomaticus Major left, Zygomaticus Major right), with the exception of channel 7 which is located on the corrugator supercilii muscle.

EMG channels 1-2. Channel 1 (Zygomaticus left) was found to be significantly higher in the smile expression (Mdn: 0.19) than the frown (Mdn: 0.03) and the surprise (Mdn: 0.02) (Smile-Surprise: Z=14.46, $p<.001$, Smile-Frown: Z=10.79, $p<.001$, Frown-Surprise: Z=9.43, $p<.001$). Similarly, Channel 2 (Zygomaticus

right) was also found to increase significantly in the smile expression (Mdn: 0.19) against the surprise (Mdn: 0.02) and the frown (Mdn: 0.03) (Smile-Surprise: $Z=14.40$, $p<.001$, Smile-Frown: $Z=11.00$, $p<.001$, Frown-Surprise: $Z=10.215$, $p<.001$).

EMG Channels 3-4. Channel 3 (Frontalis left) was found to activate higher during the surprise expression (Mdn: 0.24) than the frown (Mdn: 0.07) while remained low during smiling (Mdn: 0.03) (Surprise-Smile: $Z=14.10$, $p<.001$, Frown-Smile: $Z=10.74$, $p<.001$, Frown-Surprise: $Z=-13.30$, $p<.001$). Similarly Channel 4 (Frontalis right) was also significantly higher during the expression of surprise (Mdn: 0.24) against frown (Mdn: 0.07) and smile (Mdn: 0.03) (Surprise-Smile: $Z=14.30$, $p<.001$, Frown-Smile: $Z=11.49$, $p<.001$, Surprise- Frown: $Z=13.35$, $p<.001$).

EMG Channels 5-6. Channel 5 (Orbicularis Oculi left) was significantly higher in the expression of smile (Md:0.12) than frown (Md:0.04) and surprise (Md: 0.03) (Smile-Surprise: $Z=13.92$, $p<.001$, Smile- Frown: $Z=9.75$, $p<.001$, Frown-Surprise: $Z=4.96$, $p<.001$). Channel 6 (Orbicularis Oculi right) followed the same pattern as channel 5, activating significantly higher in the smile expression (Md:0.11) than the frown (Md:0.04) and surprise (Md: 0.04) expressions (Smile-Surprise: $Z=13.73$, $p<.001$, Smile- Frown: $Z=9.66$, $p<.001$, Frown- Surprise: $Z=4.45$, $p<.001$). These channels were expected to activate predominantly during smiling, as in genuine 'Duchenne' smile (section 2.3.2).

EMG Channel 7: Channel 7 (Corrugator supercilii) was activated significantly higher in the frown expression (Md:0.09) against the surprise (Md:0.08) and smile (Md:0.02) expressions (Surprise-Smile: $Z=12.85$, $p<.001$, Frown-Smile: $Z=13.18$, $p<.001$, Frown- Surprise: $Z=3.33$, $p=.001$).

Overall, the results showed significantly higher RMS activation of channels 1,2,5,6 during the smile expressions, channels 7,3,4 for the frown expression and 3,4 for the expression of surprise which was in alignment with our expectations.

Comparisons between the active and passive groups.

As a next step, RMS activations were compared by averaging the three expressions across channels but for the active and passive groups separately. Notably, median

RMS activation were significantly higher across expressions for the active group (Mdn: 0.12) compared to the passive group (Mdn: 0.09), as assessed with a Mann-Whitney U test (U: 6457.00, p <.001). When comparing the groups for the expressions separately, the following results were found. The Mann-Whitney U tests per expression determined that indeed the activation during the expression of smile (Active Mdn.: 0.16, Passive Mdn.: 0.08, U=5,375.00, $p<.001$) and surprise (Active Mdn.: 0.12, Passive Mdn.: 0.08, U = 6,806.00, $p<.001$) were statistically higher within the active group by comparison to the passive. There was no significant group difference in median activations for the frown condition (Active Mdn.: 0.08, Passive Mdn.: 0.07, $p>.05$).

Figure 72 shows the median RMS activation for each channel and expression while comparing the active and the passive group. Group differences in RMS activations seem to be enhanced for the 'expressive' channels. For example, on channel 1 and 2 (Zygomaticus) and channel 5 and 6 (Orbicularis oculi) during smiling. These group differences were tested using Mann-Whitney U tests and significant differences were found between the two groups on channels 1,2,3,5.6 for the smiling expression, channels 1,2,3,4,7 for the frowning expression, and channels 2,3,4,5,6 for the surprise expression (see

Table **23**).
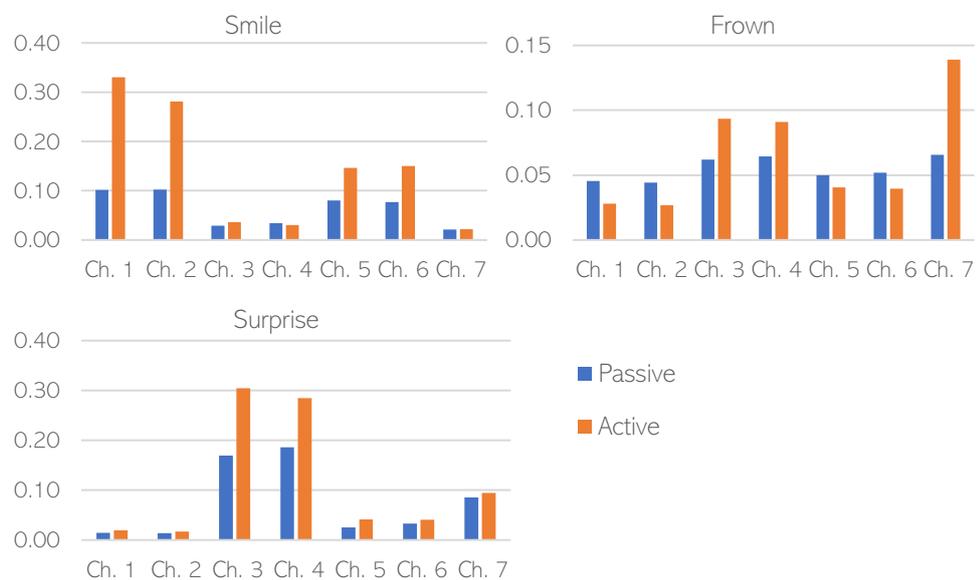
Figure 70. Bar plots showing median activation of each channel per expression for each group (Active (orange), Passive (blue).

Table 23. Table presenting the findings for the comparisons of the active and passive groups for each EMG channels and for each of the three expressions. Data were analysed using Mann-Whitney U tests.

Expression	Channels	U	Sig.(p)
	EMG_channel 1	5,278	**<.001**
	EMG_channel 2	5,821	**<.001**
	EMG_channel 3	8,422	**.009**
Smile	EMG_channel 4	10,272	.995
	EMG_channel 5	6,646	**<.001**
	EMG_channel 6	6,144	**<.001**
	EMG_channel 7	9,265	.153
	EMG_channel 1	12,912	**<.001**
	EMG_channel 2	12,308	**.004**
	EMG_channel 3	7,352	**<.001**
Frown	EMG_channel 4	8,335	**.006**
	EMG_channel 5	11,379	.114
	EMG_channel 6	11,288	.146
	EMG_channel 7	6,294	**<.001**
	EMG_channel 1	8,932	.057
	EMG_channel 2	8,675	**.023**
	EMG_channel 3	6,368	**<.001**
Surprise	EMG_channel 4	7,441	**<.001**
	EMG_channel 5	7,532	**<.001**
	EMG_channel 6	7,880	**.001**
	EMG_channel 7	9,093	.094

Bold at a significance level of .05 / Asymptotic Sig. (2-sided)

Due to the nature of the experiment participants completed the voluntary expression mimicry step after the VR experience. We did not expect to observe significant differences on the overall EMG activation between the two groups, as the group selection was randomised, and the experimental procedure was followed similarly for both groups. It is possible that the overall EMG activation differences between the participants was due to the effect of the VR experience and the interactivity it entailed. Participant of the passive group could be less inclined to perform salient expressions with respect to the participants of the active group. Perhaps such activation difference in the active group could be resulted from the intense muscle activation during the VR experience, which acted like a 'warm up' exercise (similar to those that actors do before a show).

To investigate further this between group difference, correlation analysis was conducted between the mean activation across channels and mean scores of enjoyability (1 value, ranging from 1='I didn't enjoy', to 9= 'I enjoyed very much'), presence scores, the overall duration spent in the VR experience (which was variable for each individual; expressed in seconds), and the expressivity scores for all users. A statistically significant positive correlation (Spearman's correlation) was found

between the mean RMS activation and the mean presence scores (r = .152, *p*=.010) and a significant positive correlation with the mean enjoyment scores (r = .120, *p*=.042). The mean RMS activation was not found to significantly correlate with the duration of the experience and the expressivity scores. These observations allow us to assume that participants who attained higher level of presence (as in the active interaction mode) and the ones who enjoyed the VR experience were more inclined to perform stronger facial expressions, regardless of the time spent in VR or their expressivity scores.

Classification of voluntary expressions

Having analysed the effect of voluntary expressions of the EMG signals which served as a test for the equipment used [485], we examined the feasibility of distinguishing the three voluntary expressions with further classification tests. A standard classification protocol was tested using the voluntary expression. The role of this first analysis was to evaluate the feasibility of decoding facial activations form our EMG signals data before applying it to the spontaneous affect detection data (section 6.3.4) collected during the VR experiences in the VEs. The classification tests and critical results for the automatic voluntary expression detection within VR will be presented in this subsection.

Procedure. The mean and standard deviation of the RMS activation values per expression across all participants were labelled based on corresponding expression (1=smile, 2=frown, 3=surprise) and were fed into three competitive classifiers, namely a Support Vector Machine (SVM) with radial basis function (RBF) kernel, a K-nearest Neighbour (KNN) and a Naïve Bayers (NB). The performance of the three classifiers was tested through via 10-fold cross validation.

To assess the importance of the features derived from the EMG channels towards the classification of the three expressions, the Minimum Redundancy Maximum Relevance (MRMR) algorithm was used [486]. The aim of the MRMR was to measure the ability of the predictors (features) to identify the classification label, by quantifying the redundancy and relevance of the features. These concepts are defined by *the mutual information* of the features. In short, the mutual information measures how much uncertainty of one variable that can be reduced by knowing the other variable [486]–[488]. More specifically, in this approach a large MRMR score represents confidence in the selected feature for predicting the class

label. In practice, by reducing the number of features in a subset, the required computation cost can be significantly reduced. This algorithm has been used in similar studies, e.g. [489], [490].

Classification results. The out-of-sample, 10-fold cross-validated accuracy for the detection of the three expressions achieved was 80.02% using SVM, 79.67% using KNN, and 75.38% using NB, using all features from all the EMG channels (Table 24). The features included average amplitudes (RMS) per channel, i.e., 'mCh1' (mean value for channel 1), and the standard deviation per channel, i.e., 'sdCh1' (standard deviation of values for channel 1). Reducing the features (the predictors) to 8 out of 14 (by setting a threshold based on the standard deviation of the predictor MRMR scores, see Figure 73) minimally reduced the accuracy per each classifier to 79.09% with SVM, 79.21% with KNN, and increased the accuracy to 75.61% with NB   (
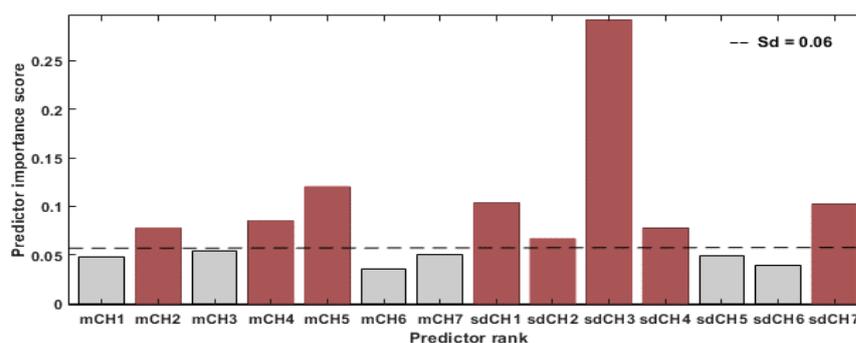
Figure 73. Predictor ranks as computed with MRMR feature selection algorithm. The selected features are the ones above the threshold (line, red filled bars).

Table **25**). Overall, the accuracies of the classifiers ranged from 75% to 80%, with the SVM classifier tending to provide the best classification accuracy. These classification rates indicate that the EMG data recorded across individuals from the interface prototype can be used to discriminate and classify the voluntary facial expressions of smiling, frowning and surprise.

Table 24. Confusion matrix and out-of-sample accuracies per classifier (expressed in percentages) together with the corresponding F-scores.

<div align="center">Confusion Matrix (%)</div>

Classifier	Expressions	Smile	Frown	Surprise	Accuracy	F-scores
SVM	Smile	90.24	6.27	3.48	80.02	[0.84, 0.73, 0.83]
	Frown	20.56	68.99	10.45		

	Surprise	4.18	14.98	80.84		
KNN	Smile	89.90	5.92	4.18		
	Frown	20.21	64.11	15.68	79.67	[0.84, 0.71, 0.83]
	Surprise	3.480	11.50	85.02		
Naïve Bayes	Smile	83.97	14.98	1.05		
	Frown	17.42	66.90	15.68	75.38	[0.83, 0.65, 0.78]
	Surprise	1.74	23	75.26		
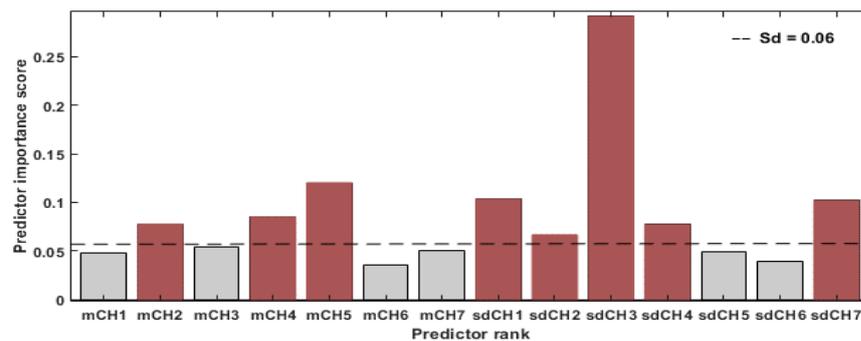
Figure 71. Predictor ranks as computed with MRMR feature selection algorithm. The selected features are the ones above the threshold (line, red filled bars).

Table 25. Confusion matrix and out-of-sample accuracies per classifier after feature selection (expressed in percentages) along with the corresponding F-scores.

Confusion Matrix -50% features (after MRMR)

Classifier	Expressions	Smile	Frown	Surprise	Accuracy	F-scores
SVM	Smile	92.68	3.83	3.48		
	Frown	24.04	64.81	11.15	79.09	[0.84, 0.71, 0.82]
	Surprise	5.23	14.98	79.79		
KNN	Smile	88.85	6.97	4.18		
	Frown	22.30	66.20	11.50	79.21	[0.82, 0.71, 0.83]
	Surprise	4.53	12.89	82.58		
Naïve Bayes	Smile	82.93	16.03	1.05		
	Frown	15.68	71.08	13.24	75.61	[0.83, 0.67, 0.78]
	Surprise	1.39	25.78	72.82		

To summarise the data analysis of the physiological measures and the classification analyses for the voluntary facial expressions, the results were as expected. The facial expressions of smile, surprise, and frown elicited distinct patterns of EMG channel activations when using the EmteqVR device. These differences between channels were statistically significant between the three expressions, and they allowed for the classification of those facial expression with 80% accuracy across multiple users

(**Hypothesis 6**). Additionally, the classifiers performed equally well when using only a subset of the data, deriving from the higher ranked features (79% accuracy), which could reduce computational cost in automatic expression detection in the future.

The active and passive groups showed significant differences in the overall EMG, with the active group achieving an overall higher EMG activation for all three voluntary facial expressions. This effect was unexpected, as there was not variation in the procedure for both groups. This effect might have been induced by the VR interaction mode, which in turn affected the participants' presence and enjoyment scores, along with potentially their motivation to perform stronger emotional facial expressions.

The observed effects for voluntary facial expressions informed the analysis and processing of the spontaneous physiological responses within VR, the analysis of which is divided into two sections, starting with the data analysis of physiological measures in section 6.3.3, followed by the classification tests in section 6.3.4. More specifically, as facial muscle activation can be subjected to interindividual differences due to potential variation in musculature [122] as well as experimental conditions (as in the case of active vs passive group's mean EMG activation difference, see section 6.3.2), a subject-specific normalisation method for EMG was chosen for the classification of spontaneous muscle activations. The data recorded during the three voluntary facial expression (which targeted the facial muscles underlying the skin when the EMG sensors were positioned) were used to inform the subject specific normalisation method, following the process of utilising maximum voluntary muscle contractions for the normalisation of EMG signals suggested by [248] and [342].

Next, the effects of spontaneous affective responses on physiological measures within VR are investigated. The affective impact of the VE events on the EMG and PPG measures was examined for the active and passive groups, related to **Hypotheses 7-9**.

6.3.3. Event-based analysis of continuous physiological data from naturalistic affect elicitation in VR

In this section, three main hypotheses are investigated: (a) that EMG sensors can reliably detect spontaneous valence changes in passive and active settings (**Hypothesis 7**), (b) that PPG and/or EMG sensors can reliably detect spontaneous arousal changes in passive and active VR settings (**Hypothesis 8**), and (c) that affect detection is more reliable in the active compared to the passive group due to higher presence and interactivity levels, despite higher noise levels expected due to room-scale locomotion (**Hypothesis 9**).

For this section, the event-based analysis approach was taken. This means that signals were epoched based on event-onsets (epoch duration based on start and end of event e.g.-250ms to min +10000ms) and then averaged across all events within each VE. Afterwards, a database was created consisting of mean activations for each physiological measure (referred to as features) for each VE. Please note, out of the 291 data sets, 288 were analysed for this physiological data analysis due to technical difficulties, i.e. no event-markers were recorded in three EMG data sets.

Tests of normality showed that all EMG data were non-parametric. Initial ANOVA results are also reported given its robustness in large sample sizes against moderate violations of normality (explained in section 3.6.4). Significant findings were followed up with non-parametric Friedman tests, Wilcoxon signed rank/Mann-Whitney U tests. All correlations were Spearman correlations.

6.3.3.1. EMG analysis comparing VE conditions in active and passive groups

Figure 74 shows the EMG activity for all channels combined. The data show EMG activity for all three VEs as a total across both groups but also separately for both groups. When looking at the figure one can see that EMG activity is less strongly modulated by VE in the passive group compared to the active group. In the active group EMG activity increases from the neutral to the positive to the negative VE condition. This raise is less pronounced or absent in the passive group.
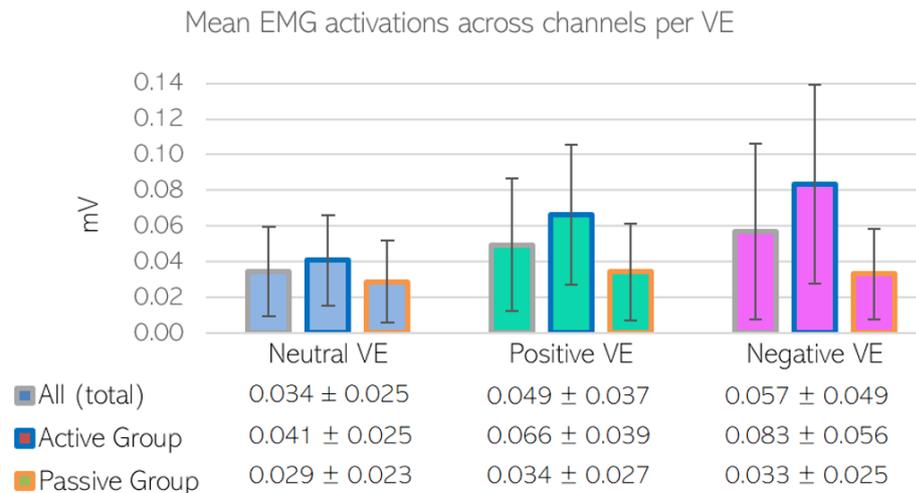
Figure 72. Mean EMG activation across channels per VE (Neutral, Positive and Negative) for (a) all users (Grey outline), (b) users in the Active group (blue outline), (c) subjects in the Passive group (orange outline). The table under the figure shows the means and standard deviations (also depicted in error bars).

The combined EMG activity across all channels, was analysed using a mixed 3x2 ANOVA with the factors VE Condition (neutral, positive, negative) and Group (active vs. passive group). The analysis revealed a significant main effect of Group ($F_{(1,285)}=87.268$, $p<.001$) meaning that EMG activity was higher in the active compared to the passive group. There was also a significant main effect of VE condition ($F_{(1.642,468.037)}=73.932$, $p<.001$), showing that for both groups combined the EMG activity was lowest in the neutral condition, intermediate in the positive condition, and highest in the negative condition. All three conditions were significantly different from each other. In addition, the interaction between the factors VE condition and Group was also significant ($F_{(1.642, 468.037)} = 48.320$, $p<.001$) showing that the modulation of the EMG activity across VEs was more pronounced for the active compared to the passive group.

In the next step, we analysed all EMG channels separately by using mixed 3x2 ANOVAs with the factors VE Condition (neutral, positive, negative) and Group (active vs. passive group). The findings are shown in Table 26. In short, for all channels there was a significant main effect of Group (all F-values >=7.614; all p Value <=.006) showing that EMG activity was always higher in the active compared to the passive group. There was also always a significant main effect of the VE condition (all F-values >=5.940; all p <=.003). The non-parametric post-hoc tests for both groups combined are presented in table 25 below. Finally, there was also a

significant interaction between the factors group and condition (all F-values >= 5.443; all $p<=.006$), with the exception of the findings for EMG channel 7. Here, the interaction was only marginally significant (F(1.693, 484.124)= 2.901, $p=.065$) but it showed a similar pattern as seem for all other channels. This pattern was already described above, EMG activity showed a stronger modulation by VE condition in the active compared to the passive group. A detailed description of the non-parametric post-hoc tests for the active and passive groups is presented in section 6.3.3.3.

Table 26. Mixed ANOVAs findings for each EMG channel.

EMG Channel	Main Effect Group	Main Effect VE Condition	Interaction VE Condition x Group
Combined EMG	Sign (*p*<.001**)	Sign (*p*<.001**)	Sign (*p*<.001**)
Channel 1	Sign (*p*<.001**)	Sign (*p*<.001**)	Sign (*p*<.001**)
Channel 2	Sign (*p*<.001**)	Sign (*p*<.001**)	Sign (*p*<.001**)
Channel 3	Sign (*p*<.001**)	Sign (*p*<.003**)	Sign (*p*<.006**)
Channel 4	Sign (*p*<.006**)	Sign (*p*<.001**)	Sign (*p*<.001**)
Channel 5	Sign (*p*<.001**)	Sign (*p*<.001**)	Sign (*p*<.001**)
Channel 6	Sign (*p*<.001**)	Sign (*p*<.001**)	Sign (*p*<.001**)
Channel 7	Sign (*p*<.006**)	Sign (*p*<.001**)	n.s. (*p*=.065)

* significance level $p <.05$, ** significance level $p <. 01$

6.3.3.2. PPG sensor analysis comparing VE Conditions between the active and passive group

The activity of the PPG sensor was analysed by extracting several HR related features such as IBI (N-N intervals), RMSSD (RMS of N-N intervals), SDNN (SD of N-N intervals), and rBPM (BPM-baseline BPM). Initial ANOVA analysis were conducted to determine whether these features can dissociate between the VE conditions and whether this effect is different for the active and passive groups (see Table 27). For all features, the main effect of Group was significant (all F-values>=12.895; all $p<.001$) meaning that enhanced values for the active compared to the passive group. The main effect of condition was only significant for the IBI (F(1.827, 522.659) = 9.667, $p<.001$) and the rBPM feature (F(2, 572) = 19.706, $p<.001$). This will be further explored with Friedman and post-hoc tests below and in Table 26. The interaction between VE Condition and Group was only significant

for the IBI (F(1.827, 522.659) =3.752, *p*=.028) and the rBPM feature (F(2, 572) =3.774, *p*=.024), meaning that the features behave differently for the active and passive groups across VE conditions. This was further evaluated when conducting Friedman and post-hoc tests for the active and passive groups separately. The findings for this are presented in Table 29 and 28.

Table 27. Mixed ANOVAs findings for the PPG features.

HR Feature	Main Effect Group	Main Effect VE Condition	Interaction VE Condition x Group
IBI	Sign (*p*<**.001****)	Sign (*p*<**.001****)	Sign. (*p*=**.028***)
RMSSD	Sign (*p*<**.001****)	n.s. (*p*=.417)	n.s. (*p*=.087)
SDNN	Sign (*p*<**.001****)	n.s. (*p*=.245)	n.s. (*p*=.094)
rBPM	Sign (*p*<**.001****)	Sign (*p*<**.001****)	Sign (*p*=**.024***)

6.3.3.3. Comparisons between VEs per sensor / feature (both groups combined).

In order to investigate the Main Effect of Condition for all EMG channels and HR features in more detail, Friedman tests and follow up Wilcoxon tests were conducted for each separately. They are presented in Table 28. To facilitate the interpretation of the findings median scores for each VE environment are provided and significant findings are displayed in bold.

Table 28. EMG and HR analysis for the combined data from the active and passive groups. Results of Friedman tests and post-hoc pairwise comparisons between the three VE conditions when significant differences are found for the Friedman tests. Significant post-hoc tests are displayed in bold. Neutral ('Neu.'), Positive ('Pos.'), Negative ('Neg.'). Descriptives include the median (Mdn) scores for each VE.

Friedman's Tests			Descriptives		Pairwise comparisons (Wilcoxon, Bonferroni corr.)		
Features	Chi-Square (χ^2)	Sig.(*p*)	VE	Mdn.	Pairs	Stat. (Z)	*p*
EMG sensors							
EMG Channel1	92.79	<.001	Neu.	.015	**Pos. - Neu.**	**-9.84**	**<.001****
			Pos.	.028	**Neg. - Neu.**	**-8.85**	**<.001****
			Neg.	.027	Neg. - Pos.	-1.31	.189
EMG Channel2	108.67	<.001	Neu.	.015	**Pos. - Neu.**	**-9.11**	**<.001****
			Pos.	.029	**Neg. - Neu.**	**-9.11**	**<.001****

			Neg.	.030	Neg. - Pos.	-1.83	.194
EMG Channel3	2.650	.266					
EMG Channel4	3.092	.213					
EMG Channel5	124.55	<.001	Neu.	.019	**Pos. - Neu.**	**-10.0**	**<.001****
			Pos.	.035	**Neg. - Neu.**	**-9.69**	**<.001****
			Neg.	.034	Neg. - Pos.	-1.34	.180
EMG Channel6	123.83	<.001	Neu.	.021	**Pos. - Neu.**	**-2.54**	**<.001****
			Pos.	.032	**Neg. - Neu.**	**-10.07**	**<.001****
			Neg.	.038	**Neg. - Pos.**	**-9.24**	**.033***
EMG Channel7	3.43	.180					
PPG sensor							
HR. IBI	25.86	<.001	Neu.	.88	**Pos. - Neu.**	**-2.66**	**.024***
			Pos.	.86	**Neg. - Neu.**	**-6.33**	**<.001****
			Neg.	.85	**Neg. - Pos.**	**-3.01**	**.009****
HR. RMSSD	2.63	.268					
HR. SDNN	4.42	.110					
HR. rBPM	28.46	<.001	Neu.	2.30	**Pos. - Neu.**	**-2.79**	**.015***
			Pos.	5.33	**Neg. - Neu.**	**-6.23**	**<.001****
			Neg.	7.18	**Neg. - Pos.**	**-3.27**	**.003****

* significance level $p <.05$, ** significance level $p <. 01$

Overall, affective conditions (positive and negative VE) were related to enhanced EMG activations for the channels 1-2 (Zygomaticus Left-Right), 5 - 6 (Orbicularis Oculi Left and Right), reduced IBI intervals and enhanced rBPM levels when compared to the neutral condition (see 'Pairwise Comparisons (Wilcoxon)' section of Table 28; see also descriptive section). This finding aligns with reports from previous research suggesting that affective stimuli increase heart-rates and enhance the activation of facial muscles (e.g. [492], [493] extended review of externalisation mediums of affect in Chapter 2). Please note, the Friedman's test was not significant for EMG channels 3 and 4 even when the main effect of the VE condition was significant for these channels in the ANOVAs presented in Table 26. As our data are non-parametric, we decided to use the findings from the more conservative Friedman's test. Secondly, the findings did not only show that these physiological measures were able to dissociate between affective and neutral VEs (EMG measures, IBI, rBPM), they were also able to dissociate between the affective positive and negative VEs (EMG channel 6, IBI, rBPM).

6.3.3.4. **Comparisons between VEs per sensor / feature per group.**

We decided to compare VE conditions for each group separately to gain a closer understanding of the significant interactions between VE environment and Group

that were present for all but one EMG channel. For each analysis, Friedman tests were conducted. If they were significant, they were further followed up with Wilcoxon post-hoc tests.

Active Group. In this section, we test the sensitivity of our physiological sensors and features to the VE manipulation for the active group. There was one additional measure called distance (Dis) because participants of this groups were able to walk towards or away to/from a stimulus in the virtual space, the mean distance from stimulus 'Dis' was calculated per user in each VE (equal to $Dis(t_{end\ of\ event\text{-}N}) - Dis(t_{start\ of\ event\text{-}N})$).

Following the same approach described in the previous section, the results from related samples Friedman's tests are presented on Table 29. Bonferroni corrected pairwise comparisons Wilcoxon's test were conducted when Friedman tests were significant.

Table 29. EMG and HR analysis for the 'active' group. Results of Friedman tests and post-hoc pairwise comparisons between the three VE conditions when significant differences are found for the Friedman tests. Neutral ('Neu.'), Positive ('Pos.'), Negative ('Neg.'). Descriptives include the median (Mdn) scores for each VE.

Friedman's Tests			Descriptive		Pairwise comparisons		
Sensors/Features	χ^2	Sig.(p)	VE	Mdn	Pairs	Z	p
EMG Sensors							
EMG Channel1	106.50	**.001****	Neu.	.022	**Pos. - Neu.**	**-8.88**	**<.001****
			Pos.	.048	**Neg. - Neu.**	**-8.70**	**<.001****
			Neg.	.056	**Neg. - Pos.**	**-2.60**	**.018***
EMG Channel2	117.228	**<.001****	Neu.	.025	**Pos. - Neu.**	**-8.57**	**<.001****
			Pos.	.057	**Neg. - Neu.**	**-9.07**	**<.001****
			Neg.	.069	**Neg. - Pos.**	**-5.73**	**.036***
EMG Channel3	9.304	**.010***	Neu.	.051	Pos. - Neu.	-1.91	.168
			Pos.	.049	**Neg. - Neu.**	**-3.23**	**.003****
			Neg.	.052	**Neg. - Pos.**	**-2.58**	**.030***
EMG_Channel4	12.194	**.002****	Neu.	.043	Pos. - Neu.	-1.27	.618
			Pos.	.042	**Neg. - Neu.**	**-3.43**	**.003****
			Neg.	.049	**Neg. - Pos.**	**-2.97**	**.009****
EMG_Channel5	114.978	**<.001****	Neu.	.025	**Pos. - Neu.**	**-8.77**	**<.001****
			Pos.	.052	**Neg. - Neu.**	**-8.99**	**<.001****
			Neg.	.057	**Neg. - Pos.**	**-2.71**	**.021***
EMG_Channel6	105.555	**<.001****	Neu.	.028	**Pos. - Neu.**	**-8.44**	**<.001****
			Pos.	.061	**Neg. - Neu.**	**-8.92**	**<.001****
			Neg.	.071	**Neg. - Pos.**	**-3.18**	**.003****
EMG_Channel7	19.388	**<.001****	Neu.	.016	**Pos. - Neu.**	**-2.66**	**.024***
			Pos.	.017	**Neg. - Neu.**	**-4.51**	**.003****
			Neg.	.026	**Neg. - Pos.**	**-2.76**	**.018***
PPG Sensor							
HR. IBI	24.826	**<.001****	Neu.	.930	**Pos. - Neu.**	**-3.16**	**.006****

			Pos.	.915	**Neg. - Neu.**	**-5.58**	**<.001****
			Neg.	.885	Neg. - Pos.	-1.32	.561
HR. RMMSD	.391	.822					
HR. SDNN	1.696	.428					
			Neu.	3.50	**Pos. - Neu.**	**-3.50**	**<.001****
HR. rBPM	25.739	**<.001****	Pos.	9.83	**Neg. - Neu.**	**-5.55**	**<.001****
			Neg.	10.57	Neg. - Pos.	-2.26	.072
Movement							
Distance	5.666	.059					

The RMS signal calculated from the EMG channels showed significant differences between the three affective VEs, as shown in Table 29. As expected, significant RMS differences were found between all VE conditions for most channels (1, 2, 5, 6 and 7). EMG activation on channel 3 and 4 (positioned on the frontalis muscle, whose activation is linked to mental workload and fatigue [132]) were found to be significantly different between the affective conditions (positive and negative), and also between the negative and the neutral VE. They could not, however, significantly discriminate between the neutral and the positive VE conditions.

Continuing to the analysis of the heart-rate features of IBI and r-BPM (BPM – baseline BPM) showed significant differences between affective and arousing VEs and the neutral, low arousing VE. Unfortunately, both HPV features extracted (SDNN, RMSSD) did not offer any discriminatory differences between the three conditions. The distance feature (distance of user from events) was also not significantly different between conditions (p = .059).

Since the active participants required to explore by physically walking around the area, the physiological signals recorded were expected to carry higher body and head movement-related noise which could also potentially result into crosstalk between sensors. However, the results for the active group showed larger changes between all three VE conditions compared to the results across all users described earlier. This additional sensitivity effect could be resulted from the overall emotional intensity that users could have in the active group experience, which was overall more energetic and exciting for the users than the passive group. These findings support Hypothesis 9.

Passive Group. As for the active group, the ability of each sensor/feature to discriminate between the VE conditions was tested for the passive group. As a reminder of hypothesis 9, for this group we expected weaker affective changes between VEs on both EMG and PPG measures despite the reduced noise levels.

Table 30 presents results from the related samples Friedman's test and the corresponding pairwise post-hoc comparisons with Bonferroni corrections.

Table 30. EMG and HR analysis for the 'passive' group. Results of Friedman tests and post-hoc pairwise comparisons between the three conditions VE when significant differences are found for the Friedman tests; Neutral ('Neu.'), Positive ('Pos.'), Negative ('Neg.'). Descriptives include the median (Mdn) scores for each VE.

Friedman's Tests			Descriptive		Pairwise comparisons (Wilcoxon)		
Features	χ^2	**Sig.(p)**	**VE**	**Mdn.**	**Pairs**	**Z**	***p***
EMG Sensors							
EMG Channel1	13.276	**.001****	Neu.	.008	**Pos. - Neu.**	**-4.11**	**<.001****
			Pos.	.014	**Neg. - Neu.**	**-2.66**	**.008***
			Neg.	.013	Neg. - Pos.	-1.07	.285
EMG Channel2	18.893	**<.001****	Neu.	.010	**Pos. - Neu.**	**-4.40**	**<.001****
			Pos.	.015	**Neg. - Neu.**	**-3.72**	**<.001****
			Neg.	.013	Neg. - Pos.	-.51	.613
EMG Channel3	.487	.784					
EMG Channel4	2.913	.233					
EMG Channel5	27.013	**<.001****	Neu.	.013	**Pos. - Neu.**	**-4.83**	**<.001****
			Pos.	.018	**Neg. - Neu.**	**-3.80**	**<.001****
			Neg.	.016	Neg. - Pos.	-.927	.354
EMG Channel6	32.437	**<.001****	Neu.	.014	**Pos. - Neu.**	**-4.08**	**<.001****
			Pos.	.019	**Neg. - Neu.**	**-4.48**	**<.001****
			Neg.	.020	Neg. - Pos.	-.004	.997
EMGChannel7	3.139	.208					
PPG Sensor							
HR. IBI	9.566	**.008****	Neu.	.843	Pos. - Neu.	-.12	.903
			Pos.	.819	**Neg. - Neu.**	**-2.95**	**.003****
			Neg.	.821	**Neg. - Pos.**	**-3.24**	**.001****
HR. RMSSD	6.171	**.046***	Neu.	.218	Pos. - Neu.	1.73	.083
			Pos.	.237	Neg. - Neu.	.044	.965
			Neg.	.240	Neg. - Pos.	1.45	.147
HR. SDNN	8.895	**.012***	Neu.	.162	**Pos. - Neu.**	**2.22**	**.026***
			Pos.	.178	Neg. - Neu.	.15	.882
			Neg.	.174	Neg. - Pos.	-1.91	.056
HR. rBPM	6.493	**.039***	Neu.	1.69	Pos. - Neu.	.099	.921
			Pos.	3.46	**Neg. - Neu.**	**3.06**	**.002****
			Neg.	4.28	**Neg. - Pos**.	2.57	**.010***

* significance level $p <.05$, ** significance level $p <. 01$

RMS signal differences were observed between the affective, high arousing VE conditions (negative and positive) and the neutral, low arousing VE condition for the EMG channels 1, 2 5, 6. Interestingly, none of the EMG channels showed significant differences that would enable discrimination between the positive and the negative condition, including channels 3-4 and channel 7. In general, we saw lower activation differences between VE conditions for the passive than for the active group for all EMG channels.

Concerning the PPG features, significant differences between the positive and negative affective conditions were observed for the IBI and rBPM features in the passive group. These differences were not present in the active group. Results showed that heart-rate features IBI and rBPM also revealed significant differences for the negative condition compared to the neutral VE conditions. In addition, the SDNN activity was significantly lower in the neutral VE compared to the positive VE. This difference was not found for the active group. The low SDNN found in the neutral stimuli relates to lower variability caused by unpleasant stimuli compared to pleasant ones [494], [495]. However, the difference between the neutral and negative events was not pronounced for this measure, rendering it less reliable compared to the other HR measures. By comparison, the IBI and the rBPM showed good discriminatory ability for the negative events against positive and neutrals. This may be explained by the higher CASR arousal ratings found on the negative events compared to the positive and neutral ones for this group (see 6.3.1.2). The differences indicated that these measures showed similar sensitivity for the passive group as for the active group in detecting increases in negative affect (as a parameter of arousal) but not for the positive events. As seen previously in section 6.3.1.2, the affective impact of the positive events was reduced for this passive compared to the active group, which can explain the reduced discriminatory power of the HR-measures.

In summary, the analysis above clearly shows that physiological measures are able to detect spontaneous affective changes in VEs. We did show that all EMG channels were able to dissociate between VE conditions and that this dissociation was even more pronounced for the active compared to the passive group, despite the enhanced activity (or noise) levels in the EMG channels. In the active group, all EMG channels showed significant differences between VE conditions while for the passive only channels 1,2,5,6 were found to show significant differences between affective and neutral conditions. These results show that spontaneous affect and specifically valence detection in VR settings can be reliable for both active and passive conditions (**Hypothesis 7**). The HR features (IBI and rBPM) showed higher discriminatory power for the affective / high arousing VEs vs the low arousing, neutral VE condition for both groups (**Hypothesis 8**), especially between the negative and the neutral VE condition. Additionally, the IBI and rBPM features differ between the two affective conditions for the passive group, showing an overall weak affective manipulation for the positive VE condition. This agrees with the

effect observed in the CASR valence and arousal ratings for the positive events for the passive group (see section 6.3.1.2, Figure 67).

The majority of studies detecting affect outside VR record data from users in seated positions facing one direction (e.g. [9], [333], [366], [496]). In our study, despite physical movement, the ability of physiological features to detect discriminate between the VEs was stronger overall for the active group compared to the passive (**Hypothesis 9**), which makes the active setting a potential better candidate for future affect elicitation studies. Physical movement can directly affect the physiological signals (e.g. raise heartrate) and degrade signal quality by introducing undesired variations and artefacts in the signal [497], especially in upright positions [498]. This effect was found predominantly in the HR and PRV-related features whose computation can be severely affected by the motion noise [487][499] which can explain the weak effect found on PRV features for the active group. Thus, the computation of PVR measures in VR is more sensitive to motion changes compared to IBI, rBPM and EMG features.

6.3.4. Valence and Arousal Classification from spontaneous affective responses recorded within VEs

In this section we will address the general case, which is the automatic identification of spontaneous affective states which can involve a wide range of expressions using classification methods. Spontaneous expressions can differ drastically from posed or voluntary expressions, and in some cases were found to be less intense and form different muscle configurations (see section 2.3.2 on '*Facial Expressions*'). The aim of this section is to validate the ability of the detection system to automatically distinguish affective changes in valence (**Hypothesis 6**) and arousal (**Hypothesis 7**) in both active (standing) and passive (seated) VR settings from continuous physiological measures. There are several advantages for such an automated affect detection system.

Firstly, the 222event-based data-window approach allowed the analysis of physiological activations for certain events. This event-based approach accounts for only the data collected during the user's interaction with a tagged event in the VR simulation, which allowed the confirmation of the study design to induce the predesigned affective states. However, such approach could be proven difficult to be applied for automatic affect detection in VR as it relies on developers/content creators to tag the events of interest, while potentially interesting data windows (including the transitions between events) would be overlooked. For this reason, a continuous rolling time-window data segmentation approach was applied for the calculation of physiological features recorded per VE condition.

Secondly, machine-learning (ML) classification approaches allow for automatic human affect recognition, bringing us a step close towards natural human-computer interaction. ML methods enable computers to learn directly from data examples, overcoming the requirement to provide an explicit model [500], while taking the simultaneous activity changes and regularities across channels into account. Today, with automatic data analysis and machine-learning approaches we can map data feature combinations to certain labels (i.e., ground truth), thus developing the models which can be deployed to automatically detect those labels in new data (in a dynamic manner) [501]. As such, machine learning approaches can be used to predict future events and to classify existing data. Such techniques have been commonly used in Affective computing, where have shown promise in building automatic detection systems from physiological data [305].

For the purpose of the classification analysis, three classification approaches were performed for the detection of each affective dimension (valence and arousal), as explained in Methods: a) a user-independent approach where all users' data were pooled together and then randomly divided for the training and testing of the models, here referred to as "mixed-subjects" , b) a second user-independent approach in which data from users were kept separately for training and testing of the models referred to as "separated-subjects" (to investigate **Hypothesis 10**), and c) a user-dependent approach in which a model was created for each participant (to investigate **Hypothesis 11**). The classifications tests were performed, firstly, on the data from the total participant sample and afterwards for each group in order to explore the performance of affect detection for each interaction mode (active against passive group) (**Hypothesis 9**).

Following this structure this section is divided in the subsections below:
a) User-Independent Classification (mixed-subjects).
b) User-Independent Classification (separated-subjects).
c) User-dependent Classification -– for each participant separately.

Data structure and Classification Procedure. For these classification tests, additional features were calculated from the EMG and PPG signals. Table 31 contains the list of features extracted from the PPG and the EMG signals which were used for the classification of arousal and valence responses.

Table 31. List of features and features measures extracted per modality

Original signal	Extracted features	Measures from individual features
PPG	IBI	Mean (average) value
	RMSSD	Maximum value
	SDNN	Minimum value
	BPM	Standard Deviation
	Peaks (loc)	Mean of previous 5 seconds
EMG	RMS for EMG Channels 1-7	Mean (average) value
		Maximum value
		Minimum value
		Standard Deviation
		Mean of previous 5 seconds

After filtering (see section 6.2.5), the data were segmented into rolling-windows, where each epoch contained 5000 samples (equal to 5 seconds) with 2500 samples overlap (forward step). Since the duration of the recordings and the interaction with the VE per user was different, the number of epochs per participant also varied.

The 'user-independent 'and 'user-dependent' classification tests were performed separately for binary (two levels: positive/negative) and three levels (negative/ neutral/positive) valence classifications and arousal classification. Each test was run using three classifiers, a C-SVM, KNN, and NB as executed for the expression classification.

By comparison, in the 'separated-subjects' user-independent approach, the 10-fold cross validation was devised by splitting the dataset into folds while also keeping the entirety of each participant's data; leaving 10% entire subjects out in turn for testing while training with the rest (90%), until all subjects have been tested.

For the user-dependent approach, a model was trained per participant, using the first 70% of the data for training and the subsequent 30% for testing that is, the validation respects causality as it is performed on future data. The means accuracies across models overall and per group were calculated. This user-centred approach cannot be generalised to new users, it however was expected to provide a better performance accuracy for valence and arousal detection.

Model testing. As explained in Section 6.2.6 (Machine learning strategy &) the data were divided into sets. In the mixed-subject *user-independent* approach (a), a random selection of 30% of the participants were used to tune the hyperparameters of the classifiers and the 70% were used to classify valence and arousal using 10-fold cross-validation, by concatenating all data from the subjects. This means that during the data division, parts of a participant's data could end-up in the training set and the remaining in the testing sets. This approach was expected to yield good performance accuracy but have lower generalizability to newer subject's data (see section 3.6).

In the *user-independent-separated-subjects* approach (b), after tuning the hyperparameters with the 30% of the participants, the remaining users were divided into 10-folds. Each fold (10% of total participants) was used as testing set while the remaining 90% of the subjects' data were used for training of the model. In this way, the entirety of each participant's data were kept in either the training or testing set,

until all users were tested. This approach is expected to have high generalisability yet to yield lower accuracies than the mixed-subjects approach.

In the *user-dependent* pipeline (c), data form each individual subject is tested separately and preserving causality in the data. That is, the first 70% of the data recorded per subject is used as training, whilst the following 30% for testing. The accuracy from the testing sets was calculated (together with F-scores and area under curve AUC per model which are available in the Appendix A).

Ground-truth class labels. The CASR self-rating scores of each participant were used as the 'ground-truth' labelling for the valence and the arousal classes. For the binary classification of arousal and valence, the CASR scores were bipolarised in high/positive and low/negative using the median (Mdn) score (low< Mdn >high) of each dimension across all subjects. The median division point for valence was 1.05 and for arousal was 1.11 (both ratings ranged from [0 to 2]). The number of time-windows corresponding to each label (class) was balanced: 51% negative and 48% positive valence; 51% for low arousal and 49% for high arousal. For the 3-classes classification the scores were divided using the median plus and minus the standard deviation (Sd: 0.53) divided by 10 (i.e., negative/low class corresponded to ratings between 0 and Mdn-(Sd/10), neutral/medium between Mdn-(Sd/10) and Mdn+(Sd/10), and the positive/high between Mdn+(Sd/10) to 2). The number of segments per class were also balanced for valence (negative 33%, neutral 29%, positive 38%), and for arousal (low 36%, average 30%, high 33%).

Normalisation. The training data sets were normalized by subtracting the median value and dividing by the mean absolute standard deviation over each feature column [294]. The testing data were normalised similarly, using the median values and absolute standard deviation of the training data. The same normalisation approach was followed in the 10-fold cross validation (mixed-subjects) and the 90:10 split folds (separated-subjects, and user-independent). The results for the classification sessions are presented in the following sections.

6.3.4.1. **User-Independent Classification (mixed-subjects)**

A total of 45 models were developed and tested with data from users from the active and the passive group. The out-of-sample accuracies achieved per model are

described in Table 32 and Table 33. The best performance classification for 2-classes of valence was 93.62% in the Active and 93.33% in Passive group, while for 3-classes our models reached 88.91% in the Active and 90.17% in the Passive. For 2-classes arousal detection, our models performed similarly when using only PPG features (86.10% Active, 80.54% Passive). The fusion of features from both EMG and PPG yielded similar accuracies to unimodal approaches (i.e., PPG-based features) for the arousal detection but not for the valence detection, as seen in Figure 75. Interestingly, when EMG features were used alone for the classification of arousal, the classifiers were able to perform arousal detection with higher accuracy than when using PPG-based features (90.93% Active group, 91.71% Passive group). The fusion with PPG features for arousal detection slightly lowered the accuracy performances (80.86% Active, 80.67% Passive).

Table 32. Out-of-sample accuracies per classifier (SVM, KNN, NB) for binary (2 cl) and ternary (3 cl) valence detection using 10-fold cross validation (user-independent, mixed-subjects)

	Valence detection Accuracies (%) (CV-User-independent)					
	All subjects		Active group		Passive group	
	2cl (EMG)	3cl (EMG)	2cl (EMG)	3cl (EMG)	2cl (EMG)	3cl (EMG)
SVM	84.19	73.95	71.89	63.65	80.61	75.75
KNN	94.71	90.95	93.62	88.91	93.33	90.17
NB	72.36	58.88	62.97	53.58	68.49	57.58

Table 33. Out-of-sample accuracies per classifier for arousal detection using 10-fold cross validation (user-independent)

	Arousal detection Accuracies (%) (CV-User-independent)								
	All subjects			Active group			Passive group		
	2cl (PPG)	2cl (EMG)	2cl (fusion)	2cl (PPG)	2cl (EMG)	2cl (fusion)	2cl (PPG)	2cl (EMG)	2cl (fusion)
SVM	66.79	83.45	74.22	69.63	72.28	67.98	67.70	80.09	75.68
KNN	83.53	93.37	81.47	86.10	90.93	80.87	80.54	91.71	80.67
NB	61.36	71.52	59.23	62.25	62.84	53.53	62.50	63.82	58.92
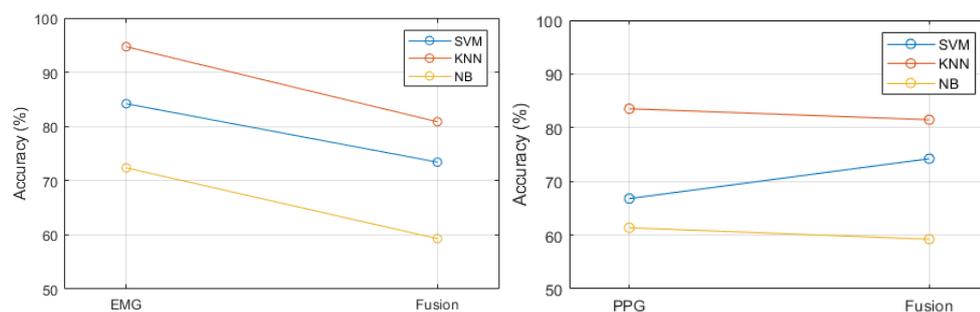
Figure 73. Accuracies per classifier (combined groups) for valence detection (left) and arousal detection (right) using unimodal-derived features against the fusion of both features (EMG and PPG).

Paired t-tests with Bonferroni corrections showed significant differences between the three classifiers on the mean accuracies for all subjects, with KNN (M_{KNN}=88.06) achieving a significantly higher accuracy than the SVM (M_{svm} = 73.86, $t(14)$ = 10.35, $p<.001$) an the NB (M_{NB}=61.99, $t(14)$=20.33, $p<.001$), and SVM achieving a higher accuracy than NB ($t(14)$ = 11.25, $p<.001$). In summary, the KNN classifier achieved the best performance for this classification approach. The classifiers performed statistically equal for the active and the passive group.

6.3.4.2. **User-Independent Classification (separated-subjects)**

A total of 45 models were trained, only this time following the separated-subjects splitting approach.
Table **34** shows the mean accuracies achieved per classifier for valence and arousal detection. The resulted means (and standard deviations) were calculated from the accuracies achieved for all the folds.

Table 34. Mean detection accuracies across folds (separated-subjects, leave-one-fold-out) per classifier (SVM, KNN, NB) for all subjects and for each group, Active and Passive.

<table>
<tr><td colspan="5" align="center">Detection Accuracies (%) (Standard Deviation)</td></tr>
<tr><td></td><td></td><td>All subjects</td><td>Active group</td><td>Passive group</td></tr>
<tr><td>Valence</td><td>SVM</td><td>63.29 (±2.71)</td><td>56.43 (±2.02)</td><td>54.13 (±3.23)</td></tr>
<tr><td></td><td>KNN</td><td>53.19 (±6.21)</td><td>55.71 (±7.17)</td><td>56.98 (±5.41)</td></tr>
<tr><td></td><td>NB</td><td>65.28 (±3.57)</td><td>55.49 (±4.87)</td><td>55.08 (±3.07)</td></tr>
<tr><td></td><td>mean</td><td>60.59 (±4.16)</td><td>55.88 (±4.69)</td><td>55.40 (±3.91)</td></tr>
</table>

Arousal	SVM	60.70 (±4.17)	53.76 (±2.65)	51.58 (±3.30)
	KNN	51.99 (±4.95)	55.21 (±4.47)	52.18 (±3.77)
	NB	61.24 (±0.78)	54.31 (±1.61)	51.79 (±1.55)
	mean	57.98 (±3.30)	54.43 (±2.91)	51.85 (±2.87)

To compare the effects of the classifier and the affective dimension on the classification performances a repeated measures ANOVA was used on the accuracy scores from all users (both groups combined) (classifier (3 levels) x affective dimension (2 levels: valence, arousal)). The results yielded significant main effects ($F(2,18) = 13.607$, $p<.001$). Indeed all pairs compared with pairwise comparisons with Bonferroni correction showed significant difference with each other (val_svm – val_knn: $t(9) = 5.10$, $p=.009$, val_knn – val_nb: $t(9)=-5.69$, $p<.001$, arou_svm – arou-knn: $t(9)=6.15$, $p<.001$, arou_knn – arou_nb: $t(9) = -6.32$, $p<.001$). The SVM and NB were found to perform similarly well for both valence and arousal classification, yielding no significant differences between their performances ($p>.05$).

The main effects of the affective dimension detected (valence and arousal) on the classifiers' performances was not statistically significant ($F(1,9)=4.89$, $p=.054$). The test however indicated a strong interaction effect between the classification performance and the affective dimension ($F(2,18)=31.01$, $p<.001$). Pairwise comparisons between the performance for valence and arousal detection did not yield significant differences on any of the classifiers. A second repeated measures ANOVA on the accuracies scores between the two groups (active and passive) with 3 factors (group (2 levels: active and passive) x affective dimension (2 levels: valence, arousal) x classifiers (3 levels)) showed that the effect of the group was not significant ($p>.05$).

These results showed that when comparing the performances of the classifiers, the KNN and the NB performed better than the KNN for arousal and valence detection. Using this classification approach, no differences were found between the detection accuracies for the active and passive group.

6.3.4.3. **User-Dependent Classification**

Six models for each participant were trained: three for binary valence detection (negative, positive) and three binary arousal detection (low, high) using the three

classifiers: SVM, KNN, and NB. Table 35 shows the mean prediction accuracies, using the testing sets of each user (the last 30% of each recording session). The averaged accuracies across participants are presented for all users combined, and for each group (active, passive), accompanied by the standard deviation of the accuracies achieved across users.

Table 35. Mean prediction accuracies across users per classifier, for (a) valence detection and (b) arousal detection. The columns correspond to the database used (from all users, the active and the passive group).

		Prediction Accuracies (Standard Deviation)		
		All subjects	Active group	Passive group
Valence	SVM	66.51 % (±22.10)	65.81 % (±23.84)	67.28 % (±20.67)
	KNN	69.76 % (±21.06)	69.25 % (±22.94)	70.31 % (±18.84)
	NB	69.67 % (±19.51)	69.42 % (±20.88)	69.93 % (±17.94)
	mean	68.65 % (±20.90)	68.01 % (±22.66)	75.68 % (±16.28)
Arousal	SVM	72.33 % (±19.36)	72.83 % (±20.93)	71.65 % (±16.69)
	KNN	77.80 % (±16.65)	76.60 % (±16.40)	79.68 % (±16.95)
	NB	71.96 % (±16.96)	69.49 % (±18.13)	75.84 % (±14.18)
	mean	74.03 % (±17.66)	72.98% (±18.76)	76.55% (±16.30)

We observe a consistently similar performance between classifiers for all groups, with KNN achieving slightly higher accuracies. To compare the performance differences between the three classifiers, Friedman's tests was conducted on the accuracies for all users combined (as the data were non-parametric), firstly for valence and then for arousal. When significant effects were found, the tests were followed up post-hoc Wilcoxon tests.

The mean differences between the accuracies achieved on Valence detection for each classifier were significantly different ($\chi^2(2)$=10.97, p=.004). Pairwise comparisons using Wilcoxon's signed-ranks test showed significant differences between the accuracies of the KNN and the SVM classifier ($Z = 2.58$, p=.10) and between the NB and SVM classifier ($Z = 2.24$, p=.25), with the SVM performing lower than the other two classifiers. The accuracies achieved for each classifier for each group (active and passive) were also compared using Mann-Whitney U tests, yielding no significant differences (p>.05).

Similarly, the accuracies achieved between the three classifiers for arousal detection were significantly different ($\chi^2(2) = 28.06$, p<.001). Pairwise comparisons

indicated that the accuracies achieved by the KKN classifier were both significantly higher than SVM ($Z=4.21$, $p<.001$) and NB ($Z=4.18$, $p<.001$). The difference between NB and SVM was not found significant ($p>.05$). The accuracies achieved for the active and the passive group were compared using Mann-Whitney U tests, showing the arousal detection accuracies for the passive group were higher than the active group for the NB classifier ($U=5949.50$, $p=.010$).

In summary, the user-dependent approach yielded high valence and arousal detection performances. For the three classifiers used, the KNN was found to achieve the best performances for both valence and for arousal detection. A similar finding was observed for the mixed-subjects' generalised classifications, where KNN outperformed the competing classifiers. By comparison, the KNN performed poorly for the separated-subjects generalised approach, which also yielded the lower performance accuracies. Additional investigation on classification methods and optimisation techniques could be used to improve the attained accuracies, although they were comparable with previous results in different settings [294], [313], [323], [502]. The splitting method used for the separate-subjects approach yielded a lower classification performance but is more appropriate to be generalized to new users' data, as the testing sets included the entirety of new participants data not seen by the classifier (Table 32). By comparison, the mixed-subjects approach although performing significantly better, required an initial subset of all subjects' data to be used for the training of the model (Tables 30, 31). The user-dependent classification results showed that developing a user-centred model provides the high prediction results to future data recorded by the same user (Table 33), albeit requiring an initial subset for the training of each model. Most importantly, these findings confirm the feasibility of automatic affect detection from the physiological features recorded from the system prototype.

6.4. **Chapter Discussion and Conclusions**

The purpose of this study was to explore the feasibility of automatically detecting the affective states evoked while a user is immersed in Virtual Reality using head-mounted wearable technologies. Three pre-validated virtual experiences were used as tool for the induction of neutral, positive and negative affective states. The experiences contained interactive and static events which were triggered by the gaze of the user. Two interaction modes for the VR experiences were explored: the interactive first-person participation in room-scale VR (active participant group), and the passive observation of virtual stimuli from a third-person perspective in VR (passive participant group). The physiological sensor set-ups used (incl. EMG and PPG sensors) and the movement tracking sensors were the same across groups. Participants reported their felt valence and arousal levels in two ways for each VE; firstly, throughout the VE experience using the continuous self-rating tool (CASR) and once at the end of the experience using a 9-point rating scale (post-VE).

Using the data recorded from this study, we investigated the validation of the study design, the effects of the used interaction mode (active vs passive group), and the sensitivity of the sensor modalities to detect changes in expression and affect along the dimensions of valence and arousal. Lastly, the development of a valence detection system and an arousal detection system were also investigated.

Firstly, the self-ratings data were analysed to assess the ability of the three VEs and their events to induce the predefined variations of valence and arousal to all participants in VR. This step served as a manipulation check. The findings showed that the VEs induced the targeted ranges of valence and arousal across individuals which confirmed our first hypothesis (**H1**).

Differences were found between the two groups as reflected in their self-ratings. In line with our expectations, the passive group was less susceptible to the affective manipulations, reporting overall lower arousal ratings than the active group (**H2**) reduced memory accuracy scores (**H3**) and reduced presence scores (**H4**). The affective VEs, positive and negative, were found to be generally more arousing and more presence inducing (**H5**) than the neutral VR. These findings confirm the relationship between affect and presence in VR (section 2.3.3). In the case of this study, the interaction modes examined showed to have a significant effect on a person's experience with VR content, thus affecting the intensity of the emotion induction.

Notably, the continuous affect self-rating (CASR) method for self-rating was more sensitive than the post-VE ratings, revealing additional difference between the two groups. More particularly, the post-VE ratings across individuals showed that the negative VE was rated as the more arousing than the positive VE for both groups. This asymmetry was somewhat expected as suggested by the evaluative space model [87], [88] as negative stimuli generate higher arousal responses than the equivalent positive ones [503], [504] due to the so-called 'negative bias [74]. Interestingly, this asymmetry was not as prominent for the CASR ratings of the active group because the active group did rate the positive events more positively and more arousing than the passive group. By comparison, the passive group's event ratings were sometimes positive but also sometimes slightly negative with very low arousal levels. This was surprising because all of these events were pre-validated with our online survey (see section 5.3). So, why did this effect occur in the passive condition? There are several explanations.

Firstly, valence and arousal ratings are likely to be modulated by the interactivity of the experience. As such, it appears that the interaction mode had a detrimental impact on the perception of positive stimuli and the positive valenced ratings. An advantage of CASR is that physiological measures and self-ratings can be measured in parallel. Hence, a change in ratings over time or with specific events, should also be seen in the physiological data. As such, future studies could investigate in depth the effect of interactivity on the perception of positive stimuli in VR, and potentially analyse the separate clusters of events in the positive VE to see whether these differences are also reflected in the physiological responses.

Secondly, differences were observed between the mean facial activation of the two groups during the post-VR voluntary expressions, showing reduced overall EMG activation for the passive group. This unexpected difference in the findings was analogous to the effect of fatigue on facial muscles, where reduced EMG activation overall is observed [505]. Therefore, the effect of the interaction mode not only showed overall decreased affective stimulation for the passive group and increased negative stimulation (which especially affected some of low arousing events of the positive VE), but also had long lasting effects on the overall the expressivity intensities of the participants.

The virtual objects and events within those VEs evoked mean valence-arousal rating combinations which when graphed into the 2-D affective space followed a V-shape relation which was extremely prominent for the ratings of the

active group. The V-shape relation or 'boomerang shape' [506] considerably evident on ratings of the IAPS picture library, e.g. [405], shows that arousal reflects the intensity of valence, towards higher pleasure (positive) or displeasure (negative). Interestingly, a similar effect was observed in the physiological data recorded during the spontaneous affect elicitation within the three VEs. Heart-rate measures which were traditionally used for the detection of arousal responses, had good discriminatory power between high arousing VEs against the low arousing neutral VE (esp. for the passive group). Similarly, the EMG channel activation measures not only discriminated between the positive and negative conditions (esp. for the active group) but also between the affective conditions and the neutral one for both groups. These findings not only agreed with our hypotheses (**H7** and **H8**) but showed a potential for valence and arousal detection from EMG alone. This observation was further confirmed in the classification experiments, showing high accuracy for valence and arousal using only EMG derived measures.

This study was one of the first to include 3D interactive events in free-walking VE using HMDs for affect stimulation, as the affective libraries with affective content were limited at that time to 2D stimuli. Notably, the event-based analysis of physiological signals showed that data from the active group allowed for stronger discrimination between the VEs, compared to the passive group, despite the increased noise levels (**H9**). This, together with the stronger affective responses observed from self-ratings in the active group, suggests that free-walking interactive VR experiences can outmatch traditional passive seated experiences when it comes to affect elicitation in VR. This finding also suggests that emotion stimulation in VR using virtual passive stimuli, and potentially 360 degree videos (as used in previous VR research, see sections 2.3.3 and 6.1), could be hampered by the choice of the stimuli presentation medium and the reduced user interactivity. The findings for this feasibility study suggest that future research in affective stimulation, can not only manipulate low-level audio-visual features but also use interactive events and objects with affective contextual information to evoke different intensities of valence and arousal. Based on our findings, we would certainly strongly recommend the usage of these interactive VR environments over passive video and VR settings in future studies.

This study also tested an automated detection for voluntary facial expressions and spontaneous affective responses using classification methods. Starting with voluntary expressions, the three posed expressions analysis (smile,

frown and surprise) showed distinct EMG channel activations patterns, yielding 80% classification accuracy across participants (**H6**). These findings confirmed the feasibility of detecting voluntary expressions of positive and negative emotions, as expressed by the activation of zygomaticus major and corrugator (see section 2.3.1). From the three classification experiments, the user-independent mixed-subjects approach (Table 32, 31) yielded the best accuracy in detecting two and three levels of valence, and two and three levels of arousal for both active and passive groups (**H10**). The best competing classifier was KNN achieving 95% accuracy in binary valence, and 93% for arousal detection. Interestingly, arousal detection was achieved from EMG measures alone, at an average 91% accuracy for both groups.

By comparison, the user-independent separated-subjects approach (Table 34) showed reduced accuracies for both dimensions, with an average out of sample accuracies of 65% in binary valence, and 61% in binary arousal across folds, with NB being the best performing classifier. This approach although less performing, provides more conservative results for new unknown users (generalisation) since the entire data from the user's session is new to the classifier. Thus, the development of generalised models from spontaneous physiological signals, still presents challenges when applied to new subjects. This difficulty has been argued in past research using physiological signals for naturalistic affect detection [323]. Nevertheless, the user-independent experiments show promising potential for noisy, naturalistic settings, where users are freely moving their heads and their body. The current experimentations were made using out of box classifiers commonly used in automatic expression detection due to their ease of execution and swiftness in training. Additional exploration of more sophisticated supervised and unsupervised detection models, and optimisation approaches could be applied to yield better performing models.

To mitigate the issue observed in the use-independent approaches, a novel user-dependent approach [323] was presented in which causality was preserved. A model was trained per user, using continuous sequences of data recorded in the study, and using the remaining data of that user for the testing of the model. Overall, high accuracy scores were observed, with all three classifiers performing 69% prediction accuracy on average for valence and 74% prediction accuracy on average for arousal. For this approach, the KKN classifier yielded the higher performing models (70% in binary valence, 78% for binary arousal). Similar performances were observed for both active and passive groups. This approach was further reenforced by preserving

the sequence of the data points using the training, thus preserving causality. The accuracies were higher than the user-independent approach with separated subjects (**H11**). The findings from these experiments confirmed the feasibility of developing user-centred affect detection approaches for VR-based settings.

The introduction of such wearable sensor set-up could open new avenues in affect detect in different settings, including extended realities. Their unintrusive, integrated, miniaturised and mobile nature makes them perfect candidates for virtual reality as they are non-invasive and less distracting than traditional tethered versions. In this study, we showed how virtual experiences could be used efficiently as a mood induction tool. Such experiences could be used in future research studies to elicit naturalistic, spontaneous responses, in passive but also in interactive settings. Future work could investigate the performance of different classifiers in different settings, such as when using different locomotion techniques and in different contexts such as in simulations involving cognitive tasks. Additionally, ensemble and unsupervised ML techniques can be employed to deal with the complexity of physiological data mapping to affective states in adaptive settings [313], [500]. Further to this, affect detection in VR could be benefited from understanding more about the effect of subjective traits on the affect elicitation such as alexithymia and expressivity. Most importantly, we demonstrated the feasibility of designing context-aware system architectures for automatic affect detection VR. This in addition to automatic affect detection can assist future applications in entertainment, simulation training and health-care interventions using VR, by providing an objective assessment of the user's state.

Summary of key findings for Chapter 6

- This was one of the first VR studies to run in fully up-to-date immersive settings with free walking capabilities while comparing behavioural and physiological responses to affective virtual scenarios

- The study achieved the collection of one of the largest biometric datasets outside laboratory settings to our current best knowledge. Participants were recruited from a diverse population and they were highly motivated.

- Participants who actively experienced the VR environment (by freely walking and exploring) had a stronger affective experience and higher feelings of presence compared to the passive group.

- The negative scenarios elicited higher intensities of arousal than the positive scenarios which was expected due to the so-called 'negative bias'.

- Most importantly, we showed the feasibility of detecting affective state in VR settings with the novel sensor set-up: User-independent classification experiments yielded 77-94% accuracy when using mixed-subjects CV, reduced to 55-65% when using separated-subjects CV (high-generalisability). A user-dependent approach achieved mean prediction accuracy of 69-80% for 2-levels of valence and arousal, showcasing the robustness of subject-specific models (low-generalisability).

- We confirmed feasibility to detect affect in seated and in free-walking VR settings.

- VR can indeed be used as a reliable affect induction tool.

Chapter 7

Conclusions & Future work

7.1. Summary of findings

Virtual Reality (VR) is an excellent way to extent cognitive, affective and behavioural research to laboratory settings that are closer to real life environments. One of the key questions in VR research is how people do perceive emotions in this setting and how does that affect the interaction between emotion, cognition and behaviour. One of the key challenges is to develop a VR system that can detect affective responses without having to rely on subjective affective ratings by the user. These ratings can be rather disruptive.

The aim of this thesis was to overcome this challenge by testing the feasibility of combining physiological readings and machine learning approaches to reliably detect changes in affect for VR users. Changes in affective states were explored as a function of valence and arousal. We saw VR as a platform which can offer the ideal laboratory for cognitive and behavioural sciences. As part of our collaboration with Emteq Labs, our team contributed towards the development and testing of the EmteqVR prototype system. This system comprised of a wearable sensor interface and software development tools specifically designed for use in VR settings. As part of this thesis, we proposed that the EMG and PPG integrated sensors on this interface could reliably contribute towards the detection of valence and arousal.

Towards this goal, three complementary feasibility studies were conducted with human participants. The first two feasibility studies explored the detection and sensitivity of the sensor set-up in controlled laboratory conditions using conventional 2D video stimuli. After a series of improvements on the prototype system, a third, larger study was conducted, this time using custom-made 3D interactive virtual environments. As such, the existing methodology was applied in seated and room-scale exploratory VR settings. The findings from this work, not only confirmed the

potential use of VR as an affect induction tool, but also corroborate the use of the wearable sensor set-up for valence and arousal detection in VR settings.

In the subsection 7.1.1 we describe and discuss the development of the initial theoretical design of the first affect detection system framework, to the implementation of this system architecture. The early prototype of the system was tested in the first two feasibility studies which are discussed in section 7.1.2. As part of that section, we also describe the video validation study. The application of the system in highly-immersive VR settings, the creation of the VR stimulus materials, and the findings from the last feasibility study are discussed in sections 7.1.3-7.1.5.

7.1.1.   Affect detection system architecture for VR settings.

The first research aim of this thesis was the identification of an affect recognition model and potential measures for affective state inference in VR for our research (for the complete list of aims, please refer to Chapter 1). The affective sciences literature uses two categories of models to describe emotions/affect. The first model is heavily rooted in distinct emotion state detection, such as the 'basic emotions', which are detected from facial configurations and behavioural patterns. Emotions are highly personal experiences with complex patterns and connections to cultural contexts, which are difficult to be broken down to smaller parts. This model was recently heavily criticised in the affective neuroscience literature and it has been shown that physiological and neuroscience measures are often not able to clearly distinguish specific emotions. By comparison, a more commonly used model to describe emotions in the affective computing and affective neuroscience literature are the dimensional valence and arousal model [305]. Affect is the core function of emotional responses, the energy flow that precedes and produces those altogether experiences. In the dimensional model, physiological and psychological changes can be measures with valence (positive to neutral to negative), and an arousal intensity (low to high) (see section 2.3.1). This way of defining affect, is more useful as it can better relate to other models and related evidence, such as the conceptual act model (see section 2.3.1.3). These models are supported by a recent neuroscientific evidence [507]. Our team decided to mostly build their research on valence and arousal detection methodologies and metrics, with the ultimate aim of combining and applying them in VR settings. Thus, our first affect detection for VR system framework was designed.

Past research on affect detection proposed multiple methods, ranging from

audio-visual, to physiological and motion data for the measurement of the core affect dimensions, valence and arousal (section 2.3.2). However, until now, only few studies have utilised those methods for arousal detection in VR and even less for valence detection (section 2.4). Valence can be determined from facial configurations, e.g., positive expressions versus negative expressions, and behavioural cues, such as approach or withdrawal from stimulus. In fact, the face is visibly for most people one of the richest sources of valenced emotional information on our body [508]. Although, advances in computer vision have brought novel ways of automatically detecting changes on facial configurations, such approaches are not easily applicable in HMD-enabled VR. An alternative approach is the usage of surface EMG sensors to effectively detect the activation of the facial muscles and facial configurations (section 2.3.1). Thus, in our experiments, eight EMG sensors were positioned on the HMD frame on the positions suggested by Boxtel [122] to infer affective states. For the measurement of arousal, a PPG sensor was embedded onto the sensor set-up to provide continuous measurement of the heart systolic peaks, as a parameter of the autonomous nervous system (section 2.3.2.1). The proposed system architecture was depicted in Figure 13, explained in section 3.2.

7.1.2. Confirming sensitivity of sensors to affect changes

The second aim of this thesis was to design a system architecture including an experimental protocol for affect induction in VR. VR was a novel technology at the time when the thesis started, thus only a few potential affect detection methodologies could be applied in VR (see section 2.3.3.3). Besides the technological constraints of working with early prototypes, out team experienced challenges in designing a controlled affect induction paradigm for the study of affect using VR. This was mainly due to the lack of existing immersive VR environments for emotion induction, and the uncertainty over the effect of VR as a technological medium on the affective impact of stimuli material.

For the initial testing of the VR headset prototypes, our team designed a paradigm using affect eliciting stimuli in traditional settings with the prototype sensor set-up. Affect self-ratings tools and audio-visual stimuli materials were explored, and the CASR self-rating tool was designed for our studies. This tool included visualisation of the affective scale onto the user's screen which was linked to an input device (later implemented using a VR controller). Video stimuli were

chosen instead of images/sounds as the closest validated stimuli to the nature of VR experiences (see section 3.5). Moreover, videos from a validated video library [397] were carefully selected using an online survey based on their ability to induce affective qualities corresponding to the four quadrants of the affective model and neutral low arousing affect. This careful approach allowed us to evaluate the data quality of the physiological sensors and their ability to detect different affective states in a systematic way, before using more complex virtual environments which was the third aim of the thesis.

The third aim was to conduct quantitative studies to study the effect of affect changes on the physiological readings for the sensor set-up, and pilot automatic valence-arousal detection. In the feasibility studies discussed in Chapter 4, we attempted to systematically check the sensitivity of the physiological measures to affective changes induced by the selected video stimuli. We started with conventional video clips, before advancing to 3-D animated objects in VR in a later study.

In the first study (Fedem 1) we recorded 35 seated participants and their physiological responses to 40 video clips, and mapped those responses to their valence self-ratings, achieving a mean mixed-subjects accuracy of 82.5% when detecting 3 levels of valence (negative, neutral and positive) using C-SVM (section 1154.2.3). This finding confirmed the sensitivity of the EMG sensors to spontaneous valence changes using traditional stimuli.

In the second study (Fedem 2, section 4.4), the location and sensitivity of a facial PPG sensor to arousal changes was investigated. Data were collected from a subgroup of participants in the same settings described in Fedem 1 study. Readings from the PPG sensor incorporated on the frame along with an ECG belt were recorded from 11 seated participants following the same experimental protocol as in Fedem 1. We chose to position the PPG sensor on the VR insert frame over the temporal artery and vein which was a set-up that had not been tested before. As ECG is the golden standard in detecting heart-rate changes, we compared the performances of the ECG-based and PPG-based derived features.

Our findings showed that as expected the ECG outperformed the PPG in the user-independent approach (across all subjects). However, the PPG-based measures offered a similar or improved performance against the ECG for 9 out of 11 participants when developing classification model per user (user-dependent approach). As expected from similar previous work on valence and arousal detection, user-dependent models performed higher than user-independent (generalised)

models (see section 3.6.4). This effect is attributed mainly due to subjective differences in physiological responses to similar stimuli but also in high inter-subject variability when self-rating affect levels. Additionally, differences were observed in the actual fitting of the interface on different individuals which was relative to facial features, musculature, and bone structure. The main drawback identified was the sensitivity of the PPG sensor to different mask fits on each individual, due to variations in the face shape and sizes observed. Therefore, these findings concluded, albeit from a small sample, that arousal detection from facial-PPG is possible but the set-up needs some improvements. Further pilot studies were executed as part of the prototype's development to improve the mask fit. The findings from these pilot studies were not included in this thesis but the improved set-up was used in future studies.

The first two studies validated the feasibility of the prototype set-up design for valence and arousal detection in ready-for-VR settings. The empirical observations acquired from these pilot studies, combined with the need for careful signal quality assessment for each user informed the development of the next generation of the Faceteq interface, the 'EmteqVR' insert (where softer materials and different forehand curvatures were added onto the interface) and the 'EmteqVR_app' software for real-time signal monitoring. After confirming the sensitivity of our sensor set-up and the performance of the system architecture, we continued by applying them in immersive VR settings.

7.1.3. Validity of the approach in immersive, free-walking VR

The fourth aim of this thesis was to investigate the effectiveness of VR in inducing emotional responsesand the efficiently of reading those via our system set-up (Aims a-d). To this goal, we custom designed and developed a selection room-sized stimuli and environments, given the aforementioned lack in room-scale immersive affect stimulus library (section 5.1). Four versions of 3-D replica of an office room (where the first studies took place) were designed and were populated with context-based 3-D affective events (involving visual, audio and user-triggered animations). A custom gaze-based interaction system and an event-marker annotation system were developed. These systems allowed the continuous tracking of the user's attentional focus on the surrounding environment, and the automatic annotation of the physiological data with object/events specific markers. This interaction systems facilitated the users' free movement within the VR room-scale

environments. Special effort was put in the design of each environment from 3D objects, 3D-sounds, animations, and textures, to physics and light/shadow rendering, to create an overall realistic space for the user to feel presence. In the following section, we discuss our findings on VR being a powerful mood induction medium and that interactivity plays a large role in the affect intensity induced

7.1.4.   Creation of stimulus materials for affect-inducing VR.

Three VEs were developed for the VR study, i.e., a neutral, a positive and negative affect inducing VE. Each VE consisted of the same 3D spatial structure and 14 VE-specific objects/events. The VEs and the event stimuli were validated using with an online survey, where the VE and the individual object/events for each VE were rated by 67 participants on their perceived valence and arousal levels. Participants were also asked which objects/events the memorised for each VE. These memory accuracy scores as a second level of validation, since affective stimuli tend to be more memorable than neutral, see section 2.3.3.2. The ratings from the participants showed that the videos of the VE-scenario conditions induced the expected mean arousal and valence ratings, which were significantly different from each other. More specifically, the neutral scenario elicited neutral valence and low arousal, while the positive and negative scenario elicited highly positive and highly negative valence respectively, and above-average arousal. The negative VE was the most arousing condition, as expected due to the 'negativity bias', discussed earlier in section 1.

The valence and arousal ratings for the specific objects and events within each VE allowed us to inspect the effect of the environmental properties in more detail, and to find and exclude those objects and events that did not induce the desired affect. Interestingly, the stimuli which were rated as more arousing and highly valenced, were the VE-specific interactive events compared to VE-specific static objects and elements that were present in all three VEs. The interactive events were also the most memorable, as confirmed by the memory accuracies scores. Significant correlation was found between memory accuracies, valence and arousal scores, in line with past literature on the link between memory recall and affect (see section 2.3.3.2).

Presence ratings were assessed in the online survey and in the VR study because the **part of the forth aim (4c) was to assess relationship between presence and affective intensities**. As expected, presence ratings of the VEs in the online

survey were low, due to the non-immersive medium used for their presentation. Still, the negative VE provoked higher presence ratings than the other scenarios. The finding also suggest a strong relationship between presence and arousal ratings for both positive and negative VEs, which is in line with the previous findings suggesting that presence increases in physiologically active and arousing experiences or tasks suggesting that presence increases in physiologically active and arousing experiences or tasks (see section 2.3.3). This finding also strengthens the existing work on the link between emotion elicitation and presence (see section 2.3.3). It also suggests that future work on immersive VR may be able to manipulate presence levels in VR by incorporating affect inducing elements into the VR simulation.

In addition, subjective difference such as alexithymia levels were expected to have an impact on self-ratings and physiological readings (see section 2.3.3.2). High alexithymia causes difficulty in recognising and regulating one's emotions which was expected to influence the affect self-ratings. We used the TAS-20 questionnaire to assess alexithymia levels in our study. Interestingly, alexithymia levels were not found to have any effect on the valence and arousal self-ratings. Hence, further investigation on the effect of alexithymia on affect ratings and physiological responses was not followed as part of this thesis.

In sum, the online survey validated the stimuli and the VEs and three ways: firstly, by confirming the mean VE ratings, secondly, by acquiring self-rating per stimulus/event, and thirdly, by recording corresponding memory accuracy scores. Our next step was to test the affect induction approach in immersive settings using HMD-enabled VR.

7.1.5. Achieving automatic affect detection in seated and room-scale VR.

Finally,Aim 4.d. of this thesis was to investigate the feasibility of applying automatic affect detection in immersive VR. To achieve this goal, a large VR study was conducted at the Science Museum in London. The objectives for this study were to validate the use of immersive VR scenes as a reliable affect induction tool, explore the effects of affect, presence and interactivity on physiological readings, and confirm the feasibility of automatic affect detection in seated and free-walking VR settings. The pre-validated VE and stimulus materials from the online survey

were used in this VR study. Two interaction modes for the VR experiences were explored: the interactive first-person participation in room-scale VR (active participant group), and the passive observation of virtual stimuli from a third-person perspective in VR (passive participant group). Continuous physiological readings from the sensor set-up, along with subjective continuous self-ratings (via CASR) and post-VE ratings were recorded from 291 participants.

The effect of interactivity on subjective ratings. The self-ratings revalidated the reliability of the VEs for affect induction, agreeing with what was previously reported in the online survey, making interactive virtual scenes a powerful affect induction tool. As expected, the ability of room-scale, freely moving (compared to vicarious passive viewing) amplified the affective responses. This was shown through differences in arousal self-ratings, between the two groups, but also through memory accuracy scores and presence, with the active group achieving higher scores than the passive group (see section 5.3.2). This finding was not directly due to the level of immersion imposed by the medium, as both groups experienced the stimuli using the same HMD technologies. The difference was perhaps rather in the overall visual perception of the 3D world, the interactivity (head-view coordination) and high fidelity in the simulated world which allowed for a more naturalistic exploration in active condition [509]. Perceptual biases in perceiving 3-D shapes and spaces can be eliminated in moving observes, (compared to static observes), which in turn increases ecological validity of the simulated environment. This in turn can permit an overall more naturalistic interaction with the content, which can reflect how we reach with the world (see section 2.3.3).

The effect of interactivity was not the only factor inducing stronger affective responses and stronger presence. The self-ratings also showed that the intensity of the affective content itself was found to be linked to higher levels of presence. In our study, the positive and negative VEs were also the most arousal and presence inducing across all users, agreeing with the suggested relationship between affect and presence (section 2.3.3). Again, in the VE examples we explored, the neutral VE contained non-interactive, neutral objects compared to the affective VEs. It is possible that interactive events add another dimension in affective elicitation, in a similar way that moving pictures do compared to static images (as described in sections 2.4.1 and 4.1). The results suggest that interactive affective experiences hold a better potential in eliciting higher intensity responses and a higher sense of 'being-there' in VR.

Continuous ratings can reveal finer-grained and event-related differences. We developed a continuous CASR rating scale to assess subjective valence and arousal changes throughout the users' interaction in the VEs. The CASR scale was easy to use. As expected, CASR ratings showed a higher sensitivity to affective changes compared to post-VE ratings. When averaged per VE scene, the mean CASR ratings showed similar arousal levels between the positive and negative scene for the active group, showing no effect of negativity bias on the ratings.

More importantly, we were also able to link affective ratings to individual events which were visible to the user at certain points during the experience (gaze-based detection, explained in section 5.2.3), because of the nature of the CASR ratings. For example, the event-specific analysis showed that the passive interaction mode affected the experience of positive stimuli. The passive group's mean VE ratings were more negative compared to the active group ratings. This was because several positive stimuli reported as slightly negative and low arousing in the passive environment, despite being rated as positive in the online survey. Low arousing negative ratings are usually attributed to tiredness and boredom (see dimensional model in Figure 1). It is hard to pinpoint the reason why those 'positive' events were rated so differently than the others but potential explanations could be (1) inhibition of approach [510] generated by the passiveness of the condition, and (2) cognitive inhibition in stimulus recognition [511] due to perceptual biases and method of visualisation. Passively observing a virtual experience through the eyes of someone else may also pertain to subjective traits (e.g. empathy levels) and other media-related ramifications that were not controlled for [512], [513]. These media type effects should be further and systematically explored in future studies.

Overall, the virtual objects and events within those VEs evoked mean valence-arousal combinations, which when graphed into the 2-D affective space followed a V-shape relation. This was most prominent for the ratings of the active group. The V-shape relation, or as often also called 'inverse U-shaped curve' [506], is considerably evident on the ratings of the IAPS picture library [405] which shows that arousal ratings increase with the intensity of valence towards higher pleasure (positive) or displeasure (negative). Therefore, arousal may be expressed as a function of valence and they are not truly independent; although this relationship can be affected by the individual parameters, and the context in which they were reported (in this case e.g., interactive versus passive environments).

Validity of automatic detection of voluntary facial expressions recognition. As part of the VR study, our team explored the sensitivity of the f-EMG sensors in detecting posed, voluntary expressions in a facial emotion mimicry exercise. Data from the voluntary expressions, of smile (as 'being happy'), frown (as 'being angry') and brow raise (as 'being surprised'), were analysed per channel. The expressions were carefully chosen to induce intense activations on the facial muscles underlying the f-EMG sensors (zygomaticus major, corrugator, orbicularis oculi and frontalis) (see section 6.3.2). From those expressions, we computed the maximum muscle contractions were which were used for the normalisation of the spontaneous EMG responses. The voluntary expressions analysis showed distinct EMG channel activations patterns for each expression. For example, positive and negative expressions were distinguishable, as expressed by the activation of zygomaticus major and corrugator (which agreed with previous research, discussed in section 2.3.1).

More importantly, as expected, the expressions were distinguishable from all users using a generalised classification model yielding 80% classification accuracy across participants. The performance scores are comparable to expression classification work by Hamedi et al., [514] using EMG RMS feature with SVM classifier (85.5% accuracy). Direct comparison with other studies is difficult, since the number of sensors, the facial expressions and features vary between different studies [515]. Our work was further improved in a different study in 2019 [516] where we achieved 86% accuracy in detecting five expression using the same sensor set-up. These findings confirmed the feasibility of detecting voluntary expressions with this setup with high accuracy.

Validity of spontaneous affect detection in VR settings. In the final study participants of the active and the passive group experienced the three VEs while wearing the sensor set-up. We wanted to assess the overall feasibility of inferring spontaneous affective changes from physiological readings within VR, which was the last research aim of this thesis. Therefore, physiological data recorded from EMG and the PPG sensors were analysed to validate the reliable measurement of affective states in different VEs, for both the passive and the active group. An event-based analysis of physiological metrics showed changes between the VE conditions, which were more prevalent for the active group. Despite the increased noise accounted by the movement of the participants in the active scenario, EMG measures showed strong detection potential for the negative, positive, and neutral conditions which can

be attributed to valence changes (explained in section 2.3.1) in spontaneous, naturalistic settings.

By comparison, the HR measures dissociated between the affective and neutral conditions for both groups, showing closer link to arousal rating differences (see section 2.3.2.1). This differentiation was stronger for the negative versus the neutral VE, which was expected as the negative VE was rated as overall more arousing than the other conditions (post-VE ratings, see Figure 64). The PRV measures showed good discriminatory power for arousing against neutral condition in the passive group but did not provide a similar effect in the active group. One reason for this could be the method of data feature extraction. More specifically, the extraction of those features was made using short time windows which have been suggested to be sufficient for time-domain HRV calculation metrics [517]–[521]. However, these metrics could have been heavily affected by movement noise in the active scenario, rendering the PVR metric calculation erroneous (further explained in section 2.3.2.1 and 4.4). Therefore, the calculation of PRV metrics in active (walking) settings should be further improved in future studies, potentially by applying an averaging approach between multiple short time windows as in [517]. Nevertheless, our result show that the features of IBI, rBPM (and potentially EMG RMS) could be used instead to detect changes in affect intensity in future which is an encouraging first step in the right direction.

Automatic affect detection can be attained in seated and free-walking conditions using classification methods. The feasibility of employing classification methods for the detection of valence and arousal changes was explored. Such methods allow the mapping of physiological responses to self-ratings of valence and arousal, via the identification of patterns in the data. In our classification experiments, continuous EMG measures from both groups were used to classify two and three classes of valence and arousal using classifiers commonly used in physiological data processing. For the training and validation of the classifiers for this study, we devised a user-independent approach as it is commonly used for the creation of generalisable models (see section 3.6.4), applying a mixed-subjects cross-validation (CV) and a separated-subjects cross validation (CV) (the later has been supported that can generalise better to new incoming data although the accuracies may be reduced [10], [323]). Additionally, we devised a user dependent approach, by developing a personalised model per user.

Our findings for such user-independent models showed that the best classifier can discriminate between two and three levels of valence and arousal with 77-94% accuracy when using mixed-subjects CV. The accuracies however drop to 55-65% when using separated-subjects CV. This demonstrates the intricacies of detecting naturalistic affective states in free-walking VR conditions (similar to [323]). Additionally, we computed the out-of-sample classification for each user separately (user-dependent approach) (section 6.3.4.3). The temporal sequence of the data was preserved when fed to the classifiers to account for time-based changes on the data (e.g., impedance, carry-over effects etc.) as well as preserve the causality of the changes. The best classifiers reached a mean prediction accuracy between participants of 69-80% for 2-levels of valence and arousal, showcasing once again the robustness of subject-specific models, which lack the ability however to generalise for new users (section 3.6.4, 6.2.6).

Between the three classifiers which were chosen based on previous research (see review by [10]), we observed that the KNN performed higher for the user-independent across mixed-users CV, NB performed higher for separated-subject CV user-independent approach (for discrimination of valence) and KNN performed significantly higher than other two classifiers in the user-dependent prediction approach. The results from the classification of valence and arousal demonstrated the feasibility of automatic affect detection using wearable technologies in free-walking immersive VR settings which to our best knowledge, has not been explored before.

Affect detection from naturalistic, non-stationary physiological signals pose many challenges, however understanding their nature could open new avenues for deployment in real-world affective computing applications. The performance of such automatic affect detection systems could be improved with the use of ensemble [522] and deep learning (DL) methods. Increasing number of research on affective computing promote the use of deep artificial neural network architectures for valence and arousal recognition from image[523], speech[524] and physiological signals[525]–[527] in controlled conditions outside VR. These methods overcome the difficulties of manual ad-hoc feature extraction and have the ability to reduce signal resolution across different layers, potentially saving computational resource [528]. Existing datasets with physiological data recorded in controlled conditions (e.g., DEAP[365], AMIGOS[529]) were classified with high accuracies (>75%) using convolutional and deep neutral networks [530], [531]. However, such methods perform lower (~65%) in data recorded in non-stationary settings [532]. Recently,

DL approaches have started to emerge for stress and fear detection systems using immersive VR technologies [533], [534]. The future investigation of successful employing DL approaches for real-time affect detection in VR could allow for affect-responsive applications and enhance the capabilities of VR based therapies and training applications.

Future work on affective computing in VR could be enriched by exploring the effect of (inter)active experiences, outside controlled seated conditions. By accounting for the factors contributing to the feeling of presence and embodiment, affective responses in VR could reliably reflect real-world, naturalistic responses. As sensor methods are miniaturising, becoming less obtrusive and even wireless, the physical movement constraints will be reduced, allowing for more realistic user interaction with content. Overall, the active condition in our study induced stronger affective responses visible on self-ratings, and in spontaneous and voluntary facial muscle activations. The larger changes between conditions noted in the active group, could be the effect of realistic affective responses as a result of place illusion (presence) and plausibility illusion and the factors pertaining it [535]. Generally, the findings show that there are inherent differences between the two modes of virtual stimuli presentation that can influence the participants responses drastically.

Overall, these results agree with previous research using VR stimuli against video clips, and video-clips against static images [294], as VR can indeed be used as a reliable affect induction tool. Our findings suggest that moving experiences, as in room-scale VR, and the interactive element of features inside the simulated world, can not only induce negative and positive affect but show stronger arousal responses to stimuli overall compared to non-interactive passive experiences. Additionally, sensor data recorded from the prototype set-up can be used to dissociate affective conditions, and ultimately detect valence and arousal.

7.2. **Discussion**

The detection of involuntary, naturalistically elicited emotional states is a challenging task which is still in early stages [8], [68]. Most of the studies in the past focused on the detection of affect in controlled laboratory conditions, which was a necessary step for the investigation of suitable methods for automatic affect detection. However, it is argued that such constrained elicitation may not generalise to real-world conditions [323], [536], [537]. The detection of spontaneous elicited affect in naturalistic conditions is superior to voluntary [538] (e.g., posing or

imitating an actor expressing an emotion) and controlled laboratory induced affect [323], [539], as they can represent more realistically the effects of actual emotions on our behaviour and physiology in real-life.

In our work, we employed physiological methods for the detection of two dimensions of affect (EMG on facial muscles for valence, and heart-rate sensors for arousal) on a novel wearable insert for virtual reality (VR) that had not been tested before. Physiological changes are less sensitive to induced affective states and the influence of social masking [322], and the methods generally utilised to detect them are of high temporal resolution which can show subtle changes that are not visible by eye otherwise [128], [537], [540]. We envisioned that this novel interface could continuously record physiological responses, while becoming 'invisible' to the senses of the wearer who is being immersed in VR. Today, as sensors are becoming increasingly smaller, we are witnessing of large expansion into wearable technologies which are incorporating physiological sensors [4]. These sensors are progressively more cost-effective and less intrusive to the wearer, making them perfect candidates for every-day interventions [541], potentially applied in areas as education, entertainment, well-being and healthcare treatments, social and computer interactions, adaptive control and security. Although the hardware technologies are evolving, we have still to enlarge our understanding on the intricate factors (e.g. [6], [542]) and the non-emotional influences (e.g. coughing, walking) [328], [543] (or lack thereof) underlying physiological activation during data collection. One of the main challenges is inter-subject and mixed-subjects differences on affect elicitation and detection [75], [322], [323]. Several external and internal factors could play a role. External factors could include the environment, the task performed, the technology involved, the stimuli presentation interface, and the degrees of freedom of the setup (standing, moving, seated, lying). Examples for internal factors could include personality, gender, cultural background, alexithymia levels, and generally the individual differences in the quality, intensity, and latency of affect expression as a response to the same stimulus. As such, future work will try to address the effect of individual differences on naturalistic affect elicitation and experiences in virtual reality.

The development of automatic affect detection models for application in real-world scenarios, will be greatly benefited by more systematic investigation of suitable and best performing classification methods tailored towards the detection of affective states from physiological signals in naturalistic conditions [10]. For example, it is difficult to determine which is the best classifier for the different types

of physiological signals for each data collection condition and degrees of freedom [323]. Additionally, there is a big discrepancy between the data collection conditions (e.g., contextual information), protocols and methodologies (e.g., classification models) used between studies, a lack in available public databases for naturalistic affect elicitation and only a small body of research on affect elicitation in VR using different types of stimuli. These together with the fact that technologies change rapidly (even within a year), makes the replication of such studies complicated. Effort should be put for future classification experiments on attempting developing classification models that can apply in the real-world scenarios, either for user-dependent models or for user-independent models, emphasising holistically on the factors that could influence emotion elicitation and how those models developed could be used for the classification of new data from the same or many new users.

Past research on spontaneous affect detection within and outside VR revolved around the detection of two or three level of valence and arousal from physiological signals (similar to our studies) by utilising post-experience self-ratings and ratings from external observers, in controlled laboratory conditions (e.g. [294], [333], [538], [544]–[546]). Our work attempted to expand the scope of past research by (a) inducing spontaneous, stimuli and context induced, affective states in naturalistic conditions, (b) not limiting movement involving walking in space in VR, (c) bringing the research study to the participants (Science Museum) instead of taking them to the laboratory, (d) utilising the immersive abilities of VR to induce a state of Presence, (e) testing the feasibility of novel sensor equipment for head-mounted tracking using EMG, and (f) allowing for triple labelling of the data (self-rating throughout the experience, retrospective (post-VE) scores per user, experiment-specific labels via validation from online surveys). Finally, as part of this thesis we also classified physiological responses in distinct levels of valence and arousal (two and three) from 1) seated laboratory-based affect induction from videos, 4) passive seated VR-based vicariously induced affect, and we took one step further, attempting affect detection in 3) actively-walking VR-based naturally induced affect. Our findings showed overall the superiority of interactive immersive experiences for spontaneous, naturalistic affect elicitation in VR, and its effects on presence. Enabling reliable automatic affect detection in such conditions forecast a promising future in applying such affective systems into real-world applications.

Such affect-enabled sensing VR technologies could be applied in three main contexts, in research, healthcare and commercials applications. VR can enable researchers and developers to directly design the experience of the wearer by

providing direct control over what is seen, heard, and interacted with. Researchers could study the effects of different parameters (e.g., changes in colours or shapes) or scenarios VR, which could enhance our understanding of the emotional impact of those parameters on different users and facilitate the development of better practices. Immersive VR can also provide high ecological validity which can in turn induce naturalistic responses in users. For this reason, VR has been successfully adapted in clinical research settings and for mental-health therapies and meditation interventions. In those settings, affect detection could provide an additional layer of information on the emotional state of a patient, which for example could inform and assist therapists and mental care practitioners delivering VRETs[547]. These sensing metrics could also be used as adaptive feedback to the simulation, thus providing more personalised content to the needs of the user. As such, training and educations applications could also benefit from such methodologies. In the post-COVID-19 era, where new practices and regulations are developed [548], we expect VR to play a bigger role in delivering those therapies and personalised content to the user [2], [549]. Similarly, the quantification of affective changes could be applied in commercial applications including market research and entertainment.

7.3. Limitations & Challenges for Future Work

Reliability of self-ratings. Mapping the physiological responses to the adequate labels or annotations is typically challenging. In our studies, we used the CASR self-ratings throughout the virtual experiences for the labelling of the data. Generally, self-reported labels may not occur simultaneously with the physiological responses, due to delays and/or errors in reporting which may not be easy to locate nor fix ([228], [477]). Although we tried to compensate for simple reaction-time delays by adjusting the CASR and physiological data synchronisation during processing, the exact and precise localisation of the labels could be subjected to individual differences (e.g. cognitive load, attention deficit disorders, mental fatigue [550]–[552]). Perhaps we could next complement the self-ratings with additional judgements from external observers and by investigating additional self-rating processing models such as in [325]. Without doubt, continuous self-rating without pre-training could also increase the cognitive load to the users [334] and distract them from the main task with potential impact on the feeling of presence. Future research could investigate the effect of simultaneous rating on the feeling of

presence, and the effect of VR on reaction times. Next, in our work we plan to compare the presence levels of users who were self-rating their affect against users who only rated retrospectively, from multiple shorter virtual stimuli experiences. Note that recording of the self-ratings is important for the development of the automatic affect detection systems. Once the system is successfully deployed, such ratings will not be necessary.

Generalisation to open environments. In our VR study, we used context-based stimuli in indoor virtual environments (office replicas). It is unclear whether the results could generalise to different environments or different stimuli e.g., using avatars. As such further research could focus on incorporating additional VR corpora that can simulate various situations and environments (e.g., outdoors, involving low-level architectural elements, additional sound, abstract synthesis, realistic social environments). Additionally, in our study the event-stimuli were activated by the user's gaze. Our event-based algorithm could be greatly benefited from the addition of accurate eye-tracking technologies which are currently emerging in the market ([553]–[556]).

To allow the users to move naturally and explore the virtual spaces, we wanted to reduce the number of factors contributing towards their distraction and discomfort. This led to the development of a wireless version of the main apparatus used for physiological signal detection, which was found non-intrusive nor hindering movement. However, the HMD VR headset with motion tracking we used, was heavy and required connection via cable to the computer. VR technologies are now launching wireless, smaller headsets with motion tracking capabilities which could greatly enhance the overall user experience. Additionally, the insertion of supplementary sensorimotor contingencies, e.g., a virtual body (avatar) in the place where the user's physical one should be in VR, could contribute to the feeling of embodiment and the subjective feeling of presence [213], [557], which could augment naturalistic affect elicitation.

Limitations of recordings and analysis approaches. The feasibility studies showed an overall promising result for valence and arousal detection. We used only a subset of features which are commonly utilised in existing related literature, in order to control for complexity added from feature multidimensionality [323]. The careful addition of extra features and additional modalities could potentially improve the accuracy performance, and as such it will be explored in our future work. Further

solutions to increase robustness and accuracy of our models will be investigated, such as the implementation of other classification models, multimodal fusion techniques, deep learning and ensemble combination strategies [558], on spontaneously elicited responses in naturalistic conditions.

Another challenge in this area of research can be obtaining reliable annotation of affective states [334] and data filtering from long-term recordings in VR. Movement and non-emotional influences can cause artefacts issues and data-missing occurrences in the data streams, which can naturally be increased with the addition of multimodal sensors. Future research should not only systematically study the effectiveness of classifications methods on affect detection in various settings, but also tackle common data-collection issues such as corrupted and missing data.

We have observed that detection accuracies on spontaneously elicited affective responses decrease, especially when using a separated-users validation approach. Natural expressions are involving subtle changes and are highly context-sensitive and subject-dependent. VR as a method of induction is very recent, and thus could have additional, unexplored effects on the experience of distinct and complex affective states. For that reason, we anticipate a substantial volume of future research will be dedicated for this endeavour.

Summary of all key findings

In this EngD thesis, we have:

- Developed and validated a novel sensor set-up specifically designed for head-mounted VR settings using physiological sensors. This set-up was able to reliably detect valence and arousal responses in seated and active/walking conditions using affective videos.

- Designed a system architecture for automatic affect detection in VR (utilising the dimensional model of affect).

- Created and validated affective VR 3D scenarios and stimuli, which were able to successfully induce variations of valence and arousal levels.

- Conducted on of the largest VR data set to our knowledge in fully up-to-date immersive settings with free walking capabilities outside laboratory settings whilst recording affective states. Physiological responses from this study confirmed the feasibility of automatic affect detection in fully immersive VR settings. VR can indeed be used as a reliable affect induction and detection tool.

Future directions:

- PPG and EMG sensor placement improvements

- Investigate effects of the intricate factors and non-emotional influences.

- Employ different VR environments to extend the range of settings confirming reliable affect detection.

- Investigation of suitable and best performing classification methods.

- Explore additional features, modalities and user newer, lighter VR-HMD headsets

References

[1] T. D. Parsons and A. A. Rizzo, "Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: A meta-analysis," *J. Behav. Ther. Exp. Psychiatry*, 2008, doi: 10.1016/j.jbtep.2007.07.007.

[2] M. Smits, J. Bart Staal, and H. Van Goor, "Could Virtual Reality play a role in the rehabilitation after COVID-19 infection?," *BMJ Open Sport and Exercise Medicine*, vol. 6, no. 1. BMJ Publishing Group, p. e000943, Oct. 23, 2020, doi: 10.1136/bmjsem-2020-000943.

[3] R. W. Picard, "Affective Computing," *MIT Press*, no. 321, pp. 1–16, 1995, doi: 10.1007/BF01238028.

[4] I. Mavridou, M. Perry, E. Seiss, T. Kostoulas, and E. Balaguer-Ballester, "Emerging Affect Detection Methodologies in VR and future directions," in *Virtual Reality International Conference VRIC 2019*, 2019, p. 35, [Online]. Available: https://www.laval-virtual.com/wp-content/themes/lavalvirtual/assets/files/Proceedings_VRIC_Laval_Virtual_2019_ConVRgence_bd.pdf#page=34.

[5] S. Cosentino, S. Sessa, and A. Takanishi, "Quantitative Laughter Detection, Measurement, and Classification - A Critical Survey," *IEEE Rev. Biomed. Eng.*, vol. 9, pp. 148–162, 2016, doi: 10.1109/RBME.2016.2527638.

[6] S. H. Fairclough, "Fundamentals of physiological computing," *Interact. Comput.*, vol. 21, no. 1–2, pp. 133–145, 2009, doi: 10.1016/j.intcom.2008.10.011.

[7] A. M. Turing, "Computing machinery and intelligence," in *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, vol. 59, no. 236, 2009, pp. 23–65.

[8] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, 2010, doi: 10.1109/T-AFFC.2010.1.

[9] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, 2017, doi: 10.1016/j.inffus.2017.02.003.

[10] H. Al Osman and T. H. Falk, "Multimodal Affect Recognition: Current Approaches and Challenges," in *Emotion and Attention Recognition Based on Biological Signals and Images*, InTech, 2017.

[11] S. Triberti, C. Repetto, and G. Riva, "Psychological factors influencing the

effectiveness of virtual reality-based analgesia: A systematic review," *Cyberpsychology, Behavior, and Social Networking*, vol. 17, no. 6. Mary Ann Liebert Inc., pp. 335–345, Jun. 01, 2014, doi: 10.1089/cyber.2014.0054.

[12] P. Ekman and H. Oster, "Facial Expressions of Emotion," *Annu. Rev. Psychol.*, vol. 30, no. 1, pp. 527–554, Jan. 1979, doi: 10.1146/annurev.ps.30.020179.002523.

[13] U. Dimberg, M. Thunberg, and K. Elmehed, "Unconscious Facial Reactions to Emotional Facial Expressions," *Psychol. Sci.*, vol. 11, no. 1, pp. 86–89, Jan. 2000, doi: 10.1111/1467-9280.00221.

[14] P. Ekman, "Basic Emotions," in *The Handbook of Cognition and Emotion*, T. D. and T. Power, Ed. 1999, pp. 45–60.

[15] C. E. Izard, "Four systems for emotion activation: Cognitive and noncognitive processes.," *Psychol. Rev.*, vol. 100, no. 1, pp. 68–90, 1993, doi: 10.1037/0033-295X.100.1.68.

[16] R. Adolphs, "Recognizing Emotion from Facial Expressions: Psychological and Neurological Mechanisms," *Behav. Cogn. Neurosci. Rev.*, vol. 1, no. 1, pp. 21–62, 2002, doi: 10.1177/1534582302001001003.

[17] S. B. Hamann and S. H. Kim, "Neural correlates of positive and negative emotion regulation.," *J. Cogn. Neurosci.*, 2007, doi: 10.1162/jocn.2007.19.5.776.

[18] F. Dolcos, A. D. Iordan, and S. Dolcos, "Neural correlates of emotion - Cognition interactions: A review of evidence from brain imaging investigations," *J. Cogn. Psychol.*, 2011.

[19] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements," *Psychol. Sci. Public Interes.*, vol. 20, no. 1, pp. 1–68, Jul. 2019, doi: 10.1177/1529100619832930.

[20] S. Richir, K. Tcha-Tokey, E. Loup-Escande, and O. Christmann, "A Questionnaire to Measure the User Experience in Immersive Virtual Environments Ecole Nationale Supérieure d'Arts et Métiers A Questionnaire to Measure the User Experience in Immersive Virtual Environments," doi: 10.1145/2927929.2927955.

[21] A. Vermeeren, E. Law, V. Roto, … M. O.-P. of the 6th, and undefined 2010, "User experience evaluation methods: current state and development needs," *dl.acm.org*, Accessed: Sep. 16, 2019. [Online]. Available: https://dl.acm.org/citation.cfm?id=1868973.

[22] K. Gurusamy, R. Aggarwal, … L. P.-B. J. of, and undefined 2008, "Systematic review of randomized controlled trials on the effectiveness of virtual reality training for laparoscopic surgery," *Wiley Online Libr.*, Accessed: Sep. 03, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/bjs.6344.

[23] A. Alaraj, M. Lemole, … J. F.-S. neurology, and undefined 2011, "Virtual reality training in neurosurgery: review of current status and future applications,"

ncbi.nlm.nih.gov, Accessed: Sep. 03, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3114314/.

[24]   A. Rizzo, F. D. Rose, B. M. Brooks, and A. A. Rizzo, "STRIVE: Stress Resilience in Virtual Environments View project Virtual Reality and Brain Injury View project Virtual Reality in Brain Damage Rehabilitation: Review," *CYBERPSYCHOLOGY Behav.*, vol. 8, 2005, doi: 10.1089/cpb.2005.8.241.

[25]   S. V Adamovich, G. G. Fluet, E. Tunik, and A. S. Merians, "Sensorimotor training in virtual reality: A review," *NeuroRehabilitation*, vol. 25, pp. 29–44, 2009, doi: 10.3233/NRE-2009-0497.

[26]   S. R. Harris, R. L. Kemmerling, and M. M. North, "Brief Virtual Reality Therapy for Public Speaking Anxiety," *CyberPsychology Behav.*, 2002, doi: 10.1089/109493102321018187.

[27]   A. Gorini and G. Riva, "Virtual reality in anxiety disorders: the past and the future," *Expert Rev. Neurother.*, vol. 8, no. 2, pp. 215–233, Feb. 2008, doi: 10.1586/14737175.8.2.215.

[28]   "• Global VR headset sales by brand 2016-2017 | Statista." https://www.statista.com/statistics/752110/global-vr-headset-sales-by-brand/ (accessed Sep. 05, 2019).

[29]   "Infographic: What is Extended Reality (XR)? - Visual Capitalist." https://www.visualcapitalist.com/extended-reality-xr/ (accessed Sep. 10, 2019).

[30]   "Emotion Analytics Market to Rake in Around US $25 Billion by 2023 with 17% CAGR - Reuters." https://www.reuters.com/brandfeatures/venture-capital/article?id=32690 (accessed Sep. 03, 2019).

[31]   R. Mandryk, "Physiological Measures for Game Evaluation," in *Game Usability: Advice from the Experts for Advancing the Player Experience*, 2008.

[32]   R. L. Mandryk and M. S. Atkins, "A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies," *Int. J. Hum. Comput. Stud.*, vol. 65, no. 4, pp. 329–347, 2007, doi: 10.1016/j.ijhcs.2006.11.011.

[33]   R. Cowie *et al.*, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, 2001, doi: 10.1109/79.911197.

[34]   K. R. Scherer, "Neuroscience projections to current debates in emotion psychology," *Cogn. Emot.*, vol. 7, no. 1, pp. 1–41, Jan. 1993, doi: 10.1080/02699939308409174.

[35]   B. L. Fredrickson, "Positive emotions.," in *Handbook of positive psychology*, C. R. Snyder & S. J. Lopez, Ed. New York: Oxford University Press, 2002, pp. 120–134.

[36]   J. E. Ledoux, "Cognitive-Emotional Interactions in the Brain," *Cogn. Emot.*, vol. 3, no. 4, pp. 267–289, Oct. 1989, doi: 10.1080/02699938908412709.

[37]   E. Hudlicka, "Guidelines for Designing Computational Models of Emotions," *Int. J. Synth. Emot.*, vol. 2, no. 1, pp. 26–79, 2011, doi: 10.4018/jse.2011010103.

[38]   L. Feldman Barrett *et al.*, "Emotions Are Real," vol. 12, no. 3, pp. 413–429, 2011,

doi: 10.1037/a0027555.

[39] W. J.-P. review and undefined 1894, "Discussion: The physical basis of emotion.," *psycnet.apa.org*, Accessed: Sep. 05, 2019. [Online]. Available: https://psycnet.apa.org/record/2006-01676-004.

[40] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biol. Psychol.*, vol. 84, pp. 14–41, 2010, doi: 10.1016/j.biopsycho.2010.03.010.

[41] S. Tomkins, *Affect imagery consciousness: Volume I: The positive affects*. 1962.

[42] C. E. Izard, *The face of emotion*. East Norwalk, CT: Appleton-Century-Crofts, 1971.

[43] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.," *J. Pers. Soc. Psychol.*, 1971, doi: 10.1037/h0030377.

[44] P. Ekman, "An Argument for Basic Emotions," *Cogn. Emot.*, vol. 6, no. 3–4, pp. 169–200, May 1992, doi: 10.1080/02699939208411068.

[45] J. Panksepp, "Criteria for basic emotions: Is DISGUST a primary '"emotion"'?," doi: 10.1080/02699930701334302.

[46] R. Plutchik and H. Kellerman, *Theories of emotion*. .

[47] R. E. Jack, O. G. B. Garrod, and P. G. Schyns, "Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time," *Curr. Biol.*, vol. 24, no. 2, pp. 187–192, Jan. 2014, doi: 10.1016/j.cub.2013.11.064.

[48] P. L. B.-E. Ekman, "Facial expression and emotion," *Am. Psychol.*, 1993, doi: 10.1037/0003-066X.48.4.384.

[49] L. F. Barrett *et al.*, "Of Mice and Men: Natural Kinds of Emotions in the Mammalian Brain? A Response to Panksepp and Izard," *Perspect. Psychol. Sci.*, vol. 2, no. 3, pp. 297–312, 2007, doi: 10.1111/j.1745-6916.2007.00046.x.

[50] M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan, "Automatic Recognition of Facial Actions in Spontaneous Expressions," 2006.

[51] H. Gunes, M. P.-I. J. of S. Emotions, and undefined 2010, "Automatic, Dimensional and Continuous Emotion Recognition," *dl.acm.org*, Accessed: Apr. 12, 2020. [Online]. Available: https://dl.acm.org/doi/abs/10.4018/jse.2010101605.

[52] M. G. Calvo and L. Nummenmaa, "Perceptual and affective mechanisms in facial expression recognition: An integrative review," *Cogn. Emot.*, vol. 30, no. 6, pp. 1081–1106, 2016, doi: 10.1080/02699931.2015.1049124.

[53] G. Recio, A. Schacht, W. S.-B. psychology, and undefined 2014, "Recognizing dynamic facial expressions of emotion: Specificity and intensity effects in event-related brain potentials," *Elsevier*, Accessed: Apr. 12, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301051113002500.

[54] M. Pantic and I. Patras, "Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments From Face Profile Image Sequences," *CYBERNETICS*, vol. 36, no. 2, p. 433, 2006, doi: 10.1109/TSMCB.2005.859075.

[55]   M. Kumar, A. Veeraraghavan, and A. Sabharval, "DistancePPG: Robust non-contact vital signs monitoring using a camera," 2015. Accessed: Sep. 05, 2019. [Online]. Available: http://www.vitalsignscamera.com/.

[56]   J. Kranjec, S. Beguš, G. Geršak, and J. Drnovšek, "Non-contact heart rate and heart rate variability measurements: A review," *Biomedical Signal Processing and Control*, vol. 13, no. 1. Elsevier BV, pp. 102–112, 2014, doi: 10.1016/j.bspc.2014.03.004.

[57]   J. Durán and J. Fernández-Dols, "Do emotions result in their predicted facial expressions? A meta-analysis of studies on the link between expression and emotion," 2018, Accessed: Feb. 22, 2021. [Online]. Available: https://psyarxiv.com/65qp7/download?format=pdf.

[58]   J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980, doi: 10.1037/h0077714.

[59]   J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant.," *J. Pers. Soc. Psychol.*, vol. 76, no. 5, pp. 805–819, 1999, doi: 10.1037/0022-3514.76.5.805.

[60]   D. Watson, L. A. Clark, and A. Tellegen, "Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales," *J. Pers. Soc. Psychol.*, vol. 54, no. 6, pp. 1063–1070, 1988.

[61]   A. Tellegen and D. Watson, "On the dimensional and hierarchical structure of affect," *Psychol. Sci.*, 1999.

[62]   H. M. McCormack, D. J. de L. Horne, and S. Sheather, "Clinical applications of visual analogue scales: a critical review," *Psychol. Med.*, vol. 18, no. 04, p. 1007, Nov. 1988, doi: 10.1017/S0033291700009934.

[63]   R. Likert, "A Technique for the Measurements of Attitudes," Accessed: Aug. 09, 2017. [Online]. Available: https://s3.amazonaws.com/academia.edu.documents/31930406/A_Technique_for_t he_Measurements_of_Attitudes.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53 UL3A&Expires=1502297940&Signature=1EgJUCXpqXJxXdtAi9vy4v26qiU%3D &response-content-disposition=inline%3B filename%3DA_Technique_for_the_Measurements_of_Atti.pdf.

[64]   M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Ther. Exp. Psychiatry*, 1994, doi: 10.1016/0005-7916(94)90063-9.

[65]   A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," *Behav. Res. Methods*, vol. 45, no. 4, pp. 1191–1207, Dec. 2013, doi: 10.3758/s13428-012-0314-x.

[66]   N. A. Remington, L. R. Fabrigar, and P. S. Visser, "Reexamining the circumplex model of affect.," *J. Pers. Soc. Psychol.*, 2000, doi: 10.1037/0022-3514.79.2.286.

[67]   M. Mortillaro, B. Meuleman, and K. R. Scherer, "Advocating a Componential Appraisal Model to Guide Emotion Recognition," *Int. J. Synth. Emot.*, vol. 3, no. 1, pp. 18–32, Aug. 2012, doi: 10.4018/jse.2012010102.

[68]   Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2009, doi: 10.1109/TPAMI.2008.52.

[69]   M. Soleymani, S. Koelstra, I. Patras, and T. Pun, "Continuous Emotion Detection in Response to Music Videos." Accessed: Feb. 07, 2020. [Online]. Available: http://www.neurobs.com.

[70]   N. Frijda, "Emotions and action," in *Feelings and emotions. The Amsterdam Symposium*, 2004.

[71]   N. H. Frijda, "The emotions: Studies in emotion and social interaction," *Cambridge Univ. Press*, p. 560, 1986, doi: 10.1093/0199253048.001.0001.

[72]   R. S. Lazarus, "Progress on a cognitive motivational relational theory of emotion," *Am. Psychol.*, 1991, doi: Doi 10.1037//0003-066x.46.8.819.

[73]   R. L.-T. nature of emotion: F. questions and undefined 1994, "Human emotion: A functional view."

[74]   K. R. Scherer, "Appraisal theory.," *Handbook of cognition and emotion*. 1999, doi: 10.1002/0470013494.ch30.

[75]   H. Leventhal and K. Scherer, "The Relationship of Emotion to Cognition: A Functional Approach to a Semantic Controversy," *Cogn. Emot.*, vol. 1, no. 1, pp. 3–28, Mar. 1987, doi: 10.1080/02699938708408361.

[76]   N. H. Frijda, "Emotion, cognitive structure, and action tendency," *Cogn. Emot.*, vol. 1, no. 2, pp. 115–143, 1987, doi: 10.1080/02699938708408043.

[77]   N. H. Frijda, "What might emotions be? Comments on the Comments," *Soc. Sci. Inf.*, vol. 46, no. 3, pp. 433–443, Sep. 2007, doi: 10.1177/05390184070460030112.

[78]   R. Lazarus, S. F.-T. handbook of behavioral medicine, and undefined 1984, "Coping and adaptation," *New York Guilford*.

[79]   P. Ellsworth, K. S.-H. of affective sciences, and   undefined 2003, "Appraisal processes in emotion," *ict.usc.edu*, Accessed: Sep. 05, 2019. [Online]. Available: http://ict.usc.edu/~gratch/CSCI534/Readings/EllsworthScherer03.PDF.

[80]   A. Moors, P. C. Ellsworth, K. R. Scherer, and N. H. Frijda, "Appraisal Theories of Emotion: State of the Art and Future Development," *Emot. Rev.*, vol. 5, no. 2, pp. 119–124, Apr. 2013, doi: 10.1177/1754073912468165.

[81]   B. L. Fredrickson, "The role of positive emotions in positive psychology. The broaden-and-build theory of positive emotions.," *Am. Psychol.*, 2001, doi: 10.1037/0003-066X.56.3.218.

[82]   S. Schachter and J. Singer, "Cognitive, social, and physiological determinants of emotional state.," *Psychol. Rev.*, vol. 69, no. 5, pp. 379–399, 1962, doi:

10.1037/h0046234.

[83]   N. H. Frijda and B. Mesquita, "The social roles and functions of emotions," *Emot. Cult.*, pp. 51–87, 1994, doi: 10.1037/10152-002.

[84]   L. F. Barrett, "The Conceptual Act Theory: A Précis," *Emot. Rev.*, vol. 6, no. 4, pp. 292–297, Oct. 2014, doi: 10.1177/1754073914534479.

[85]   L. W. Barsalou, "Situated Conceptualization: Theory and Application," 2003.

[86]   Lisa Feldman Barrett, *How Emotions Are Made by Lisa Feldman Barrett*. 2018.

[87]   J. T. Cacioppo, G. G. Berntson, C. J. Norris, and J. K. Gollan, "The Evaluative Space Model," *Handb. Theor. Soc. Psychol. Vol. 1*, no. August 2014, pp. 50–73, 2012, doi: 10.4135/9781446249215.n4.

[88]   J. T. Cacioppo and G. G. Berntson, "Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates.," *Psychol. Bull.*, 1994, doi: 10.1037/0033-2909.115.3.401.

[89]   J. T. Larsen, A. P. McGraw, and J. T. Cacioppo, "Can people feel happy and sad at the same time?," *J. Pers. Soc. Psychol.*, 2001, doi: 10.1037//0022-3514.81.4.684.

[90]   M. Ptaszynski, R. Rzepka, and K. Araki, "On the Need of Context Processing in Affective Computing," *Proc. 26th Fuzzy Syst. Symp. ({FSS}2010), Organ. Sess. Emot.*, pp. 920–924, 2010.

[91]   I. B. Mauss and M. D. Robinson, "Measures of emotion: A review," *Cognition and Emotion*, vol. 23, no. 2. pp. 209–237, 2009, doi: 10.1080/02699930802204677.

[92]   P. Ekman, R. Levenson, W. F.- Science, and undefined 1983, "Autonomic nervous system activity distinguishes among emotions," *science.sciencemag.org*, Accessed: Aug. 29, 2019. [Online]. Available: https://science.sciencemag.org/content/221/4616/1208.short.

[93]   P. J. Lang, "The emotion probe: Studies of motivation and attention," *Am. Psychol.*, vol. 50, no. 5, pp. 372–385, 1995, doi: 10.1037/0003-066X.50.5.372.

[94]   K. Scherer, T. Johnstone, and G. Klasmeyer, "Vocal expression of emotion," *Handbook of Affective Sciences*. p. 1199, 2003.

[95]   K. Scherer, "What are emotions? And how can they be measured?," *Soc. Sci. Inf.*, 2005.

[96]   R. P.-H. of emotions and undefined 1993, "Emotions and their vicissitudes: Emotions and psychopathology."

[97]   P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto: Consulting Psychologists Press, 1978.

[98]   A. O. Cohen *et al.*, "The impact of emotional states on cognitive control circuitry and function," *J. Cogn. Neurosci.*, vol. 28, no. 3, pp. 446–459, Mar. 2016, doi: 10.1162/jocn_a_00906.

[99]   R. El Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Real-Time Vision for Human-Computer*

Interaction, Springer US, 2005, pp. 181–200.

[100] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognit.*, vol. 36, no. 1, pp. 259–275, Jan. 2003, doi: 10.1016/S0031-3203(02)00052-3.

[101] H. Aviezer, Y. Trope, and A. Todorov, "Body cues, not facial expressions, discriminate between intense positive and negative emotions," *Science (80-.).*, vol. 338, no. 6111, pp. 1225–1229, Nov. 2012, doi: 10.1126/science.1224313.

[102] A. P. Atkinson, W. H. Dittrich, A. J. Gemmell, and A. W. Young, "Emotion perception from dynamic and static body expressions in point-light and full-light displays," *Perception*, vol. 33, no. 6, pp. 717–746, 2004, doi: 10.1068/p5096.

[103] A. Atkinson, M. Tunstall, W. D.- Cognition, and undefined 2007, "Evidence for distinct contributions of form and motion information to the recognition of emotions from body gestures," *Elsevier*, Accessed: Sep. 10, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010027706001235.

[104] J. Montepare, E. Koff, D. Zaitchik, and M. Albert, "The use of body movements and gestures as cues to emotions in younger and older adults," *J. Nonverbal Behav.*, vol. 23, no. 2, pp. 133–152, Jun. 1999, doi: 10.1023/A:1021435526134.

[105] F. E. Pollick, H. M. Paterson, A. Bruderlin, and A. J. Sanford, "Perceiving affect from arm movement." Accessed: Aug. 29, 2019. [Online]. Available: www.elsevier.com/locate/cognit.

[106] H. G. Wallbott, "Bodily expression of emotion," *Eur. J. Soc. Psychol.*, 1998, doi: 10.1002/(SICI)1099-0992(1998110)28:6<879::AID-EJSP901>3.0.CO;2-W.

[107] N. Dael, M. Mortillaro, and K. R. Scherer, "The Body Action and Posture Coding System (BAP): Development and Reliability," *J. Nonverbal Behav.*, vol. 36, no. 2, pp. 97–121, Jun. 2012, doi: 10.1007/s10919-012-0130-0.

[108] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*. 2003, doi: 10.1016/S0167-6393(02)00084-5.

[109] K. S.-P. bulletin and undefined 1986, "Vocal affect expression: A review and a model for future research.," *psycnet.apa.org*, Accessed: Sep. 10, 2019. [Online]. Available: https://psycnet.apa.org/record/1986-16849-001.

[110] G. Duchenne, *Mécanisme de la physionomie humaine: où, Analyse électro-physiologique de l'expression des passions*. 1876.

[111] U. Hess and A. Fischer, "Emotional mimicry: Why and when we mimic emotions," *Soc. Personal. Psychol. Compass*, vol. 8, no. 2, pp. 45–57, 2014, doi: 10.1111/spc3.12083.

[112] P. Ekman and E. L. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. 2012.

[113] S. Namba, S. Makihara, R. S. Kabir, M. Miyatani, and T. Nakao, "Spontaneous Facial Expressions Are Different from Posed Facial Expressions: Morphological

Properties and Dynamic Sequences," *Curr. Psychol.*, vol. 36, no. 3, pp. 593–605, Sep. 2017, doi: 10.1007/s12144-016-9448-9.

[114] P. Ekman, W. V Friesen, and R. J. Davidson, "The Duchenne Smile: Emotional Expression and Brain Physiology II," 1990. Accessed: Oct. 24, 2020. [Online]. Available: https://psycnet.apa.org/journals/psp/58/2/342.html?uid=1990-14081-001.

[115] D. De Boulogne, *The Mechanism of Human Facial Expression*. 1990.

[116] U. Dimberg, M. Thunberg, and K. Elmehed, "Unconscious Facial Reactions to Emotional Facial Expressions," *Psychol. Sci.*, vol. 11, no. 1, pp. 86–89, 2000, doi: 10.1111/1467-9280.00221.

[117] K. Waters, "A Muscle Model for Animating Three-Dimensional Facial Expression," 1987.

[118] A. Van Boxtel, "Facial EMG as a tool for inferring affective states," *Proc. Meas. Behav.*, vol. 2010, no. August 24-27, pp. 104–108, 2010.

[119] S. Brown, G. S.-B. psychology, and undefined 1980, "Relationships between facial electromyography and subjective experience during affective imagery," *Elsevier*, Accessed: Sep. 10, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0301051180900265.

[120] J. T. Larsen, C. J. Norris, and J. T. Cacioppo, "Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii," in *Psychophysiology*, Sep. 2003, vol. 40, no. 5, pp. 776–785, doi: 10.1111/1469-8986.00078.

[121] T. S. H. Wingenbach, M. Brosnan, M. C. Pfaltz, P. Peyk, and C. Ashwin, "Perception of Discrete Emotions in Others: Evidence for Distinct Facial Mimicry Patterns," *Sci. Rep.*, vol. 10, no. 1, pp. 1–13, Dec. 2020, doi: 10.1038/s41598-020-61563-5.

[122] A. Van Boxtel, "Facial EMG as a tool for inferring affective states," *Proc. Meas. Behav.*, 2010.

[123] J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, "Psychophysiological science," in *Handbook of Psychophysiology*, 2000.

[124] J. T. Larsen, C. J. Norris, and J. T. Cacioppo, "Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii," in *Psychophysiology*, 2003, doi: 10.1111/1469-8986.00078.

[125] P. Ekman, "Lie Catching and Micro Expressions," *The Philosophy of Deception*. Oxford University Press., pp. 188–133, 2009, Accessed: Oct. 13, 2020. [Online]. Available: https://philpapers.org/rec/EKMLCA.

[126] A. F.-M.-E. of and undefined 2012, "Microexpression and macroexpression," *Oxford Elsevier/Academic Press*.

[127] D. Cernea and A. Kerren, "A survey of technologies on the rise for emotion-enhanced interaction," *J. Vis. Lang. Comput.*, vol. 31, pp. 70–86, Dec. 2015, doi:

10.1016/j.jvlc.2015.10.001.

[128]   J. Cacioppo, R. Petty, … M. L.-J. of personality, and undefined 1986, "Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions.," *psycnet.apa.org*, Accessed: Sep. 10, 2019. [Online]. Available: https://psycnet.apa.org/record/1986-13864-001.

[129]   L. G. Tassinary and J. T. Cacioppo, "Unobservable Facial Actions and Emotion," *Psychol. Sci.*, 1992, doi: 10.1111/j.1467-9280.1992.tb00252.x.

[130]   G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, "Boredom, engagement and anxiety as indicators for adaptation to difficulty in games," *Proc. 12th Int. Conf. Entertain. media ubiquitous era - MindTrek '08*, p. 13, 2008, doi: 10.1145/1457199.1457203.

[131]   R. L. Hazlett and J. Benedek, "Measuring emotional valence to understand the user's experience of software," *Int. J. Hum. Comput. Stud.*, 2007, doi: 10.1016/j.ijhcs.2006.11.005.

[132]   C. N. Huang, C. H. Chen, and H. Y. Chung, "The review of applications and measurements in facial electromyography," *Journal of Medical and Biological Engineering*. 2005.

[133]   Y. Golland, A. Hakim, T. Aloni, S. Schaefer, and N. Levit-Binnun, "Affect dynamics of facial EMG during continuous emotional experiences," *Biol. Psychol.*, vol. 139, pp. 47–58, Nov. 2018, doi: 10.1016/j.biopsycho.2018.10.003.

[134]   J. M. Kivikangas *et al.*, "A review of the use of psychophysiological methods in game research," *J. Gaming Virtual Worlds*, 2011, doi: 10.1386/jgvw.3.3.181_1.

[135]   E. Dalla-Toffola, D. Bossi, M. Buonocore, C. Montomoli, L. Petrucci, and E. Alfonsi, "Usefulness of BFB/EMG in facial palsy rehabilitation," *Disability and Rehabilitation*, vol. 27, no. 14. Taylor & Francis, pp. 809–815, Jul. 22, 2005, doi: 10.1080/09638280400018650.

[136]   H. M. Ryu *et al.*, "Study on the validity of surface electromyography as assessment tools for facial nerve palsy," *J. Pharmacopuncture*, vol. 21, no. 4, pp. 258–267, 2018, doi: 10.3831/KPI.2018.21.029.

[137]   G. Naik, D. Kumar, … V. S.-… on U. of vision in human, and undefined 2006, "Hand gestures for HCI using ICA of EMG," *dl.acm.org*, Accessed: Sep. 10, 2019. [Online]. Available: https://dl.acm.org/citation.cfm?id=1273397.

[138]   G. Yang, Y. Lin, and P. Bhattacharya, "A driver fatigue recognition model based on information fusion and dynamic Bayesian network," *Inf. Sci. (Ny).*, vol. 180, no. 10, pp. 1942–1954, May 2010, doi: 10.1016/j.ins.2010.01.011.

[139]   G. Krantz, M. Forsman, and U. Lundberg, "Consistency in physiological stress responses and electromyographic activity during induced stress exposure in women and men," *Integr. Physiol. Behav. Sci.*, vol. 39, no. 2, pp. 105–118, 2004, doi: 10.1007/BF02734276.

[140] J. R. Cram, "Surface EMG recordings and pain-related disorders: A diagnostic framework," *Biofeedback Self. Regul.*, vol. 13, no. 2, pp. 123–138, Jun. 1988, doi: 10.1007/BF01001496.

[141] A. J. Fridlund and J. T. Cacioppo, "Guidelines for Human Electromyographic Research," *Psychophysiology*, vol. 23, no. 5, pp. 567–589, 1986, doi: 10.1111/j.1469-8986.1986.tb00676.x.

[142] K. P. Miyapuram, C. Xing, M. Nascimben, and T. Z. Ramsøy, "A Minimal Setup for Spontaneous Smile Quantification Applicable for Valence Detection," *Front. Psychol. | www.frontiersin.org*, vol. 11, p. 566354, Dec. 2020, doi: 10.3389/fpsyg.2020.566354.

[143] M. Sharma, S. Kacker, and M. Sharma, "A Brief Introduction and Review on Galvanic Skin Response," *Int. J. Med. Res. Prof.*, vol. 2, no. 6, 2016, doi: 10.21276/ijmrp.2016.2.6.003.

[144] N. Karim and S. Ali, "Heart rate variability-A review," *J. Basic Appl. Sci. Vol. 7,* vol. 7, no. 1, pp. 71–77, 2011, Accessed: May 10, 2020. [Online]. Available: https://www.researchgate.net/publication/228664207.

[145] M. Merone, P. Soda, M. Sansone, C. S.-E. S. with, and undefined 2017, "ECG databases for biometric systems: A systematic review," *Elsevier*, Accessed: Sep. 11, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417416305140.

[146] Q. Yousef, M. B. I. Reaz, and M. A. M. Ali, "The Analysis of PPG Morphology: Investigating the Effects of Aging on Arterial Compliance," *Meas. Sci. Rev.*, vol. 12, no. 6, 2012, doi: 10.2478/v10048-012-0036-3.

[147] C. Kim, Y. J. Kim, H. Jung, and D. S. Han, "A real-time heartbeat estimation system using PPG signals," in *2017 IEEE International Conference on Consumer Electronics, ICCE 2017*, Mar. 2017, pp. 145–146, doi: 10.1109/ICCE.2017.7889263.

[148] N. Saquib, M. T. I. Papon, I. Ahmad, and A. Rahman, "Measurement of heart rate using photoplethysmography," in *Proceedings of 2015 International Conference on Networking Systems and Security, NSysS 2015*, 2015, doi: 10.1109/NSysS.2015.7043525.

[149] B. M. Appelhans and L. J. Luecken, "Heart rate variability as an index of regulated emotional responding," *Review of General Psychology*, vol. 10, no. 3. pp. 229–240, 2006, doi: 10.1037/1089-2680.10.3.229.

[150] K. Schaaff and M. T. P. Adam, "Measuring emotional arousal for online applications: Evaluation of ultra-short term heart rate variability measures," in *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, 2013, pp. 362–368, doi: 10.1109/ACII.2013.66.

[151] K. H. Choi, J. Kim, O. S. Kwon, M. J. Kim, Y. H. Ryu, and J. E. Park, "Is heart rate variability (HRV) an adequate tool for evaluating human emotions? – A focus on the

use of the International Affective Picture System (IAPS)," *Psychiatry Res.*, vol. 251, pp. 192–196, May 2017, doi: 10.1016/j.psychres.2017.02.025.

[152]   P. Nickel and F. Nachreiner, "Sensitivity and Diagnosticity of the 0.1-Hz Component of Heart Rate Variability as an Indicator of Mental Workload," in *Human Factors*, Dec. 2003, vol. 45, no. 4, pp. 575–590, doi: 10.1518/hfes.45.4.575.27094.

[153]   N. M.-A. in psychology and undefined 1988, "Heart rate variability and mental workload assessment," *Elsevier*, Accessed: Sep. 11, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166411508623845.

[154]   J. Taelman, S. Vandeput, A. Spaepen, and S. Van Huffel, "Influence of mental stress on heart rate and heart rate variability," in *IFMBE Proceedings*, 2008, vol. 22, pp. 1366–1369, doi: 10.1007/978-3-540-89208-3_324.

[155]   C. M. A. van Ravenswaaij-Arts, "Heart Rate Variability," *Ann. Intern. Med.*, vol. 118, no. 6, p. 436, Mar. 1993, doi: 10.7326/0003-4819-118-6-199303150-00008.

[156]   J. A. J. Heathers, "Everything Hertz: Methodological issues in short-term frequency-domain HRV," *Frontiers in Physiology*, vol. 5 MAY. Frontiers Research Foundation, p. 177, May 07, 2014, doi: 10.3389/fphys.2014.00177.

[157]   S. Lu *et al.*, "Can photoplethysmography variability serve as an alternative approach to obtain heart rate variability information?," *J. Clin. Monit. Comput.*, vol. 22, no. 1, pp. 23–29, Jan. 2008, doi: 10.1007/s10877-007-9103-y.

[158]   U. R. Acharya, K. P. Joseph, N. Kannathal, C. M. Lim, and J. S. Suri, "Heart rate variability: A review," *Medical and Biological Engineering and Computing*, vol. 44, no. 12. pp. 1031–1051, Dec. 2006, doi: 10.1007/s11517-006-0119-0.

[159]   F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," *Front. Public Heal.*, vol. 5, p. 258, Sep. 2017, doi: 10.3389/fpubh.2017.00258.

[160]   D. S. Quintana, A. J. Guastella, T. Outhred, I. B. Hickie, and A. H. Kemp, "Heart rate variability is associated with emotion recognition: Direct evidence for a relationship between the autonomic nervous system and social cognition," *Int. J. Psychophysiol.*, vol. 86, no. 2, pp. 168–172, Nov. 2012, doi: 10.1016/j.ijpsycho.2012.08.012.

[161]   P. R. Steffen, T. Austin, A. DeBarros, and T. Brown, "The Impact of Resonance Frequency Breathing on Measures of Heart Rate Variability, Blood Pressure, and Mood," *Front. Public Heal.*, vol. 5, Aug. 2017, doi: 10.3389/fpubh.2017.00222.

[162]   M. de Meijer, "The contribution of general features of body movement to the attribution of emotions," *J. Nonverbal Behav.*, 1989, doi: 10.1007/BF00990296.

[163]   N. Dael, M. Mortillaro, K. S.- Emotion, and undefined 2012, "Emotion expression in body action and posture.," *psycnet.apa.org*, Accessed: Sep. 10, 2019. [Online]. Available: https://psycnet.apa.org/record/2011-25181-001.

[164]   T. Shafir, R. P. Tsachor, and K. B. Welch, "Emotion regulation through movement:

Unique Sets of Movement Characteristics are Associated with and Enhance Basic Emotions," *Front. Psychol.*, vol. 6, no. JAN, 2016, doi: 10.3389/fpsyg.2015.02030.

[165] A. S. Won, B. Perone, M. Friend, and J. N. Bailenson, "Identifying Anxiety Through Tracked Head Movements in a Virtual Classroom," *Cyberpsychology, Behav. Soc. Netw.*, vol. 19, no. 6, pp. 380–387, Jun. 2016, doi: 10.1089/cyber.2015.0326.

[166] P. Alinia, H. Ghasemzadeh, and G. Fortino, "BodyCloud: A Cloud-assisted Software Platform for Pervasive and Continuous Monitoring of Assisted Livings using Wearable and Mobile Devices View project Design of a MEMS-based, wireless sensor board for vibration monitoring View project Raffaele Gravina Università della Calabria," *Inf. Fusion*, 2016, doi: 10.1016/j.inffus.2016.09.005.

[167] F. Pollick, H. Paterson, A. Bruderlin, A. S.- Cognition, and undefined 2001, "Perceiving affect from arm movement," *Elsevier*, Accessed: Sep. 10, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010027701001470.

[168] H. Zacharatos, C. Gatzoulis, and Y. L. Chrysanthou, "Automatic emotion recognition based on body movement analysis: A survey," *IEEE Comput. Graph. Appl.*, vol. 34, no. 6, pp. 35–45, Nov. 2014, doi: 10.1109/MCG.2014.106.

[169] W. IJsselsteijn, "Elements of a multi-level theory of presence: Phenomenology, mental processing and neural correlates," *Proc. Presence 2002 2nd Int. Work. Presence*, 2002, doi: 10.1.1.12.7866.

[170] R. M. Baños, C. Botella, M. Alcañiz, V. Liaño, B. Guerrero, and B. Rey, "Immersion and Emotion: Their Impact on the Sense of Presence," *CyberPsychology Behav.*, 2004, doi: 10.1089/cpb.2004.7.734.

[171] A. Chirico, P. Cipresso, D. B. Yaden, F. Biassoni, G. Riva, and A. Gaggioli, "Effectiveness of Immersive Videos in Inducing Awe: An Experimental Study," *Sci. Rep.*, vol. 7, no. 1, Dec. 2017, doi: 10.1038/s41598-017-01242-0.

[172] A. Chirico and A. Gaggioli, "When Virtual Feels Real: Comparing Emotional Responses and Presence in Virtual and Natural Environments," *Cyberpsychology, Behav. Soc. Netw.*, vol. 22, no. 3, pp. 220–226, Mar. 2019, doi: 10.1089/cyber.2018.0393.

[173] C. J. Wilson and A. Soranzo, "The Use of Virtual Reality in Psychology: A Case Study in Visual Perception," *Computational and Mathematical Methods in Medicine*, vol. 2015. 2015, doi: 10.1155/2015/151702.

[174] J. Diemer, G. W. Alpers, H. M. Peperkorn, Y. Shiban, and A. Mühlberger, "The impact of perception and presence on emotional reactions: A review of research in virtual reality," *Frontiers in Psychology*, vol. 6, no. JAN. Frontiers Media SA, p. 26, 2015, doi: 10.3389/fpsyg.2015.00026.

[175] M. Slater and S. Wilbur, "A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments," *Presence*

Teleoperators Virtual Environ., 1997, doi: 10.1007/s10750-008-9541-7.

[176] T. Schubert, F. Friedmann, and H. Regenbrecht, "The Experience of Presence: Factor Analytic Insights," *Presence Teleoperators Virtual Environ.*, 2001, doi: 10.1162/105474601300343603.

[177] J. Steuer, "Defining Virtual Reality: Dimensions Determining Telepresence," *J. Commun.*, 1992, doi: 10.1111/j.1460-2466.1992.tb00812.x.

[178] M. Lombard and T. Ditton, "At the Heart of It All: The Concept of Presence," *J. Comput. Commun.*, 2006, doi: 10.1111/j.1083-6101.1997.tb00072.x.

[179] Y. Ling, H. T. Nefs, W. P. Brinkman, C. Qu, and I. Heynderickx, "The relationship between individual characteristics and experienced presence," *Comput. Human Behav.*, vol. 29, no. 4, pp. 1519–1530, 2013, doi: 10.1016/j.chb.2012.12.010.

[180] J. Diemer, G. W. Alpers, H. M. Peperkorn, Y. Shiban, and A. Mühlberger, "The impact of perception and presence on emotional reactions: A review of research in virtual reality," *Frontiers in Psychology*. 2015, doi: 10.3389/fpsyg.2015.00026.

[181] B. G. Witmer and M. J. Singer, "Measuring presence in virtual environments: A presence questionnaire," *Presence Teleoperators Virtual Environ.*, vol. 7, no. 3, pp. 225–240, 1998, doi: 10.1162/105474698565686.

[182] M. Slater and M. Usoh, "Presence in immersive virtual environments," in *1993 IEEE Annual Virtual Reality International Symposium*, 1993, pp. 90–96, doi: 10.1109/vrais.1993.380793.

[183] M. Lombard and T. Ditton, "At the Heart of It All: The Concept of Presence," *J. Comput. Commun.*, vol. 3, no. 2, pp. 0–0, Jun. 2006, doi: 10.1111/j.1083-6101.1997.tb00072.x.

[184] J. Lessiter, J. Freeman, E. Keogh, and J. Davidoff, "A Cross-Media Presence Questionnaire: The ITC-Sense of Presence Inventory," *Presence Teleoperators Virtual Environ.*, vol. 10, no. 3, pp. 282–297, 2001, doi: 10.1162/105474601300343612.

[185] M. J. Schuemie, P. Van der Straaten, M. Krijn, and C. A. P. G. Van der Mast, "Research on presence in virtual reality: A survey," *Cyberpsychology and Behavior*, vol. 4, no. 2. Mary Ann Liebert, Inc, pp. 183–201, 2001, doi: 10.1089/109493101300117884.

[186] B. G. Witmer and M. J. Singer, "Measuring presence in virtual environments: A presence questiionnaire," *Presence*, 1998.

[187] M. Slater, D. P. Pertaub, and A. Steed, "Public Speaking in Virtual Reality: Facing an Audience of Avatars," *IEEE Comput. Graph. Appl.*, vol. 19, no. 2, pp. 6–9, Mar. 1999, doi: 10.1109/38.749116.

[188] M. Slater, "A Note on Presence Terminology," *Presence Connect*, vol. 3, no. 3, pp. 1–5, 2003, Accessed: Aug. 11, 2017. [Online]. Available: www.cs.ucl.ac.uk/staff/m.slater.

[189] M. Slater and S. Wilbur, "A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments," *Presence Teleoperators Virtual Environ.*, 1997, doi: 10.1007/s10750-008-9541-7.

[190] C. Botella *et al.*, "Differences on presence and reality judgment using a high impact workstation and a PC workstation," *Cyberpsychology Behav.*, vol. 2, no. 1, pp. 49–52, 1999, doi: 10.1089/cpb.1999.2.49.

[191] C. Heeter, "Being There: The Subjective Experience of Presence," *Presence*, 1992, doi: 10.1109/VRAIS.1995.512482.

[192] M. Lombard, R. Reich, M. Grabe, C. Bracken, and T. Ditton, "Presence and television..," *Hum. Commun. Res.*, vol. 26, no. 1, pp. 75–98, Jan. 2000, doi: 10.1111/j.1468-2958.2000.tb00750.x.

[193] "VIVE™ United Kingdom | Buy VIVE Hardware." https://www.vive.com/uk/product/ (accessed Sep. 10, 2019).

[194] "Oculus." https://www.oculus.com/ (accessed Sep. 12, 2019).

[195] "PlayStation VR | The VR gaming system for PS4 | PlayStation." https://www.playstation.com/en-gb/explore/playstation-vr/ (accessed Sep. 12, 2019).

[196] J. Freeman, J. Lessiter, K. Pugh, and E. Keogh, "When presence and emotion are related, and when they are not," *8th Annu. Int. Work. Presence*, 2005.

[197] H. S. Wallach, M. P. Safir, and R. Samana, "Personality variables and presence," *Virtual Real.*, vol. 14, no. 1, pp. 3–13, 2010, doi: 10.1007/s10055-009-0124-3.

[198] H. Regenbrecht and T. Schubert, "Measuring the Sense of Presence and its Relations to Fear of Heights in Virtual Environments Human Penguin Project View project Removing Spatial Boundaries in Immersive Mobile Communication View project," *Artic. Int. J. Human-Computer Interact.*, vol. 10, no. 3, pp. 233–249, 1998, doi: 10.1207/s15327590ijhc1003_2.

[199] G. Riva *et al.*, "Affective Interactions Using Virtual Reality: The Link between Presence and Emotions," *CyberPsychology Behav.*, vol. 10, no. 1, pp. 45–56, 2007, doi: 10.1089/cpb.2006.9993.

[200] A. Mihai, D. Opriș, and D. David, "The Effects of Expectancies on Presence in Virtual Environments: A Brief Research Report."

[201] T. B. Sheridan, "Further musings on the psychophysics of presence," *Presence: Teleoperators and Virtual Environments*, vol. 5, no. 2. MIT Press Journals, pp. 241–246, 1996, doi: 10.1162/pres.1996.5.2.241.

[202] A. K. Seth, K. Suzuki, and H. D. Critchley, "An interoceptive predictive coding model of conscious presence," *Front. Psychol.*, vol. 3, no. JAN, 2012, doi: 10.3389/fpsyg.2011.00395.

[203] S. C. Wilson and T. X. Barber, "Vivid Fantasy and Hallucinatory Abilities in the Life Histories of Excellent Hypnotic Subjects ('Somnambules'): Preliminary Report with

Female Subjects," in *Concepts, Results, and Applications*, Springer US, 1981, pp. 133–149.

[204] H. L. G. J. Merckelbach, & Muris, H. Merckelbach, R. Horselenberg, and P. Muris, "The Creative Experiences Questionnaire (CEQ): A brief self-report measure of fantasy proneness Citation for published version (APA): The Creative Experiences Questionnaire (CEQ): a brief self-report measure of fantasy proneness," doi: 10.1016/S0191-8869(00)00201-4.

[205] K. E. O'Grady, "The absorption scale: A factor-analytic assessment," *Int. J. Clin. Exp. Hypn.*, vol. 28, no. 3, pp. 281–288, Jul. 1980, doi: 10.1080/00207148008409853.

[206] A. T. Church, "Relating the Tellegen and Five-Factor Models of Personality Structure," *J. Pers. Soc. Psychol.*, vol. 67, no. 5, pp. 898–909, 1994, doi: 10.1037/0022-3514.67.5.898.

[207] C. Johns, D. Nuñez, M. Daya, D. Sellars, J. Casanueva, and E. Blake, "The Interaction Between Individuals' Immersive Tendencies and the Sensation of Presence in a Virtual Environment," 2000, pp. 65–74.

[208] M. Lombard, T. B. Ditton, and L. Weinstein, "Measuring Presence: The Temple Presence Inventory," *Proceeding Presence 2009 12th Int. Work. Presence*, 2009.

[209] M. Meehan, B. Insko, M. Whitton, and F. Brooks, "Physiological measures of presence in virtual environments," in *Proceedings of Presence 2001 : the 4th international workshop on presence*, 2001.

[210] B. Wiederhold, D. Jang, M. K.-… : Mind, cognitions and, and undefined 2001, "An investigation into physiological responses in virtual environments: an objective measurement of presence," *academia.edu*, Accessed: May 11, 2020. [Online]. Available:
http://www.academia.edu/download/41180724/An_Investigation_into_Physiologic al_Resp20160115-14664-1kdsmrm.pdf.

[211] J. Diemer, A. Mühlberger, P. Pauli, and P. Zwanzger, "Virtual reality exposure in anxiety disorders: Impact on psychophysiological reactivity," *World Journal of Biological Psychiatry*, vol. 15, no. 6. Informa Healthcare, pp. 427–442, 2014, doi: 10.3109/15622975.2014.892632.

[212] L. J. Levine and M. A. Safer, "Sources of bias in memory for emotions," *Curr. Dir. Psychol. Sci.*, vol. 11, no. 5, pp. 169–173, 2002, doi: 10.1111/1467-8721.00193.

[213] M. Slater, "Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 364, no. 1535, pp. 3549–3557, Dec. 2009, doi: 10.1098/rstb.2009.0138.

[214] D. T. Levin and D. J. Simons, "Perceiving Stability in a Changing World: Combining Shots and Intergrating Views in Motion Pictures and the Real World," *Media Psychol.*, 2000, doi: 10.1207/S1532785XMEP0204_03.

[215] D. Villani, M. Lucchetta, A. Preziosa, and G. Riva, "The role of interactive media features on the affective response : a virtual reality study," *eMinds Int. J. HumanComputer Interact.*, vol. I, no. 5, pp. 35–55, 2009.

[216] G. Robillard, S. Bouchard, T. Fournier, and P. Renaud, "Anxiety and presence during VR immersion: a comparative study of the reactions of phobic and non-phobic participants in therapeutic virtual environments derived from computer games.," *Cyberpsychol. Behav.*, vol. 6, no. 5, pp. 467–476, 2003, doi: 10.1089/109493103769710497.

[217] R. M. Baños, C. Botella, I. Rubió, S. Quero, A. García-Palacios, and M. Alcañiz, "Presence and Emotions in Virtual Environments: The Influence of Stereoscopy," *CyberPsychology Behav.*, vol. 11, no. 1, pp. 1–8, 2008, doi: 10.1089/cpb.2007.9936.

[218] S. Bouchard, J. St-Jacques, G. Robillard, and P. Renaud, "Anxiety Increases the Feeling of Presence in Virtual Reality," *Presence Teleoperators Virtual Environ.*, vol. 17, no. 4, pp. 376–391, Aug. 2008, doi: 10.1162/pres.17.4.376.

[219] S. Estupiñán, F. Rebelo, P. Noriega, C. Ferreira, and E. Duarte, "Can virtual reality increase emotional responses (Arousal and Valence)? a pilot study," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014, vol. 8518 LNCS, no. PART 2, pp. 541–549, doi: 10.1007/978-3-319-07626-3_51.

[220] A. Felnhofer, O. D. Kothgassner, T. Hetterle, L. Beutl, H. Hlavacs, and I. Kryspin-Exner, "Afraid to Be There? Evaluating the Relation Between Presence, Self-Reported Anxiety, and Heart Rate in a Virtual Public Speaking Task," *Cyberpsychology, Behav. Soc. Netw.*, vol. 17, no. 5, pp. 310–316, May 2014, doi: 10.1089/cyber.2013.0472.

[221] M. V. Sanchez-Vives and M. Slater, "From presence to consciousness through virtual reality," *Nature Reviews Neuroscience*, vol. 6, no. 4. Nature Publishing Group, pp. 332–339, 2005, doi: 10.1038/nrn1651.

[222] R. Dietz and A. Lang, "Affective agents: Effects of agent affect on arousal, attention, liking and learning," in *Cognitive Technology Conference*, 1999.

[223] A. Lang, J. Newhagen, and B. Reeves, "Negative video as structure: Emotion, attention, capacity, and memory," *J. Broadcast. Electron. Media*, vol. 40, no. 4, pp. 460–477, 1996, doi: 10.1080/08838159609364369.

[224] F. Scholcover, S. J. Cauffman, and D. J. Gillan, "TIMING AND PHYSIOLOGICAL AROUSAL: IMPLICATIONS FOR HUMAN PERFORMANCE," 2018, doi: 10.1177/1541931218621345.

[225] "Classics in the History of Psychology -- Yerkes & Dodson (1908)." http://psychclassics.yorku.ca/Yerkes/Law/ (accessed May 25, 2020).

[226] F. S. Erthal *et al.*, "Load-dependent modulation of affective picture processing," *Cogn. Affect. Behav. Neurosci.*, 2005, doi: 10.3758/CABN.5.4.388.

[227] J. De Houwer and H. Tibboel, "Stop what you are not doing! Emotional pictures interfere with the task not to respond.," *Psychon. Bull. Rev.*, 2010, doi: 10.3758/PBR.17.5.699.

[228] M. G. Pereira *et al.*, "Emotion affects action: Midcingulate cortex as a pivotal node of interaction between negative emotion and motor signals," *Cogn. Affect. Behav. Neurosci.*, vol. 10, no. 1, pp. 94–106, 2010, doi: 10.3758/CABN.10.1.94.

[229] A. C. Savine, S. M. Beck, B. G. Edwards, K. S. Chiew, and T. S. Braver, "Enhancement of cognitive control by approach and avoidance motivational states.," *Cogn. Emot.*, 2010, doi: 10.1080/02699930903381564.

[230] J. Zhou, T. Gao, Y. Zhang, J. Liang, R. Shui, and M. Shen, "Losing control in front of a fearful face: The effect of emotional information on cognitive control," *Q. J. Exp. Psychol.*, 2011, doi: 10.1080/17470218.2010.543282.

[231] D. Palomba, A. Angrilli, and A. Mini, "Visual evoked potentials, heart rate responses and memory to emotional pictorial stimuli.," *Int. J. Psychophysiol.*, 1997.

[232] S. Schweizer and T. Dalgleish, "The impact of affective contexts on working memory capacity in healthy populations and in individuals with PTSD.," *Emotion*, 2016, doi: 10.1037/emo0000072.

[233] N. A. Ryckman, D. R. Addis, A. J. Latham, and A. J. Lambert, "Forget about the future: effects of thought suppression on memory for imaginary emotional episodes," *Cogn. Emot.*, vol. 32, no. 1, pp. 200–206, 2018, doi: 10.1080/02699931.2016.1276049.

[234] A. Sutcliffe, B. Gault, and J. E. Shin, "Presence, memory and interaction in virtual environments," *Int. J. Hum. Comput. Stud.*, vol. 62, no. 3, pp. 307–327, 2005, doi: 10.1016/j.ijhcs.2004.11.010.

[235] A. Zlomuzica, F. Preusser, C. Totzeck, E. Dere, and J. Margraf, "The impact of different emotional states on the memory for what, where and when features of specific events," *Behav. Brain Res.*, vol. 298, pp. 181–187, 2016, doi: 10.1016/j.bbr.2015.09.037.

[236] J. Bailey, J. Bailenson, A. Won, J. Flora, and K. Armel, "Presence and Memory: Immersive Virtual Reality Effects on Cued Recall," *Proc. Int. Soc. Presence Res. Annu. Conf.*, 2011.

[237] J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam, "Emotion and Decision Making," *Annu. Rev. Psychol.*, vol. 66, no. 1, pp. 799–823, Jan. 2015, doi: 10.1146/annurev-psych-010213-115043.

[238] A. Miu, R. Heilman, D. H.-B. psychology, and undefined 2008, "Anxiety impairs decision-making: psychophysiological evidence from an Iowa Gambling Task," *Elsevier*, Accessed: May 25, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301051107001962.

[239] C. Becker-Asano, D. Sun, B. Kleim, C. N. Scheel, and B. Nebel, "Outline of an

empirical study on the effects of emotions on strategic behavior in virtual emergencies."

[240] N. H. Frijda, "The Laws of Emotion," *Am. Psychol.*, 1988, doi: 10.5465/AMR.2007.25275690.

[241] I. Roseman, C. S.-A. processes in emotion: Theory, and undefined 2001, "Appraisal theory," *books.google.com*, Accessed: Sep. 16, 2019. [Online]. Available: https://books.google.co.uk/books?hl=en&lr=&id=fmtnDAAAQBAJ&oi=fnd&pg= PA3&dq=Roseman+%26+Smith+2001&ots=sqPcVUcePp&sig=XE4spEOZowLX NE02BDLxcV-2ubU.

[242] M. Falkenstein, J. Hohnsbein, J. Hoormann, and L. Blanke, "Effects of crossmodal divided attention on ERP components: Error processing in choice reaction tasks," *Electroencephalogr. Clin. Neurophysiol.*, 1991.

[243] N. Derakshan, S. Smyth, and M. W. Eysenck, "Effects of state anxiety on performance using a task-switching paradigm: an investigation of attentional control theory," *Psychon Bull Rev*, 2009, doi: 10.3758/PBR.16.6.1112.

[244] T. Letzring, L. A.-P. and I. Differences, and undefined 2015, "Personality traits and affective states: Relationships with and without affect induction," *Elsevier*, Accessed: Sep. 16, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0191886914006436.

[245] K. Muir, A. Madill, and C. Brown, "Individual differences in emotional processing and autobiographical memory: interoceptive awareness and alexithymia in the fading affect bias," *Cogn. Emot.*, vol. 31, no. 7, pp. 1392–1404, Oct. 2017, doi: 10.1080/02699931.2016.1225005.

[246] S. Gaetan, V. Bréjard, and A. Bonnet, "Video games in adolescence and emotional functioning: Emotion regulation, emotion intensity, emotion expression, and alexithymia," *Comput. Human Behav.*, vol. 61, pp. 344–349, Aug. 2016, doi: 10.1016/j.chb.2016.03.027.

[247] G. J. Taylor, "Recent developments in alexithymia theory and research," *Canadian Journal of Psychiatry*, vol. 45, no. 2. Canadian Psychiatric Association, pp. 134–142, 2000, doi: 10.1177/070674370004500203.

[248] G. Panayiotou, M. Panteli, and E. Vlemincx, "Processing Emotions in Alexithymia: A Systematic Review of Physiological Markers," in *Alexithymia*, Cambridge University Press, 2018, pp. 291–320.

[249] J. Gross, O. John, and J. Richards, *Berkeley expressivity questionnaire*. 1995.

[250] J. J. Gross and O. E. John, "Revealing Feelings: Facets of Emotional Expressivity in Self-Reports, Peer Ratings, and Behavior," 1997.

[251] "Homepage - Emotiv." https://www.emotiv.com/ (accessed Sep. 10, 2019).

[252] "Looxid Labs – Explore User Mind." https://looxidlabs.com/ (accessed Sep. 10, 2019).

[253] "Products / ENOBIO - Neuroelectrics." https://www.neuroelectrics.com/products/enobio/ (accessed Sep. 10, 2019).

[254] "Neurable." http://neurable.com/ (accessed Sep. 10, 2019).

[255] "emteq." https://emteq.net/ (accessed Sep. 10, 2019).

[256] "Home - Affectiva : Affectiva." https://www.affectiva.com/ (accessed Sep. 10, 2019).

[257] G. W. I. T. and G. (ITAG) and undefined 2016, "Virtual reality exposure therapy: To benefit those who stutter and treat social anxiety," *ieeexplore.ieee.org*, Accessed: Sep. 10, 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7782512/.

[258] "SDK - Affectiva : Affectiva." https://www.affectiva.com/product/emotion-sdk/ (accessed Sep. 10, 2019).

[259] M. Horvat, M. Dobrinić, M. Novosel, and P. Jerčić, "Assessing emotional responses induced in virtual reality using a consumer EEG headset: A preliminary report."

[260] A. Pereira *et al.*, "Cross-Subject EEG Event-Related Potential Classification for Brain-Computer Interfaces Using Residual Networks." Accessed: Sep. 10, 2019. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01878227.

[261] J. Jantz, A. Molnar, R. A.-A. S. 2017 V. Village, and undefined 2017, "A brain-computer interface for extended reality interfaces," *dl.acm.org*, Accessed: Sep. 10, 2019. [Online]. Available: https://dl.acm.org/citation.cfm?id=3089290.

[262] "Developers | Neurable INC." http://www.neurable.com/developers (accessed Sep. 10, 2019).

[263] M. M. M. Èller, A. Keil, T. Gruber, and T. Elbert, "Processing of affective pictures modulates right-hemispheric gamma band EEG activity." Accessed: Sep. 10, 2019. [Online]. Available: http://www.ub.uni-konstanz.de/kops/volltexte/2007/4346/.

[264] F. Škola and F. Liarokapis, "Examining the effect of body ownership in immersive virtual and augmented reality environments," *Vis. Comput.*, vol. 32, no. 6–8, pp. 761–770, Jun. 2016, doi: 10.1007/s00371-016-1246-8.

[265] I. Mavridou, E. Seiss, M. Hamedi, and E. Balaguer-Ballester, "Towards valence detection from EMG for Virtual Reality applications," 2018, Accessed: Sep. 10, 2019. [Online]. Available: http://eprints.bournemouth.ac.uk/31022/.

[266] L. Reidi, "A Gamified Approach to Cognitive Training using Virtual Reality and Affective Adaptation," University of Cambridge, 2018.

[267] W. Barfield, C. Hendrix, and K. E. Bystrom, "Effects of stereopsis and head tracking on performance using desktop virtual environment displays," *Presence: Teleoperators and Virtual Environments*, vol. 8, no. 2. MIT Press Journals, pp. 237–240, 1999, doi: 10.1162/105474699566198.

[268] "Empatica | Medical devices, AI and algorithms for remote patient monitoring." https://www.empatica.com/en-gb/index.html (accessed Jun. 27, 2021).

[269] A. F. Bulagang, J. Mountstephens, and J. Teo, "Multiclass emotion prediction using heart rate and virtual reality stimuli," *J. Big Data*, vol. 8, no. 1, pp. 1–12, Dec. 2021, doi: 10.1186/s40537-020-00401-x.

[270] A. Borrego, J. Latorre, … M. A.-… C. on V., and undefined 2019, "Reliability of the Empatica E4 wristband to measure electrodermal activity to emotional stimuli," *ieeexplore.ieee.org*, Accessed: Jun. 27, 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8994546/?casa_token=znwrE0K7ShY AAAAA:4fXW_LgB7nN3e_8yRXDWyCfm0Ck_evpiI1y87tQgTf76FSzQSYKZa NzWN3mVfyx6S-qvbQFO-8qp.

[271] "Our Technology - emteq labs | Measure what matters during immersive experiences." https://www.emteqlabs.com/our-technology/ (accessed Jun. 27, 2021).

[272] I. Mavridou *et al.*, "FACETEQ interface demo for emotion expression in VR," in *Proceedings - IEEE Virtual Reality*, 2017, pp. 441–442, doi: 10.1109/VR.2017.7892369.

[273] J. Blascovich, J. Loomis, A. C. Beall, K. R. Swinth, C. L. Hoyt, and J. N. Bailenson, "Immersive virtual environment technology as a methodological tool for social psychology," *Psychol. Inq.*, vol. 13, no. 2, pp. 103–124, 2002, doi: 10.1207/S15327965PLI1302_01.

[274] B. J. Li, J. N. Bailenson, A. Pines, W. J. Greenleaf, and L. M. Williams, "A public database of immersive VR videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures," *Front. Psychol.*, vol. 8, no. DEC, Dec. 2017, doi: 10.3389/fpsyg.2017.02116.

[275] J. I. Gold, K. A. Belmont, and D. A. Thomas, "The Neurobiology of Virtual Reality Pain Attenuation," *CYBERPSYCHOLOGY Behav.*, vol. 10, no. 4, 2007, doi: 10.1089/cpb.2007.9993.

[276] G. Riva, "Virtual Reality in Psychotherapy: Review," *CyberPsychology Behav.*, 2005, doi: 10.1089/cpb.2005.8.220.

[277] K. M. Malloy and L. S. Milling, "The effectiveness of virtual reality distraction for pain reduction: A systematic review," *Clinical Psychology Review*, vol. 30, no. 8. pp. 1011–1018, Dec. 2010, doi: 10.1016/j.cpr.2010.07.001.

[278] D.-P. Pertaub, M. Slater, and C. Barker, "An experiment on public speaking anxiety in response to three different types of virtual audience," *Presence Teleoperators Virtual Environ.*, vol. 11, no. 1, pp. 68–78, Feb. 2002, doi: 10.1162/105474602317343668.

[279] M. Price and P. Anderson, "The role of presence in virtual reality exposure therapy," *J. Anxiety Disord.*, vol. 21, no. 5, pp. 742–751, 2007, doi: 10.1016/j.janxdis.2006.11.002.

[280] I. Alsina-Jurnet and J. Gutiérrez-Maldonado, "Influence of personality and individual abilities on the sense of presence experienced in anxiety triggering virtual

environments," *Int. J. Hum. Comput. Stud.*, vol. 68, no. 10, pp. 788–801, 2010, doi: 10.1016/j.ijhcs.2010.07.001.

[281] S. Notzon, S. Deppermann, A. Fallgatter, … J. D.-B., and undefined 2015, "Psychophysiological effects of an iTBS modulated virtual reality challenge including participants with spider phobia," *Elsevier*, Accessed: Sep. 12, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301051115300648.

[282] A. P. Anderson, M. D. Mayer, A. M. Fellows, D. R. Cowan, M. T. Hegel, and J. C. Buckey, "Relaxation with immersive natural scenes presented using virtual reality," *Aerosp. Med. Hum. Perform.*, vol. 88, no. 6, pp. 520–526, Jun. 2017, doi: 10.3357/AMHP.4747.2017.

[283] C. Crescentini, L. Chittaro, V. Capurso, R. Sioni, and F. Fabbro, "Psychological and physiological responses to stressful situations in immersive virtual reality: Differences between users who practice mindfulness meditation and controls," *Comput. Human Behav.*, vol. 59, pp. 304–316, 2016, doi: 10.1016/j.chb.2016.02.031.

[284] N. Vaughan, T. Wainwright, V. N. Dubey, T. W. Wainwright, and R. G. Middleton, "A Review of Virtual Reality Based Training Simulators for Orthopaedic Surgery SEE PROFILE A Review of Virtual Reality Based Training Simulators for Orthopaedic Surgery," doi: 10.1016/j.medengphy.2015.11.021.

[285] A. Borrego, J. Latorre, M. Alcañiz, and R. Llorens, "Comparison of Oculus Rift and HTC Vive: Feasibility for Virtual Reality-Based Exploration, Navigation, Exergaming, and Rehabilitation," *Games Health J.*, vol. 7, no. 3, pp. 151–156, Jun. 2018, doi: 10.1089/g4h.2017.0114.

[286] S. Bronner, R. Pinsker, R. Naik, J. N.-J. of science and medicine in, and undefined 2016, "Physiological and psychophysiological responses to an exer-game training protocol," *Elsevier*, Accessed: Aug. 27, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1440244015000584.

[287] M. Bellani, L. Fornasari, L. Chittaro, and P. Brambilla, "Virtual reality in autism: State of the art," *Epidemiol. Psychiatr. Sci.*, vol. 20, no. 3, pp. 235–238, Sep. 2011, doi: 10.1017/S2045796011000448.

[288] S. Parsons and P. Mitchell, "The potential of virtual reality in social skills training for people with autistic spectrum disorders," *Journal of Intellectual Disability Research*, vol. 46, no. 5. pp. 430–443, 2002, doi: 10.1046/j.1365-2788.2002.00425.x.

[289] L. R. Valmaggia, L. Latif, M. J. Kempton, and M. Rus-Calafell, "Virtual reality in the psychological treatment for mental health problems: An systematic review of recent evidence," *Psychiatry Res.*, vol. 236, pp. 189–195, Feb. 2016, doi: 10.1016/j.psychres.2016.01.015.

[290] Y. Bian *et al.*, "A framework for physiological indicators of flow in VR games: construction and preliminary evaluation," *Pers. Ubiquitous Comput.*, vol. 20, no. 5,

pp. 821–832, Oct. 2016, doi: 10.1007/s00779-016-0953-5.

[291] D. Västfjäll, P. Larsson, and M. Kleiner, "Emotion and auditory virtual environments: affect-based judgments of music reproduced with virtual reverberation times.," *Cyberpsychol. Behav.*, vol. 5, no. 1, pp. 19–32, Feb. 2002, doi: 10.1089/109493102753685854.

[292] B. L. Fredrickson, "The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions," *Am. Psychol.*, vol. 56, no. 3, pp. 218–226, 2001, doi: 10.1037/0003-066X.56.3.218.

[293] D. R. Herring, M. H. Burleson, N. A. Roberts, and M. J. Devine, "Coherent with laughter: Subjective experience, behavior, and physiological responses during amusement and joy," *Int. J. Psychophysiol.*, vol. 79, no. 2, pp. 211–218, Feb. 2011, doi: 10.1016/j.ijpsycho.2010.10.007.

[294] J. Marín-Morales *et al.*, "Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors," *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, doi: 10.1038/s41598-018-32063-4.

[295] M. Alcañiz, R. Baños, C. Botella, B. Rey, and M. Alcañiz Raya, "The EMMA Project: Emotions as a Determinant of Presence," 2003.

[296] A. Felnhofer *et al.*, "Is virtual reality emotionally arousing? Investigating five emotion inducing virtual park scenarios," *Int. J. Hum. Comput. Stud.*, 2015, doi: 10.1016/j.ijhcs.2015.05.004.

[297] J. López-Tarruella Maldonado, J. L. Higuera Trujillo, S. Iñarra Abad, M. . Carmen Llinares Millán, J. Guixeres Provinciales, and M. A. Raya, "Virtual Reality as a Tool for Emotional Evaluation of Architectural Environments," in *Architectural Draughtsmanship*, Springer International Publishing, 2018, pp. 889–903.

[298] G. Franz, M. Von Der Heyde, and H. H. Bülthoff, "An empirical approach to the experience of architectural space in virtual reality-exploring relations between features and affective appraisals of rectangular indoor spaces," in *Automation in Construction*, 2005, vol. 14, no. 2 SPEC. ISS., pp. 165–172, doi: 10.1016/j.autcon.2004.07.009.

[299] O. Vartanian *et al.*, "Architectural design and the brain: Effects of ceiling height and perceived enclosure on beauty judgments and approach-avoidance decisions," *J. Environ. Psychol.*, vol. 41, pp. 10–18, Mar. 2015, doi: 10.1016/j.jenvp.2014.11.006.

[300] L. B. Fich, P. Jönsson, P. H. Kirkegaard, M. Wallergård, A. H. Garde, and Å. Hansen, "Can architectural design alter the physiological reaction to psychosocial stress? A virtual TSST experiment," *Physiol. Behav.*, vol. 135, pp. 91–97, Aug. 2014, doi: 10.1016/j.physbeh.2014.05.034.

[301] J. L. Higuera-Trujillo, J. López-Tarruella Maldonado, and C. Llinares Millán, "Psychological and physiological human responses to simulated and real environments: A comparison between Photographs, 360° Panoramas, and Virtual

Reality," *Appl. Ergon.*, vol. 65, pp. 398–409, Nov. 2017, doi: 10.1016/j.apergo.2017.05.006.

[302] B. Rohrmann, I. B.-J. of E. Psychology, and undefined 2002, "Subjective responses to computer simulations of urban environments," *Elsevier*, Accessed: Dec. 29, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0272494401902063.

[303] B. John, "Pupil diameter as a measure of emotion and sickness in VR," 2019, pp. 1–3, doi: 10.1145/3314111.3322868.

[304] P. Robitaille and M. J. McGuffin, "Increased affect-arousal in VR can be detected from faster body motion with increased heart rate," in *Proceedings - I3D 2019: ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, May 2019, doi: 10.1145/3306131.3317022.

[305] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010, doi: 10.1109/T-AFFC.2010.1.

[306] A. Panning *et al.*, "Multimodal affect recognition in spontaneous HCI environment," in *2012 IEEE International Conference on Signal Processing, Communications and Computing, ICSPCC 2012*, 2012, pp. 430–435, doi: 10.1109/ICSPCC.2012.6335662.

[307] J. Fleureau, P. Guillotel, and H. T. Quan, "Physiological-based affect event detector for entertainment video applications," *IEEE Trans. Affect. Comput.*, vol. 3, no. 3, pp. 379–385, Jul. 2012, doi: 10.1109/T-AFFC.2012.2.

[308] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. Affect. Comput.*, vol. 2, no. 2, pp. 92–105, 2011, doi: 10.1109/T-AFFC.2011.9.

[309] H. Gunes and M. Pantic, "Automatic, Dimensional and Continuous Emotion Recognition," *Int. J. Synth. Emot.*, 2010, doi: 10.4018/jse.2010101605.

[310] M. Singh, M. Singh, and N. Singhal, "Emotion Recognition along Valence Axis Using Naïve Bayes Classifier," 2013.

[311] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer, "Technique for automatic emotion recognition by body gesture analysis," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops*, 2008, doi: 10.1109/CVPRW.2008.4563173.

[312] K. Takahashi, "Remarks on emotion recognition from multi-modal bio-potential signals," in *Proceedings of the IEEE International Conference on Industrial Technology*, 2004, vol. 3, pp. 1138–1143, doi: 10.1109/icsmc.2003.1244650.

[313] O. AlZoubi, D. Fossati, S. D'Mello, and R. A. Calvo, "Affect detection from non-stationary physiological data using ensemble classifiers," *Evol. Syst.*, vol. 6, no. 2, pp. 79–92, 2015, doi: 10.1007/s12530-014-9123-z.

[314] D. Wu *et al.*, "Optimal arousal identification and classification for affective computing using physiological signals: Virtual reality stroop task," *IEEE Trans. Affect. Comput.*, vol. 1, no. 2, pp. 109–118, Jul. 2010, doi: 10.1109/T-AFFC.2010.12.

[315] I. Shumailov and H. Gunes, "Computational analysis of valence and arousal in virtual reality gaming using lower arm electromyograms," in *2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017*, 2018, vol. 2018-Janua, pp. 164–169, doi: 10.1109/ACII.2017.8273595.

[316] F. L. Luro and V. Sundstedt, "A Comparative Study of Eye Tracking and Hand Controller for Aiming Tasks in Virtual Reality," 2019, doi: 10.1145/3317956.3318153.

[317] O. H. M. Lutz, C. Burmeister, L. F. Dos Santos, N. Morkisch, C. Dohle, and J. Krüger, "Application of head-mounted devices with eye-tracking in virtual reality therapy," *Curr. Dir. Biomed. Eng.*, vol. 3, no. 1, pp. 53–56, Mar. 2017, doi: 10.1515/cdbme-2017-0012.

[318] J. Pinto, A. Fred, and H. P. Da Silva, "Biosignal-Based Multimodal Emotion Recognition in a Valence-Arousal Affective Framework Applied to Immersive Video Visualization," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2019, pp. 3577–3583, doi: 10.1109/EMBC.2019.8857852.

[319] J. Wagner, J. Kim, E. A.-2005 I. international conference, and undefined 2005, "From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification," *ieeexplore.ieee.org*, Accessed: Jan. 06, 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1521579/?casa_token=qqLI0Ot50NU AAAAA:o_if7JSg1pYE-9lWcXrygXCwBNL2dEgExu7Z1lRN-kSMsAWwARobMhJd3nZ2qlWMFyX8DeWsQSM.

[320] "The Lab." https://developerblog.myo.com/ (accessed Jun. 27, 2021).

[321] H. Gunes, M. Piccardi, and M. Pantic, "From the Lab to the Real World: Affect Recognition Using Multiple Cues and Modalities," in *Affective Computing*, 2008.

[322] J. Kim, "Bimodal Emotion Recognition using Speech and Physiological Changes," in *Robust Speech Recognition and Understanding*, 2007.

[323] O. Alzoubi, S. K. D'Mello, and R. A. Calvo, "Detecting naturalistic expressions of nonbasic affect using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 3, pp. 298–310, 2012, doi: 10.1109/T-AFFC.2012.4.

[324] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. Mcmahon, M. Sawey, and M. Schröder, "'FEELTRACE': AN INSTRUMENT FOR RECORDING PERCEIVED EMOTION IN REAL TIME." Accessed: Oct. 16, 2019. [Online]. Available: http://www.iscaaspeech.org/archive.

[325] G. N. Yannakakis and H. P. Martínez, "Grounding Truth via Ordinal Annotation."

Accessed: Feb. 18, 2020. [Online]. Available: http://semaine-db.eu/.

[326] R. Cowie, M. Sawey, C. Doherty, J. Jaimovich, C. Fyans, and P. Stapleton, "Gtrace: General trace program compatible with emotionML," in *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, 2013, pp. 709–710, doi: 10.1109/ACII.2013.126.

[327] D. Messinger, T. Cassel, … S. A.-J. of N., and undefined 2008, "Infant smiling dynamics and perceived positive emotion," *Springer*, Accessed: Jan. 04, 2021. [Online]. Available: https://link.springer.com/content/pdf/10.1007/s10919-008-0048-8.pdf.

[328] J. Healey, "Recording affect in the field: Towards methods and metrics for improving ground truth labels," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011, doi: 10.1007/978-3-642-24600-5_14.

[329] I. Mavridou *et al.*, "FACETEQ interface demo for emotion expression in VR," in *Proceedings - IEEE Virtual Reality*, 2017, pp. 441–442, doi: 10.1109/VR.2017.7892369.

[330] E. Maria, M. Ley, and H. Sten, "Emotion Recognition from Physiological Signal Analysis: A Review," *Electron. Notes Theor. Comput. Sci.*, vol. 343, pp. 35–55, 2019, doi: 10.1016/j.entcs.2019.04.009.

[331] B. G. Witmer and M. J. Singer, "Measuring Presence in Virtual Environments : A Presence," *Presence*, 1998, doi: 10.1026//1617-6383.15.2.69.

[332] M. Slater, "Presence and the sixth sense," *Presence: Teleoperators and Virtual Environments*, vol. 11, no. 4. pp. 435–439, Aug. 2002, doi: 10.1162/105474602760204327.

[333] S. Koelstra *et al.*, "DEAP: A Database for Emotion Analysis ;Using Physiological Signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012, doi: 10.1109/T-AFFC.2011.15.

[334] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*, 2013, doi: 10.1109/FG.2013.6553804.

[335] R. Schleicher and N. Galley, "Continuous Rating of Experienced Affect while Watching Emotional Film Clips."

[336] T. Suslow and U. S. Donges, "Alexithymia components are differentially related to explicit negative affect but not associated with explicit positive affect or implicit affectivity," *Front. Psychol.*, vol. 8, no. OCT, Oct. 2017, doi: 10.3389/fpsyg.2017.01758.

[337] P. McDonald, K. P.-P. Medicine, and undefined 1990, "The expression and perception of facial emotion in alexithymia: a pilot study.," *psycnet.apa.org*,

Accessed: Dec. 05, 2019. [Online]. Available: https://psycnet.apa.org/record/1990-25482-001.

[338] E. A. Butler, T. L. Lee, and J. J. Gross, "Does expressing your emotions raise or lower your blood pressure?: The answer depends on cultural context," *J. Cross. Cult. Psychol.*, vol. 40, no. 3, pp. 510–517, May 2009, doi: 10.1177/0022022109332845.

[339] D. Leising, T. Grande, R. F.-J. of R. in Personality, and undefined 2009, "The Toronto Alexithymia Scale (TAS-20): A measure of general psychological distress," *Elsevier*, Accessed: Oct. 16, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092656609000828.

[340] K. Geukes, S. Nestler, R. Hutteman, A. C. P. Küfner, and M. D. Back, "Running head: PERSONALITY AND WITHIN-PERSON VARIABILITY 1 Trait Personality and State Variability: Predicting Individual Differences in Within-and Cross-Context Fluctuations in Affect, Self-Evaluations, and Behavior in Everyday Life," *J. Res. Pers.*

[341] W. Revelle and K. R. Scherer, "Personality and Emotion."

[342] M. Pasupathi and J. Tsai, "Emotion and aging: Experience, expression, and control," *Artic. Psychol. Aging*, 1998, doi: 10.1037/0882-7974.12.4.590.

[343] H. Bonnemeier *et al.*, "Circadian profile of cardiac autonomic nervous modulation in healthy subjects: Differing effects of aging and gender on heart rate variability," *J. Cardiovasc. Electrophysiol.*, vol. 14, no. 8, pp. 791–799, Aug. 2003, doi: 10.1046/j.1540-8167.2003.03078.x.

[344] M. Almeida-Santos, … J. B.-F.-A. of gerontology, and undefined 2016, "Aging, heart rate variability and patterns of autonomic regulation of the heart," *Elsevier*, Accessed: Dec. 05, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167494315300881.

[345] D. L. Woods, J. M. Wyma, E. W. Yund, T. J. Herron, and B. Reed, "Factors influencing the latency of simple reaction time," *Front. Hum. Neurosci.*, vol. 9, no. MAR, Mar. 2015, doi: 10.3389/fnhum.2015.00131.

[346] U. Dimberg and L. O. Lundquist, "Gender differences in facial reactions to facial expressions," *Biol. Psychol.*, 1990, doi: 10.1016/0301-0511(90)90024-Q.

[347] A. M. Kring, A. H. Gordon, and M. Kring, "Sex Differences in Emotion: Expression, Experience, and Physiology," Gross & Mufioz, 1998.

[348] P. T. Costa and R. R. McCrae, "Stability and change in personality assessment: The Revised NEO Personality Inventory in the year 2000," *Journal of Personality Assessment*, vol. 68, no. 1. Routledge, pp. 86–94, 1997, doi: 10.1207/s15327752jpa6801_7.

[349] J. D. Henry and J. R. Crawford, "The short-form version of the Depression anxiety stress scales (DASS-21): Construct validity and normative data in a large non-clinical sample," *Br. J. Clin. Psychol.*, vol. 44, no. 2, pp. 227–239, Jun. 2005, doi:

10.1348/014466505X29657.

[350] L. Parkitny and J. McAuley, "The Depression Anxiety Stress Scale (DASS)," 2010. doi: 10.1016/S1836-9553(10)70030-8.

[351] D. Jolliffe, D. F.-J. of adolescence, and undefined 2006, "Development and validation of the Basic Empathy Scale," *Elsevier*, Accessed: Oct. 16, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140197105001090.

[352] "Unity Real-Time Development Platform | 3D, 2D VR & AR Visualizations." https://unity.com/ (accessed Oct. 15, 2019).

[353] W. Wang and Q. He, "A survey on emotional semantic image retrieval," in *Proceedings - International Conference on Image Processing, ICIP*, 2008, doi: 10.1109/ICIP.2008.4711705.

[354] J. Machajdik and A. Hanbury, "Affective Image Classification using Features Inspired by Psychology and Art Theory," *Int. Conf. Multimed.*, 2010, doi: 10.1145/1873951.1873965.

[355] B. H. Detenber and B. Reeves, "A Bio-Informational Theory of Emotion: Motion and Image Size Effects on Viewers," *J. Commun.*, 1996, doi: 10.1111/j.1460-2466.1996.tb01489.x.

[356] C. L. Larson, J. Aronoff, and J. J. Stearns, "The shape of threat: simple geometric forms evoke rapid and sustained capture of attention.," *Emotion*, 2007, doi: 10.1037/1528-3542.7.3.526.

[357] C. L. Larson, J. Aronoff, and E. L. Steuer, "Simple geometric shapes are implicitly associated with affective value," *Motiv. Emot.*, 2012, doi: 10.1007/s11031-011-9249-2.

[358] V. Yanulevskaya *et al.*, "In the Eye of the Beholder : Employing Statistical Analysis and Eye Tracking for Analyzing Abstract Paintings Categories and Subject Descriptors," *Proc. 20th ACM Int. Conf. Multimed.*, 2012, doi: 10.1145/2393347.2393399.

[359] X. Lu, P. Suryanarayan, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang, "On Shape and the Computability of Emotions," *Proc. 20th ACM Int. Conf. Multimed.*, 2012, doi: 10.1145/2393347.2393384.

[360] J. van Paasschen, E. Zamboni, F. Bacci, and Melcher, "Consistent Emotions Elicited by Low-Level Visual Features in Abstract Art," *Art Percept.*, 2014, doi: 10.1163/22134913-00002012.

[361] E. Geslin, L. Jégou, and D. Beaudoin, "How Color Properties Can Be Used to Elicit Emotions in Video Games," *Int. J. Comput. Games Technol.*, 2016, doi: 10.1155/2016/5182768.

[362] A. Skulmowski, A. Bunge, K. Kaspar, and G. Pipa, "Forced-choice decision-making in modified trolley dilemma situations: a virtual reality and eye tracking study," *Front. Behav. Neurosci.*, vol. 8, no. DEC, p. 426, Dec. 2014, doi:

10.3389/fnbeh.2014.00426.

[363]   A. T. Duchowski, V. Shivashankaraiah, T. Rawls, A. K. Gramopadhye, B. J. Melloy, and B. Kanki, "Binocular Eye Tracking in Virtual Reality for Inspection Training."

[364]   "VIVE Input Utility - Asset Store." https://assetstore.unity.com/packages/tools/integration/vive-input-utility-64219 (accessed Feb. 07, 2020).

[365]   S. Koelstra *et al.*, "DEAP: A Database for Emotion Analysis using Physiological Signals." Accessed: Feb. 07, 2020. [Online]. Available: http://www.eecs.qmul.ac.uk/mmv/datasets/deap/.

[366]   M. Soleymani and M. Pantic, "IEEE TRANS. AFFECTIVE COMPUTING 1 Multi-Modal Emotion Recognition in Response to Videos," 2011, doi: 10.1109/TAC.2011.37.

[367]   S. D'Mello and J. Kory, "Consistent but modest: A meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies," in *ICMI'12 - Proceedings of the ACM International Conference on Multimodal Interaction*, 2012, pp. 31–38, doi: 10.1145/2388676.2388686.

[368]   T. Fujimura, K. Katahira, and K. Okanoya, "Contextual modulation of physiological and psychological responses triggered by emotional stimuli," *Front. Psychol.*, vol. 4, no. MAY, p. 212, May 2013, doi: 10.3389/fpsyg.2013.00212.

[369]   S. Yang and G. Yang, "Emotion recognition of EMG based on improved L-M BP neural network and SVM," *J. Softw.*, vol. 6, no. 8, pp. 1529–1536, 2011, doi: 10.4304/jsw.6.8.1529-1536.

[370]   P. Kuppens *et al.*, "The Relation Between Valence and Arousal in Subjective Experience Varies With Personality and Culture," *J. Pers.*, vol. 85, 2017, doi: 10.1111/jopy.12258.

[371]   H. Lindman, *Analysis of Variance in Complex Experimental Designs*. 1974.

[372]   "SPSS Software | IBM." https://www.ibm.com/analytics/spss-statistics-software (accessed Oct. 30, 2019).

[373]   "MathWorks - Makers of MATLAB and Simulink - MATLAB & Simulink." https://uk.mathworks.com/ (accessed Oct. 30, 2019).

[374]   U. Narayanan, A. Unnikrishnan, V. Paul, and S. Joseph, "A survey on various supervised classification algorithms," in *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing, ICECDS 2017*, Jun. 2018, pp. 2118–2124, doi: 10.1109/ICECDS.2017.8389824.

[375]   N. Jadhav and R. Sugandhi, "Survey on Human Behavior Recognition Using Affective Computing," in *Proceedings - 2018 IEEE Global Conference on Wireless Computing and Networking, GCWCN 2018*, Mar. 2019, pp. 98–103, doi: 10.1109/GCWCN.2018.8668632.

[376]   C. Lisetti and K. Alvarez, "Emotion recognition from physiological signals for user

modeling of affect," 2014. Accessed: Jan. 30, 2020. [Online]. Available: https://www.researchgate.net/publication/229004773.

[377] S. Jerritta, M. Murugappan, K. Wan, and S. Yaacob, "Emotion recognition from facial EMG signals using higher order statistics and principal component analysis," *J. Chinese Inst. Eng.*, vol. 37, no. 3, pp. 385–394, 2014, doi: 10.1080/02533839.2013.799946.

[378] M. Asghari Oskoei, S. Member, H. Hu, and S. Member, "Support Vector Machine-Based Classification Scheme for Myoelectric Control Applied to Upper Limb," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 8, 2008, doi: 10.1109/TBME.2008.919734.

[379] M. Hamedi, S. H. Salleh, C. M. Ting, M. Astaraki, and A. M. Noor, "Robust Facial Expression Recognition for MuCI: A Comprehensive Neuromuscular Signal Analysis," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 102–115, Jan. 2018, doi: 10.1109/TAFFC.2016.2569098.

[380] M. Hamedi, I. M. Rezazadeh, and M. Firoozabadi, *Facial Gesture Recognition Using Two-Channel Bio-Sensors Configuration and Fuzzy Classifier: A Pilot Study*. .

[381] E. L. Van Den Broek, V. Lisý, J. H. Janssen, J. H. D. M. Westerink, M. H. Schut, and K. Tuinenbreijer, "Affective man-machine interface: Unveiling human emotions through biosignals," *Commun. Comput. Inf. Sci.*, vol. 52, pp. 21–47, 2010, doi: 10.1007/978-3-642-11721-3_2.

[382] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun, "Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2006, vol. 4105 LNCS, pp. 530–537, doi: 10.1007/11848035_70.

[383] E. H. Jang, B. J. Park, M. S. Park, S. H. Kim, and J. H. Sohn, "Analysis of physiological signals for recognition of boredom, pain, and surprise emotions," *J. Physiol. Anthropol.*, vol. 34, no. 1, Jun. 2015, doi: 10.1186/s40101-015-0063-5.

[384] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A Practical Guide to Support Vector Classification," 2003. Accessed: Oct. 30, 2019. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin.

[385] "LIBSVM -- A Library for Support Vector Machines." https://www.csie.ntu.edu.tw/~cjlin/libsvm/ (accessed Oct. 30, 2019).

[386] G. C. Cawley and N. L. C. Talbot, "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation," 2010.

[387] A. J. Fridlund and J. T. Cacioppo, "Guidelines for Human Electromyographic Research," *Psychophysiology*, vol. 23, no. 5, pp. 567–589, 1986, doi: 10.1111/j.1469-8986.1986.tb00676.x.

[388] T. Baumgartner, K. Lutz, C. F. Schmidt, and L. Jäncke, "The emotional power of music: How music enhances the feeling of affective pictures," *Brain Res.*, vol. 1075,

no. 1, pp. 151–164, Feb. 2006, doi: 10.1016/j.brainres.2005.12.065.

[389] M. Horvat, D. Kukolja, and D. Ivanec, "Comparing affective responses to standardized pictures and videos: A study report," in *38th International Convention MIPRO IEEE*, 2015, pp. 1394–1398, Accessed: Jun. 14, 2020. [Online]. Available: http://arxiv.org/abs/1505.07398.

[390] E. Siedlecka and T. F. Denson, "Experimental Methods for Inducing Basic Emotions: A Qualitative Review," *Emot. Rev.*, vol. 11, no. 1, pp. 87–97, Jan. 2019, doi: 10.1177/1754073917749016.

[391] C. Fernández, J. C. Pascual, J. Soler, M. Elices, M. J. Portella, and E. Fernández-Abascal, "Physiological responses induced by emotion-eliciting films," *Appl. Psychophysiol. Biofeedback*, vol. 37, no. 2, pp. 73–79, Jun. 2012, doi: 10.1007/s10484-012-9180-7.

[392] H. M. Peperkorn and A. MÜhlberger, "The impact of different perceptual cues on fear and presence in virtual reality," *Stud. Health Technol. Inform.*, vol. 191, pp. 75–79, 2013, doi: 10.3233/978-1-61499-282-0-75.

[393] A. Gorini, C. S. Capideville, G. De Leo, F. Mantovani, and G. Riva, "The role of immersion and narrative in mediated presence: the virtual hospital experience.," *Cyberpsychol. Behav. Soc. Netw.*, vol. 14, no. 3, pp. 99–105, Mar. 2011, doi: 10.1089/cyber.2010.0100.

[394] M. M. BRADLEY, B. N. CUTHBERT, and P. J. LANG, "Picture media and emotion: Effects of a sustained affective context," *Psychophysiology*, vol. 33, no. 6, pp. 662–670, Nov. 1996, doi: 10.1111/j.1469-8986.1996.tb02362.x.

[395] A. Dijksterhuis and P. K. Smith, "Affective Habituation: Subliminal Exposure to Extreme Stimuli Decreases Their Extremity," *psycnet.apa.org*, 2002, doi: 10.1037/1528-3542.2.3.203.

[396] M. M. Bradley, P. J. Lang, and B. N. Cuthbert, "Emotion, novelty, and the startle reflex: Habituation in humans," *Behav. Neurosci.*, vol. 107, no. 6, pp. 970–980, 1993, doi: 10.1037/0735-7044.107.6.970.

[397] A. C. Samson, S. D. Kreibig, B. Soderstrom, A. A. Wade, and J. J. Gross, "Eliciting positive, negative and mixed emotional states: A film library for affective scientists," *Cogn. Emot.*, vol. 30, no. 5, pp. 827–856, Jul. 2016, doi: 10.1080/02699931.2015.1031089.

[398] T. Brosch, D. Sander, and K. Scherer, "That Baby Caught My Eye... Attention Capture by Infant Faces Emotion regulation View project Brain Activity underlying Social Emotion Imitation and Regulation View project," 2007, doi: 10.1037/1528-3542.7.3.685.

[399] A. C. Samson, S. D. Kreibig, B. Soderstrom, A. A. Wade, and J. J. Gross, "Eliciting positive, negative and mixed emotional states: A film library for affective scientists," *Cogn. Emot.*, 2015, doi: 10.1080/02699931.2015.1031089.

[400] T. A. Ito, J. T. Larsen, N. K. Smith, and J. T. Cacioppo, "Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations.," *J. Pers. Soc. Psychol.*, vol. 75, no. 4, pp. 887–900, 1998, doi: 10.1037/0022-3514.75.4.887.

[401] Á. Szőllősi and M. Racsmány, "Enhanced mnemonic discrimination for emotional memories: the role of arousal in interference resolution," *Mem. Cogn.*, 2020, doi: 10.3758/s13421-020-01035-3.

[402] M. K. Doss, J. K. Picart, and D. A. Gallo, "Creating Emotional False Recollections: Perceptual Recombination and Conceptual Fluency Mechanisms," *Emotion*, 2019, doi: 10.1037/emo0000590.

[403] P. Ekman, R. W. Levenson, and W. V Friesen, "Autonomic nervous system activity distinguishes among emotions.," *Science*, 1983, doi: 10.1126/science.6612338.

[404] M. M. Bradley and P. J. Lang, "Affective reactions to acoustic stimuli," *Psychophysiology*, vol. 37, no. 2, pp. 204–215, Mar. 2000, doi: 10.1111/1469-8986.3720204.

[405] B. N. Lang, P.J., Bradley, M.M., & Cuthbert, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8.," *Tech. Rep. A-8.*, 2008, doi: 10.1016/j.epsr.2006.03.016.

[406] N. K. Smith, J. T. Cacioppo, J. T. Larsen, and T. L. Chartrand, "May I have your attention, please: Electrocortical responses to positive and negative stimuli," *Neuropsychologia*, vol. 41, no. 2, pp. 171–183, 2003, doi: 10.1016/S0028-3932(02)00147-1.

[407] S. Ribaric and I. Fratric, "Experimental Evaluation of Matching-Score Normalization Techniques on Different Multimodal Biometric Systems." Accessed: Feb. 18, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1653147/.

[408] C.-C. Acm Reference Format: Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol*, vol. 2, no. 3, p. 27, Apr. 2011, doi: 10.1145/1961189.1961199.

[409] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun, "Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2006, vol. 4105 LNCS, pp. 530–537, doi: 10.1007/11848035_70.

[410] M. K. Bailey, C. J. Smaldone, E. Chabot, and Y. Sun, "Impact of Placement of Facial PPG Sensor on Pulse-Rate Monitoring Accuracy." Accessed: Jun. 21, 2020. [Online]. Available: http://pulsesensor.com/pages/open-hardware.

[411] F. Shaffer, R. McCraty, and C. L. Zerr, "A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability," *Front. Psychol.*, vol. 5, no. September, p. 1040, 2014, doi: 10.3389/fpsyg.2014.01040.

[412] L. Nilsson, T. Goscinski, S. Kalman, L. G. Lindberg, and A. Johansson, "Combined photoplethysmographic monitoring of respiration rate and pulse: A comparison between different measurement sites in spontaneously breathing subjects," *Acta Anaesthesiol. Scand.*, vol. 51, no. 9, pp. 1250–1257, 2007, doi: 10.1111/j.1399-6576.2007.01375.x.

[413] N. D. Giardino, P. M. Lehrer, and R. Edelberg, "Comparison of finger plethysmograph to ECG in the measurement of heart rate variability," *Psychophysiology*, vol. 39, no. 2, pp. 246–253, Mar. 2002, doi: 10.1111/1469-8986.3920246.

[414] H. Fukushima, … H. K.-… C. of the, and undefined 2012, "Estimating heart rate using wrist-type photoplethysmography and acceleration sensor while running," *ieeexplore.ieee.org*, Accessed: Nov. 23, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6346570/.

[415] W. Johnston and Y. Mendelson, "Extracting heart rate variability from a wearable reflectance pulse oximeter," in *Bioengineering, Proceedings of the Northeast Conference*, 2005, pp. 157–158, doi: 10.1109/nebc.2005.1431971.

[416] N. Selvaraj, A. Jaryal, J. Santhosh, K. K. Deepak, and S. Anand, "Assessment of heart rate variability derived from finger-tip photoplethysmography as compared to electrocardiography," *J. Med. Eng. Technol.*, vol. 32, no. 6, pp. 479–484, Nov. 2008, doi: 10.1080/03091900701781317.

[417] Y. Mendelson, … C. P.-C. of the I. E. in, and undefined 2003, "Measurement site and photodetector size considerations in optimizing power consumption of a wearable reflectance pulse oximeter," *ieeexplore.ieee.org*, Accessed: Feb. 18, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1280775/.

[418] T. Tamura, Y. Maeda, M. Sekine, and M. Yoshida, "Wearable Photoplethysmographic Sensors-Past and Present," *mdpi.com*, vol. 3, pp. 282–302, 2014, doi: 10.3390/electronics3020282.

[419] Y. Y. Kim, E. N. Kim, M. J. Park, K. S. Park, H. D. Ko, and H. T. Kim, "The application of biosignal feedback for reducing cybersickness from exposure to a virtual environment," *Presence Teleoperators Virtual Environ.*, vol. 17, no. 1, pp. 1–16, Feb. 2008, doi: 10.1162/pres.17.1.1.

[420] S. Otsuka, K. Kurosaki, M. O.-T. 2017-2017 IEEE, and undefined 2017, "Physiological measurements on a gaming virtual reality headset using photoplethysmography: A preliminary attempt at incorporating physiological measurement with," *ieeexplore.ieee.org*, Accessed: Feb. 18, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8228049/.

[421] G. Bernal and P. Maes, "Emotional beasts: Visually expressing emotions through avatars in VR," in *Conference on Human Factors in Computing Systems - Proceedings*, May 2017, vol. Part F127655, pp. 2395–2402, doi:

10.1145/3027063.3053207.

[422] G. Geethavani, M. Rameswarudu, and R. R. Reddy, "Effect of Caffeine on Heart Rate and Blood Pressure," *Int. J. Sci. Res. Publ.*, vol. 4, no. 2, 2014, Accessed: Jun. 21, 2020. [Online]. Available: www.ijsrp.org.

[423] V. Putz-Anderson, J. V Setzer, and J. S. Croxton, "Effects of alcohol, caffeine and methyl chloride on man.," *Psychol. Rep.*, vol. 48, no. 3, pp. 715–725, Jun. 1981, doi: 10.2466/pr0.1981.48.3.715.

[424] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, 1997, doi: 10.1016/S0031-3203(96)00142-2.

[425] T. Kostoulas, G. Chanel, M. Muszynski, P. Lombardo, and T. Pun, "Dynamic time warping of multimodal signals for detecting highlights in movies," in *INTERPERSONAL 2015 - Proceedings of the 1st ACM Workshop on Modeling INTERPERsonal SynchrONy And infLuence, co-located with ICMI 2015*, Nov. 2015, pp. 35–40, doi: 10.1145/2823513.2823515.

[426] Y. T. Zhang, Y. L. Zheng, W. H. Lin, H. Y. Zhang, and X. L. Zhou, "Challenges and opportunities in cardiovascular health informatics," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 3, pp. 633–642, 2013, doi: 10.1109/TBME.2013.2244892.

[427] G. Borges and V. Brusamarello, "Sensor fusion methods for reducing false alarms in heart rate monitoring," *J. Clin. Monit. Comput.*, vol. 30, no. 6, pp. 859–867, Dec. 2016, doi: 10.1007/s10877-015-9786-4.

[428] J. J. Gross and R. W. Levenson, "Emotion Elicitation using Films," *Cogn. Emot.*, vol. 9, no. 1, pp. 87–108, Jan. 1995, doi: 10.1080/02699939508408966.

[429] B. Detenber, R. Simons, and G. Bennett, "Roll ' em!: The Effects of Picture Motion on Emotional Responses," *J. Broadcast. Electron. Media*, 1998, doi: 10.1080/08838159809364437.

[430] B. L. Sherman and J. R. Dominick, "Perceptions of Colorization," *Journal. Mass Commun. Q.*, vol. 65, no. 4, pp. 976–980, 1988, doi: 10.1177/107769908806500422.

[431] M. K. Uhrig *et al.*, "Emotion Elicitation: A Comparison of Pictures and Films," *Front. Psychol.*, vol. 7, Feb. 2016, doi: 10.3389/fpsyg.2016.00180.

[432] C. Mühl, E. L. Van Den Broek, F. Nijboer, A.-M. Brouwer, N. Van Wouwe, and D. Heylen, "Multi-modal Affect Induction for Affective Brain-Computer Interfaces Dopamine selectively modulates the outcome of unnatural action-valence associations View project Locked-in syndrome View project Multi-modal Affect Induction for Affective Brain-Computer Interfaces," 2011, doi: 10.1007/978-3-642-24600-5_27.

[433] R. M. Baños, C. Botella, M. Alcañiz, V. Liaño, B. Guerrero, and B. Rey, "Immersion and emotion: Their impact on the sense of presence," *Cyberpsychology Behav.*, vol. 7, no. 6, pp. 734–741, 2004, doi: 10.1089/cpb.2004.7.734.

[434] M. Slater, V. Linakis, M. Usoh, and R. Kooper, "Immersion , Presence , and Performance in Virtual Environments : An Experiment with Tri-Dimensional Chess," *Virtual Real.*, 1995, doi: 10.1.1.34.6594.

[435] M. Meehan *et al.*, "Physiological measures of presence in stressful virtual environments," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 645–652, Jul. 2002, doi: 10.1145/566654.566630.

[436] "Free Software for Students & Educators | Maya | Autodesk." https://www.autodesk.com/education/free-software/maya (accessed Jan. 03, 2020).

[437] M. White, A. Smith, K. Humphryes, S. Pahl, D. Snelling, and M. Depledge, "Blue space: The importance of water for preference, affect, and restorativeness ratings of natural and built scenes," *J. Environ. Psychol.*, vol. 30, no. 4, pp. 482–493, Dec. 2010, doi: 10.1016/j.jenvp.2010.04.004.

[438] M. V. Thoma, R. Mewes, and U. M. Nater, "Preliminary evidence: The stress-reducing effect of listening to water sounds depends on somatic complaints," *Med. (United States)*, vol. 97, no. 8, Feb. 2018, doi: 10.1097/MD.0000000000009851.

[439] D. Cracknell, M. P. White, S. Pahl, W. J. Nichols, and M. H. Depledge, "Marine Biota and Psychological Well-Being: A Preliminary Examination of Dose-Response Effects in an Aquarium Setting.," *Environ. Behav.*, vol. 48, no. 10, pp. 1242–1269, Dec. 2016, doi: 10.1177/0013916515597512.

[440] H. Noguchi and T. Sakaguchi, "Effect of Illuminance and Color Temperature on Lowering of Physiological Activity," 1999.

[441] "CAPCOM : RESIDENT EVIL 7 biohazard." http://residentevil7.com/uk/#_top (accessed Jan. 03, 2020).

[442] "AFFECTED - The Manor on Gear VR | Oculus." https://www.oculus.com/experiences/gear-vr/1058194390920956/ (accessed Jan. 03, 2020).

[443] "Freesound - Freesound." https://freesound.org/ (accessed Jan. 28, 2020).

[444] "Unity - Scripting API: Physics." https://docs.unity3d.com/ScriptReference/Physics.html (accessed Feb. 03, 2020).

[445] T. Sharot and E. A. Phelps, "How arousal modulates memory: Disentangling the effects of attention and retention," *Cogn. Affect. Behav. Neurosci.*, vol. 4, no. 3, pp. 294–306, 2004, doi: 10.3758/CABN.4.3.294.

[446] E. A. Kensinger, "Remembering the Details: Effects of Emotion.," *Emot. Rev.*, vol. 1, no. 2, pp. 99–113, 2009, doi: 10.1177/1754073908100432.

[447] N. Klein Selle *et al.*, "Memory detection: The effects of emotional stimuli Memory detection: The effects of emotional stimuli A R T I C L E I N F O," *Artic. Biol. Psychol.*, 2017, doi: 10.1016/j.biopsycho.2017.07.021.

[448] M. M. Bradley, M. K. Greenwald, M. C. Petry, and P. J. Lang, "Remembering Pictures: Pleasure and Arousal in Memory," *J. Exp. Psychol. Learn. Mem. Cogn.*,

vol. 18, no. 2, pp. 379–390, 1992, doi: 10.1037/0278-7393.18.2.379.

[449] igroup, "Presence Questionnaire 1 | igroup.org – project consortium," *igroup.org*, 2016. http://igroup.org/projects/ipq/ (accessed Jan. 06, 2020).

[450] R. R. McCrae and P. T. Costa, "Personality trait structure as a human universal.," *Am. Psychol.*, 1997, doi: 10.1037/0003-066X.52.5.509.

[451] Qualtrics, "Qualtrics." Qualtrics, Provo, Utah, USA, 2019, [Online]. Available: https://www.qualtrics.com.

[452] C. Rosakranse and S. Y. Oh, "Presence Questionaire Use Trends Measuring Presence: The Use Trends of Five Canonical Presence Questionaires from 1998-2012," *Int. Soc. Presence Res.*, pp. 25–30, 2014.

[453] N. Morina, W. P. Brinkman, D. Hartanto, and P. M. G. Emmelkamp, "Sense of presence and anxiety during virtual social interactions between a human and virtual humans," *PeerJ*, vol. 2014, no. 1, 2014, doi: 10.7717/peerj.337.

[454] T. Schubert, F. Friedmann, and H. Regenbrecht, "The experience of presence: Factor analytic insights," *Presence Teleoperators Virtual Environ.*, vol. 10, no. 3, pp. 266–281, Jun. 2001, doi: 10.1162/105474601300343603.

[455] M. Slater and M. Usoh, "Representations Systems, Perceptual Position, and Presence in Immersive Virtual Environments," *Presence Teleoperators Virtual Environ.*, vol. 2, no. 3, pp. 221–233, Jan. 1993, doi: 10.1162/pres.1993.2.3.221.

[456] C. M. Hendrix, "Exploratory Studies on the Sense of Presence in Virtual Environments as a Function of Visual and Auditory Display Parameters."

[457] A. Carlin, H. Hoffman, S. W.-B. research and therapy, and undefined 1997, "Virtual reality and tactile augmentation in the treatment of spider phobia: a case report," *Elsevier*, Accessed: Jan. 06, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000579679600085X.

[458] P. Kuppens *et al.*, "The Relation Between Valence and Arousal in Subjective Experience Varies With Personality and Culture," *J. Pers.*, vol. 85, no. 4, pp. 530–542, 2017, doi: 10.1111/jopy.12258.

[459] W. G. Graziano and N. Eisenberg, "AGREEABLENESS: A DIMENSION OF PERSONALITY," 1997.

[460] R. E. Lucas, K. Le, and P. S. Dyrenforth, "Explaining the extraversion/positive affect relation: Sociability cannot account for extraverts' greater happiness," *J. Pers.*, vol. 76, no. 3, pp. 385–414, Jun. 2008, doi: 10.1111/j.1467-6494.2008.00490.x.

[461] L. D. Smillie, A. J. Cooper, J. Wilt, and W. Revelle, "Do extraverts get more bang for the buck? Refining the affective-reactivity hypothesis of extraversion," *J. Pers. Soc. Psychol.*, vol. 103, no. 2, pp. 306–326, 2012, doi: 10.1037/a0028372.

[462] R. Trnka, K. Balcar, M. Kuška, and K. Hnilica, "Neuroticism and Valence of Negative Emotional Concepts Emotions, Creativity, and Emotional Creativity: The Link between Cognition and Emotion View project a lecture View project," *Artic.*

Soc. Behav. Personal. An Int. J., 2012, doi: 10.2224/sbp.2012.40.5.843.

[463]   G. Riva *et al.*, "Affective Interactions Using Virtual Reality: The Link between Presence and Emotions," *CYBERPSYCHOLOGY Behav.*, vol. 10, no. 1, 2007, doi: 10.1089/cpb.2006.9993.

[464]   K. Kim, M. Z. Rosenthal, D. J. Zielinski, and R. Brady, "Effects of virtual environment platforms on emotional responses," *Comput. Methods Programs Biomed.*, vol. 113, no. 3, pp. 882–893, Mar. 2014, doi: 10.1016/j.cmpb.2013.12.024.

[465]   "VIVE™ | VIVE Virtual Reality System." https://www.vive.com/us/product/vive-virtual-reality-system/ (accessed Oct. 15, 2019).

[466]   A. L. Simeone, I. Mavridou, and W. Powell, "Altering User Movement Behaviour in Virtual Environments," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 4, 2017, doi: 10.1109/TVCG.2017.2657038.

[467]   "VIVE™ | Vive Deluxe Audio Strap." https://www.vive.com/us/vive-deluxe-audio-strap/ (accessed Jan. 16, 2020).

[468]   S. Bouchard, G. Robillard, J. St-Jacques, S. Dumoulin, M. J. Patry, and P. Renaud, "Reliability and validity of a single-item measure of presence in VR," in *Proceedings - 3rd IEEE International Workshop on Haptic, Audio and Visual Environments and their Applications - HAVE 2004*, 2004, pp. 59–61, doi: 10.1109/have.2004.1391882.

[469]   W. Sato, T. Fujimura, and N. Suzuki, "Enhanced facial EMG activity in response to dynamic facial expressions," *Int. J. Psychophysiol.*, vol. 70, no. 1, pp. 70–74, Oct. 2008, doi: 10.1016/j.ijpsycho.2008.06.001.

[470]   "Find Peaks in Data - MATLAB & Simulink - MathWorks United Kingdom." https://uk.mathworks.com/help/signal/ug/find-peaks-in-data.html (accessed Jan. 13, 2020).

[471]   M. Vollmer, "A robust, simple and reliable measure of heart rate variability using relative RR intervals," in *Computing in Cardiology*, Feb. 2015, vol. 42, pp. 609–612, doi: 10.1109/CIC.2015.7410984.

[472]   J. Collins, H. Regenbrecht, T. Langlotz, Y. Said Can, C. Ersoy, and R. Butson, "Measuring Cognitive Load and Insight: A Methodology Exemplified in a Virtual Reality Learning Context," in *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Oct. 2019, pp. 351–362, doi: 10.1109/ISMAR.2019.00033.

[473]   E. Ščiglinskas, A. Mačiukas, A. Vidugirienė, and T. Krilavičius, "Investigation on the Dependencies Between HRV, Physical Training, and Focus of Attention in Virtual Environment," 2019, pp. 529–540.

[474]   D. D. Heikoop, J. C. F. de Winter, B. van Arem, and N. A. Stanton, "Effects of platooning on signal-detection performance, workload, and stress: A driving simulator study," *Appl. Ergon.*, vol. 60, pp. 116–127, Apr. 2017, doi: 10.1016/j.apergo.2016.10.016.

[475] R. T. Wilkinson and S. Allison, "Age and simple reaction time: Decade differences for 5,325 subjects," *Journals Gerontol.*, vol. 44, no. 2, 1989, doi: 10.1093/geronj/44.2.p29.

[476] A. S. Vincent, T. Roebuck-Spencer, K. Gilliland, and R. Schlegel, "Automated Neuropsychological Assessment Metrics (v4) Traumatic Brain Injury Battery: Military Normative Data," *Mil. Med.*, vol. 177, no. 3, pp. 256–269, Mar. 2012, doi: 10.7205/milmed-d-11-00289.

[477] D. L. Woods, J. M. Wyma, E. W. Yund, T. J. Herron, and B. Reed, "Factors influencing the latency of simple reaction time," *Front. Hum. Neurosci.*, vol. 9, Mar. 2015, doi: 10.3389/fnhum.2015.00131.

[478] G. EAPA Batista and D. Furtado Silva, "How k-Nearest Neighbor Parameters Affect its Performance." Accessed: Feb. 07, 2021. [Online]. Available: http://www.icmc.usp.br/.

[479] "Train multiclass naive Bayes model - MATLAB fitcnb." https://www.mathworks.com/help/stats/fitcnb.html (accessed Feb. 07, 2021).

[480] "Train support vector machine (SVM) classifier for one-class and binary classification - MATLAB fitcsvm." https://www.mathworks.com/help/stats/fitcsvm.html (accessed Feb. 07, 2021).

[481] "Fit k-nearest neighbor classifier - MATLAB fitcknn." https://www.mathworks.com/help/stats/fitcknn.html (accessed Feb. 07, 2021).

[482] I. I. a. Essa and A. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 757–763, 1997, doi: 10.1109/34.598232.

[483] D. Marzoli, "Sun-induced frowning fosters aggressive feelings Human Neurophysiology View project," doi: 10.1080/02699931.2013.801338.

[484] A. V. B.-P. of measuring behavior and undefined 2010, "Facial EMG as a tool for inferring affective states," *academia.edu*, Accessed: Aug. 27, 2019. [Online]. Available: https://www.academia.edu/download/6006600/MB2010_Proceedings_web.pdf#page=116.

[485] M. Jiang, A. Rahmani, … T. W.-… on C. and, and U. 2015, "Facial expression recognition with sEMG method," *ieeexplore.ieee.org*, Accessed: Feb. 26, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7363189/.

[486] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinform. Comput. Biol.*, vol. 3, no. 2, pp. 185–205, Apr. 2005, doi: 10.1142/S0219720005001004.

[487] J. Harju, A. Tarniceriu, J. Parak, and al -, "Physiological Measurement Related content Monitoring of heart rate and inter-beat intervals with wrist plethysmography in patients with atrial fibrillation," *Physiol. Meas*, vol. 39, p. 84001, 2018, doi:

10.1088/1361-6579/aad2c0.

[488] C. Bishop, M., *Pattern recognition and machine learning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

[489] J. Liu, H. Meng, A. Nandi, M. L.-2016 12th International, and undefined 2016, "Emotion detection from EEG recordings," *ieeexplore.ieee.org*, Accessed: Feb. 04, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7603437/.

[490] S. Berretti *et al.*, "3D facial expression recognition using SIFT descriptors of automatically detected keypoints," vol. 27, no. 11, pp. 1021–1036, 2011, doi: 10.1007/s00371-011-0611-xï.

[491] a J. Fridlund, G. E. Schwartz, and S. C. Fowler, "Pattern recognition of self-reported emotional state from multiple-site facial EMG activity during affective imagery.," *Psychophysiology*, vol. 21, no. 6, pp. 622–637, 1984, doi: 10.1111/j.1469-8986.1984.tb00249.x.

[492] C. V. O. Witvliet and S. R. Vrana, "Play it again Sam: Repeated exposure to emotionally evocative music polarises liking and smiling responses, and influences other affective reports, facial EMG, and heart rate," *Cogn. Emot.*, vol. 21, no. 1, pp. 3–25, 2007, doi: 10.1080/02699930601000672.

[493] J. Leite, S. Carvalho, S. Galdo-Alvarez, J. Alves, A. Sampaio, and Ó. F. Gonçalves, "Affective picture modulation: Valence, arousal, attention allocation and motivational significance," *Int. J. Psychophysiol.*, vol. 83, no. 3, pp. 375–381, Mar. 2012, doi: 10.1016/j.ijpsycho.2011.12.005.

[494] W. Wu and J. Lee, "Improvement of HRV methodology for positive/negative emotion assessment," 2009, doi: 10.4108/ICST.COLLABORATECOM2009.8296.

[495] S. Rezaei, S. Moharreri, N. J. Dabanloo, and S. Parvaneh, "Evaluating Valence Level of Pictures Stimuli in Heart Rate Variability Response."

[496] D. Wu *et al.*, "Optimal Arousal Identification and Classification for Affective Computing Using Physiological Signals : Virtual Reality Stroop Task," doi: 10.1109/T-AFFC.2010.12.

[497] K. T. Sweeney, S. Member, T. E. Ward, S. Member, and S. F. McLoone, "Artifact Removal in Physiological Signals-Practices and Possibilities," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 3, 2012, doi: 10.1109/TITB.2012.2188536.

[498] K. Harry Shelley *et al.*, "Can Photoplethysmography Variability Serve as an Alternative Approach to Obtain Heart Rate Variability Information?," *Artic. J. Clin. Monit. Comput.*, 2008, doi: 10.1007/s10877-007-9103-y.

[499] A. Alqaraawi, A. Alwosheel, and A. Alasaad, "Heart rate variability estimation in photoplethysmography signals using Bayesian learning approach," *Healthc. Technol. Lett.*, vol. 3, no. 2, pp. 136–142, 2016, doi: 10.1049/htl.2016.0006.

[500] P. V Rouast, S. Member, M. T. P Adam, R. Chiong, and S. Member, "Deep Learning for Human Affect Recognition: Insights and New Developments," vol. 14, no. 8,

2015, doi: 10.1109/TAFFC.2018.2890471.

[501] I. Behoora and C. S. Tucker, "Machine learning classification of design team members' body language patterns for real time emotional state detection," *Des. Stud.*, vol. 39, pp. 100–127, Jul. 2015, doi: 10.1016/j.destud.2015.04.003.

[502] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2067–2083, 2008, doi: 10.1109/TPAMI.2008.26.

[503] J. Heinström, "Negative affectivity – the emotional dimension," in *From Fear to Flow*, Elsevier, 2010, pp. 75–103.

[504] J. T. Cacioppo, J. T. Larsen, N. K. Smith, and G. G. Bertson, "The affect system: what lurks below the surface of feelings?," *Feel. Emot. Amsterdam Conf.*, pp. 223–242, 2004, Accessed: Jan. 29, 2020. [Online]. Available: https://psycnet.apa.org/record/2004-95227-014.

[505] J. H. T. Viitasalo and P. V. Komi, "Signal characteristics of EMG during fatigue," *Eur. J. Appl. Physiol. Occup. Physiol.*, vol. 37, no. 2, pp. 111–121, Jun. 1977, doi: 10.1007/BF00421697.

[506] P. Kuppens, F. Tuerlinckx, J. A. Russell, and L. F. Barrett, "The relation between valence and arousal in subjective experience," *Psychol. Bull.*, vol. 139, no. 4, pp. 917–940, 2013, doi: 10.1037/a0030811.

[507] K. A. Lindquist, T. D. Wager, H. Kober, E. Bliss-Moreau, and L. F. Barrett, "The brain basis of emotion: A meta-analytic review," *Behavioral and Brain Sciences*, vol. 35, no. 3. NIH Public Access, pp. 121–143, Jun. 2012, doi: 10.1017/S0140525X11000446.

[508] P. Ekman, "Cross-cultural studies of facial expression," in *Darwin and Facial Expression A Century of Research in Review*, 1973.

[509] P. Scarfe and A. Glennerster, "Using high-fidelity virtual reality to study perception in freely moving observers," *J. Vis.*, vol. 15, no. 9, pp. 3–3, Jul. 2015, doi: 10.1167/15.9.3.

[510] H. Veling, R. W. Holland, and A. van Knippenberg, "When approach motivation and behavioral inhibition collide: Behavior regulation through stimulus devaluation," *J. Exp. Soc. Psychol.*, vol. 44, no. 4, pp. 1013–1019, Jul. 2008, doi: 10.1016/j.jesp.2008.03.004.

[511] A. E. Ferrey, T. J. Burleigh, and M. J. Fenske, "Stimulus-category competition, inhibition, and affective devaluation: a novel account of the uncanny valley," *Front. Psychol.*, vol. 6, no. MAR, p. 249, Mar. 2015, doi: 10.3389/fpsyg.2015.00249.

[512] J. T. Lanzetta and B. G. Englis, "Expectations of Cooperation and Competition and Their Effects on Observers' Vicarious Emotional Responses," *J. Pers. Soc. Psychol.*, vol. 56, no. 4, pp. 543–554, 1989, doi: 10.1037/0022-3514.56.4.543.

[513] P. L. Lockwood, "The anatomy of empathy: Vicarious experience and disorders of

social cognition," *Behavioural Brain Research*, vol. 311. Elsevier B.V., pp. 255–266, Sep. 15, 2016, doi: 10.1016/j.bbr.2016.05.048.

[514] M. Hamedi, S. H. Salleh, M. Astaraki, and A. M. Noor, "EMG-based facial gesture recognition through versatile elliptic basis function neural network," *Biomed. Eng. Online*, vol. 12, no. 1, p. 73, Jul. 2013, doi: 10.1186/1475-925X-12-73.

[515] M. Perusquía-Hernández, M. Hirokawa, and K. Suzuki, "A Wearable Device for Fast and Subtle Spontaneous Smile Recognition," *IEEE Trans. Affect. Comput.*, vol. 8, no. 4, pp. 522–533, Oct. 2017, doi: 10.1109/TAFFC.2017.2755040.

[516] J. Lou *et al.*, "Realistic Facial Expression Reconstruction for VR HMD Users," *IEEE Trans. Multimed.*, vol. 22, no. 3, pp. 730–743, Mar. 2020, doi: 10.1109/TMM.2019.2933338.

[517] M. L. Munoz *et al.*, "Validity of (Ultra-)Short recordings for heart rate variability measurements," *PLoS One*, vol. 10, no. 9, Sep. 2015, doi: 10.1371/journal.pone.0138921.

[518] U. Nussinovitch *et al.*, "Reliability of ultra-short ECG indices for heart rate variability," *Ann. Noninvasive Electrocardiol.*, vol. 16, no. 2, pp. 117–122, Apr. 2011, doi: 10.1111/j.1542-474X.2011.00417.x.

[519] M. E. van den Berg *et al.*, "Normal values of corrected heart-rate variability in 10-second electrocardiograms for all ages," *Front. Physiol.*, vol. 9, no. APR, Apr. 2018, doi: 10.3389/fphys.2018.00424.

[520] F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," *Front. Public Heal.*, vol. 5, Sep. 2017, doi: 10.3389/fpubh.2017.00258.

[521] T. Thong, K. Li, J. McNames, M. Aboy, and B. Goldstein, "Accuracy of Ultra-Short Heart Rate Variability Measures," in *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, 2003, vol. 3, pp. 2424–2427, doi: 10.1109/iembs.2003.1280405.

[522] O. AlZoubi, D. Fossati, S. D'Mello, and R. A. Calvo, "Affect detection from non-stationary physiological data using ensemble classifiers," *Evol. Syst.*, vol. 6, no. 2, pp. 79–92, Jun. 2015, doi: 10.1007/s12530-014-9123-z.

[523] M. Ranzato, F. J. Huang, Y. L. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007, doi: 10.1109/CVPR.2007.383157.

[524] N. Cummins, A. Baird, and B. W. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*, vol. 151. Academic Press Inc., pp. 41–54, Dec. 01, 2018, doi: 10.1016/j.ymeth.2018.07.007.

[525] Z. Yin, M. Zhao, Y. Wang, J. Yang, and J. Zhang, "Recognition of emotions using multimodal physiological signals and an ensemble deep learning model," *Comput. Methods Programs Biomed.*, vol. 140, pp. 93–110, Mar. 2017, doi:

10.1016/j.cmpb.2016.12.005.

[526] J. Lin, S. Pan, C. S. Lee, and S. Oviatt, "An explainable deep fusion network for affect recognition using physiological signals," in *International Conference on Information and Knowledge Management, Proceedings*, Nov. 2019, pp. 2069–2072, doi: 10.1145/3357384.3358160.

[527] M. D. Hssayeni and B. Ghoraani, "Multi-modal Physiological Data Fusion for Affect Estimation Using Deep Learning," *IEEE Access*, pp. 1–1, Feb. 2021, doi: 10.1109/access.2021.3055933.

[528] H. Martinez, … Y. B.-I. C., and undefined 2013, "Learning deep physiological models of affect," *ieeexplore.ieee.org*, Accessed: Feb. 18, 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6496209/?casa_token=4rVcH4DC8A IAAAAA:6ytX7VbDem9W5UdkdVJmnlOONzYQJs1YOIr0ZX-O4eoSxu3-gyAHz89p5RuLsmxr6fgVscjddwWO.

[529] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A dataset for affect, personality and mood research on individuals and groups," *arXiv*. 2017, Accessed: Feb. 18, 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8554112/?casa_token=WnZbZUvIepI AAAAA:MlCn_xmRvq_OHrTKoOEgwX5f5KYI0SgEsKfHOiXtpCV4NZq1W73-scJFH7ytRW7Jbg6LU2K5sbZn.

[530] C. Li, Z. Bao, L. Li, and Z. Zhao, "Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition," *Inf. Process. Manag.*, vol. 57, no. 3, p. 102185, May 2020, doi: 10.1016/j.ipm.2019.102185.

[531] S. Tripathi, S. Acharya, R. Sharma, … S. M.-P. of the, and undefined 2017, "Using deep and convolutional neural networks for accurate emotion classification on DEAP dataset," *dl.acm.org*, Accessed: Feb. 18, 2021. [Online]. Available: https://dl.acm.org/doi/abs/10.5555/3297863.3297883.

[532] O. Alzoubi, B. Almakhadmeh, S. K. Tawalbeh, M. Bani Yassien, and I. Hmeidi, "A Deep Learning Approach for Classifying Emotions from Physiological Data," in *2020 11th International Conference on Information and Communication Systems, ICICS 2020*, 2020, pp. 214–219, doi: 10.1109/ICICS49469.2020.239534.

[533] O. Bălan, G. Moise, A. Moldoveanu, M. Leordeanu, and F. Moldoveanu, "An investigation of various machine and deep learning techniques applied in automatic fear level detection and acrophobia virtual therapy," *Sensors (Switzerland)*, vol. 20, no. 2, 2020, doi: 10.3390/s20020496.

[534] D. Quesnel, S. DiPaola, and B. E. Riecke, "Deep learning for classification of peak emotions within virtual reality systems," in *International series on information systems and management in creative eMedia*, 2017, vol. 2017, no. 2, pp. 6–11, Accessed: Feb. 18, 2021. [Online]. Available:

www.ambientmediaassociation.org/Journal.

[535] M. Slater, "Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments," in *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 2009, doi: 10.1098/rstb.2009.0138.

[536] M. Valstar, Z. Ambadar, J. F. Cohn, M. F. Valstar, and M. Pantic, "Spontaneous vs. posed facial behavior: Automatic analysis of brow actions 3D Facial Expression Understanding View project Multimodal affective computing for automated depression analysis View project Spontaneous vs. Posed Facial Behavior: Automatic Analysis of Brow Actions," 2006, doi: 10.1145/1180995.1181031.

[537] U. Hess, A. Kappas, G. J. Mchugo, R. E. Kleck, and J. T. Lanzetta, "AN ANALYSIS OF THE ENCODING AND DECODING OF SPONTANEOUS AND POSED SMILES: THE USE OF FACIAL ELECTROMYOGRAPHY," 1989.

[538] J. N. Bailenson *et al.*, "Real-time classification of evoked emotions using facial feature tracking and physiological responses," *Int. J. Hum. Comput. Stud.*, vol. 66, no. 5, pp. 303–317, 2008, doi: 10.1016/j.ijhcs.2007.10.011.

[539] R. A. Calvo and S. K. D'Mello, *New Perspectives on Affect and Learning Technologies*. 2011.

[540] I. Mavridou, "FACETEQ; A novel platform for measuring emotion in VR," in *Virtual Reality International Conference (VRIC '17)*, 2017, doi: 10.1145/3110292.3110302.

[541] I. Mavridou, E. Seiss, T. Kostoulas, C. Nduka, and E. Balaguer-Ballester, "Towards an effective arousal detection system for virtual reality," in *Workshop on Human-Habitat for Health (H3): Human-Habitat Multimodal Interaction for Promoting Health and Well-Being in the Internet of Things Era. ACM*, 2018, p. 4, Accessed: Mar. 03, 2020. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3279963.3279969.

[542] R. Picard, "The promise of Affective Computing," *Handb. Affect. Comput.*, 2014, doi: 10.1093/oxfordhb/9780199942237.013.013.

[543] J. Healey, "Physiological Sensing of Emotion," *Oxford Handb. Affect. Comput.*, p. 204, 2014, doi: 10.1093/oxfordhb/9780199942237.013.023.

[544] J. Wagner, J. Kim, and E. André, "FROM PHYSIOLOGICAL SIGNALS TO EMOTIONS: IMPLEMENTING AND COMPARING SELECTED METHODS FOR FEATURE EXTRACTION AND CLASSIFICATION."

[545] D. Novak, M. Mihelj, and M. Munih, "A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing," *Interact. Comput.*, vol. 24, no. 3, pp. 154–172, 2012, doi: 10.1016/j.intcom.2012.04.003.

[546] C. R. Snyder and S. J. Lopez, *Handbook of positive psychology*. Oxford University Press, 2002.

[547] M. Krijn, P. M. G. Emmelkamp, R. P. Olafsson, and R. Biemond, "Virtual reality exposure therapy of anxiety disorders: a review," *Clin. Psychol. Rev.*, 2004.

[548] A. Steed *et al.*, "Evaluating immersive experiences during Covid-19 and beyond," *Interactions*, vol. 27, no. 4, pp. 62–67, Jul. 2020, doi: 10.1145/3406098.

[549] R. P. Singh, M. Javaid, R. Kataria, M. Tyagi, A. Haleem, and R. Suman, "Significant applications of virtual reality for COVID-19 pandemic," *Diabetes Metab. Syndr. Clin. Res. Rev.*, vol. 14, no. 4, pp. 661–664, Jul. 2020, doi: 10.1016/j.dsx.2020.05.011.

[550] P. Zarjam, J. Epps, and N. H. Lovell, "Beyond Subjective Self-Rating: EEG Signal Classification of Cognitive Workload," *IEEE Trans. Auton. Ment. Dev.*, vol. 7, no. 4, pp. 301–310, 2015, doi: 10.1109/TAMD.2015.2441960.

[551] R. Langner, M. B. Steinborn, A. Chatterjee, W. Sturm, and K. Willmes, "Mental fatigue and temporal preparation in simple reaction-time performance," *Acta Psychol. (Amst).*, vol. 133, no. 1, pp. 64–72, 2010, doi: 10.1016/j.actpsy.2009.10.001.

[552] R. G. Vaurio, D. J. Simmonds, and S. H. Mostofsky, "Increased intra-individual reaction time variability in attention-deficit/hyperactivity disorder across response inhibition tasks with different cognitive demands," *Neuropsychologia*, vol. 47, no. 12, pp. 2389–2396, Oct. 2009, doi: 10.1016/j.neuropsychologia.2009.01.022.

[553] T. Piumsomboon, G. Lee, … R. L.-… I. S. on, and undefined 2017, "Exploring natural eye-gaze-based interaction for immersive virtual reality," *ieeexplore.ieee.org*, Accessed: Mar. 04, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7893315/.

[554] J. L. Soler-Dominguez, J. D. Camba, M. Contero, and M. Alcañiz, "A proposal for the selection of eye-tracking metrics for the implementation of adaptive gameplay in virtual reality based games," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10280, Springer Verlag, 2017, pp. 369–380.

[555] M. L. Mele and S. Federici, "Gaze and eye-tracking solutions for psychological research," *Cogn. Process.*, vol. 13, no. 1 SUPPL, 2012, doi: 10.1007/s10339-012-0499-z.

[556] A. T. Duchowski, "A breadth-first survey of eye-tracking applications," *Behav. Res. Methods, Instruments, Comput.*, vol. 34, no. 4, pp. 455–470, 2002, doi: 10.3758/BF03195475.

[557] R. Shumaker and L. Stephanie, "Virtual, Augmented and Mixed Reality: Designing and Developing Augmented and Virtual Environments - 5th International Conference, VAMR 2013, Held as Part of HCI International 2013, Proceedings," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8021 LNCS, no. PART 1.

2013, Accessed: Mar. 04, 2020. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=MBe6BQAAQBAJ&oi=fnd&pg= PR6&dq=Virtual+Augmented+and+Mixed+Reality.+Designing+and+Developing+ Augmented+and+Virtual+Environments&ots=Ljph9d3cDc&sig=8TTvkvjmWWp3 tlEFsFvBnX7SsM4.

[558] H. Al Osman and T. H. Falk, "Multimodal Affect Recognition: Current Approaches and Challenges," in *Emotion and Attention Recognition Based on Biological Signals and Images*, InTech, 2017.

[559] C. Williams, "Research Methods," 2007.

[560] J. W. Creswell, "RESEARCH DESIGN Qualitative, Quantitative. and Mixed Methods Approaches SECOND EDITION SAGE Publications International Educational and Professional Publisher Thousand Oaks London New Delhi," 2003.

[561] R. S. Kalawsky, "The Validity of Presence as a Reliable Human Performance Metric in Immersive Environments.," *Presence*, no. March, pp. 1–16, 2000, Accessed: Jan. 09, 2020. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.12.8095.

[562] "iMOTIONS, website," doi: 10.3389/FPSYG.2015.01725.\R\N\R\N&A;NBSP;&Q;,&Q;POST_TITLE&Q;:&Q ;WHAT.

A. Appendix

Chapter 6 – Additional material:

A.1 Presence correlations to other subjective ratings per VE

The relationships between presence scores, end of VE valence and arousal ratings, and memory accuracy were investigated separately for each VE, using Pearsons correlations.

For the positive VE**,** presence scores were positively correlated with post-VE valence ratings (r(291)=0.496, *p*<.001), post-VE arousal ratings (r(291)=0.311, *p*<.001). Presence scores were also positively correlated with the CASR valence ratings (r(291)=0.341, *p*<.001) and CASR arousal ratings (r(291)=0.313, *p*<.001). The weak positive correlation between presence ratings a memory accuracy was also significant (r(291)=0.133, *p*=.023).

For the negative VE, a different pattern emerged. For the post-VE ratings, presence scores were only significantly correlated with arousal ratings (r(291)=0.541, *p*=.001), but not with valence ratings. For the average CASR ratings, presence scores were correlated with valence ratings (r(291)=-0.151, *p*=.01) and arousal ratings (r(291)=-0.185, *p*=.002) but these correlations were much weaker. There was no correlation with memory accuracy (r(291)=0.078, *p*=.187).

Finally, for the neutral VE, presence scores were positively correlated with post-VE valence ratings (r(291)=0.217, *p*<.001) and the post-VE arousal ratings (r(291)=0.210, *p*<.001). The same pattern was found for the CASR ratings. Here, presence scores were also positively correlated with the CASR valence ratings (r(291)=0.183, *p*=.002) and the CASR arousal ratings (r(291)=0.182, *p*=.002). However, there was no relationship between presence ratings and memory accuracy (r=-.005, *p*=.395).

The relationship between **presence and enjoyment** was also analysed for this study. Generally, the significant correlations between presence and enjoyment scores were highest for the positive VE (r(291)=0.610, *p*<.001), at a medium level

for the neutral VE (r(291)=0.470, *p<*.001), and lowest for the negative VE (r(291)=0.279, *p<*.001).

For the positive VE, **enjoyment was correlated with all valence and arousal measure**s (all r(291)≥0.325, all *p<*.001) but there was no correlation with memory accuracy (r=0.078, *p=*.185). For the negative VE, enjoyment was correlated with post-VE valence and arousal measures (all r(291)≥0.188, all p≤.001) but there were no significant correlations between enjoyment scores and mean CASR valence ratings (r=0.095, *p=*.105), CASR arousal ratings (r=0.094, *p=*0.110) and memory accuracy (r=0.028, *p=*.635). For the neutral VE, enjoyment was correlated with all valence and arousal measures (all r(291)≥0.219, all *p<*.001) but there was no correlation with memory accuracy (r=0.093, *p=*.112).

Taken together, these findings show that there is a clear relationship between feeling present and affective value of a VE for all three VEs, meaning the more valence or arousal in a VE the higher was the presence rating too. Interestingly, higher presence levels were only related to higher memory scores in the positive VE. **Presence was correlated with enjoyment scores.** This relationship was strongest for the positive VE, less strong for the neutral VE, and least strong for the negative VE. As for the presence scores, there was a **clear relationship between enjoyment and the affective value of a VE for all three VEs**. However, enjoyment was not correlated with memory accuracy.

A.2 Classification Experiments

A.2.1 User-Independent mixed-subjects

Figure 76 shows the highest out-of-sample accuracies reached for valence and arousal detection for binary (2-classes) and ternary (3-classes) classification. These include the accuracies achieved for the active, the passive and the combination of both groups. The blue bars represent the accuracies for binary classification and the red bars for ternary. For both dimensions, the active group was achieved the highest accuracies.
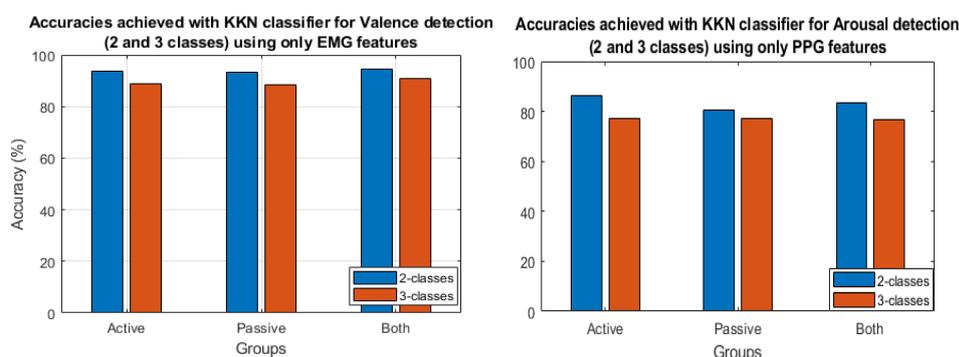

Figure 74. Best accuracies achieved per group (active, passive and combined groups) for valence detection using EMG features (left figure), and for arousal detection using PPG features (right figure).

A.2.2. User-Independent (mixed-subjects): Valence Classification

In this section, we present the confusion matrices and the Roc curves per classifier, firstly for all the subjects and then for active and the passive groups for all valence classification experiments.

Combined groups (2-classes) –Table 36 show the confusion matrices per classifier for two classes valence (negative, positive) from data of both groups, the out-of-sample accuracies obtained and the corresponding F-scores. The best performing classifier was KNN with 94.71% accuracy and the lowest out-of-sample misclassification rate (k-Loss$_{KNN}$ = 0.05, k-Loss$_{SVM}$ =0.16, k-Loss$_{NB}$ = 0.28). Figure 77 shows the ROC curves per classifier with their corresponding area under curve (AUC).

Table 36. Confusion matrixes for the SVM, KNN, NB classifiers for valence, together with their equivalent CV accuracy percentages and F-scores, using data from both groups combined. Values are expressed as percentages.

Confusion Matrix -2 classes					
Classifier	Valence	Negative	Positive	Accuracy	F-Scores
SVM	Negative	79.50	20.50	84.19	0.83, 0.85
	Positive	11.11	88.89		
KNN	Negative	4.03	6.55	94.71	0.95, 0.95
	Positive	5.47	95.97		
Naïve Bayes	Negative	76.90	23.10	72.36	0.73, 0.71
	Positive	32.16	67.84		
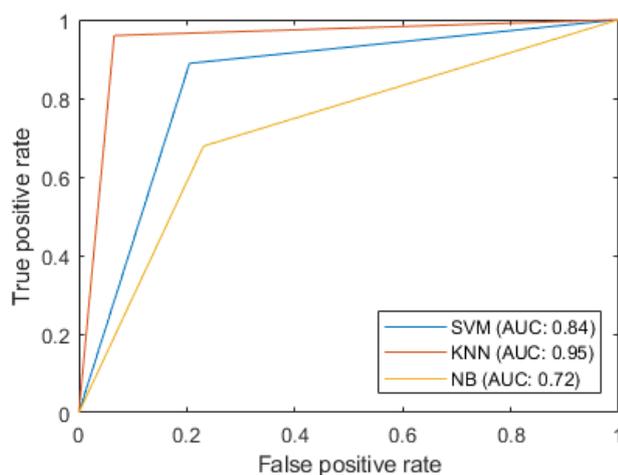
Figure 75. ROC curves per classifier (for combined groups-2 classes valence classification).

Combined groups (3-classes) – Table 37 show the confusion matrices per classifier for three classes valence (negative, neutral positive) using EMG features from data of both groups, the out-of-sample accuracies obtained and the corresponding F-scores. The best performing classifier was KNN with 90.95% accuracy and the lowest out-of-sample misclassification rate (k-Loss$_{KNN}$ = 0.09, k-Loss$_{SVM}$ =0.26, k-Loss$_{NB}$ = 0.41). Figure 78 shows the ROC curves per classifier with their corresponding area under curve (AUC).

Table 37. Confusion matrixes for the SVM, KNN, NB classifiers for 3 classes of valence, together with their equivalent accuracy percentages and F-scores, using data from both groups combined. Values are expressed as percentages.

| Confusion Matrix -Valence -3 classes					
Classifier	Valence	Negative	Neutral	Positive	Accuracy	F-scores
SVM	Negative	79.27	8.69	12.04		
	Neutral	15.33	48.43	36.25	73.95	0.80, 0.57, 0.78
	Positive	5.57	7.41	87.03		
KNN	Negative	92.24	4.31	3.45		
	Neutral	5.59	83.96	10.46	90.95	0.94, 0.85, 0.92
	Positive	0.44	4.86	94.70		
Naïve Bayes	Negative	71.95	9.23	18.82		
	Neutral	35.53	23.23	41.24	58.88	0.65, 0.32, 0.66
	Positive	20.14	7.94	71.92		
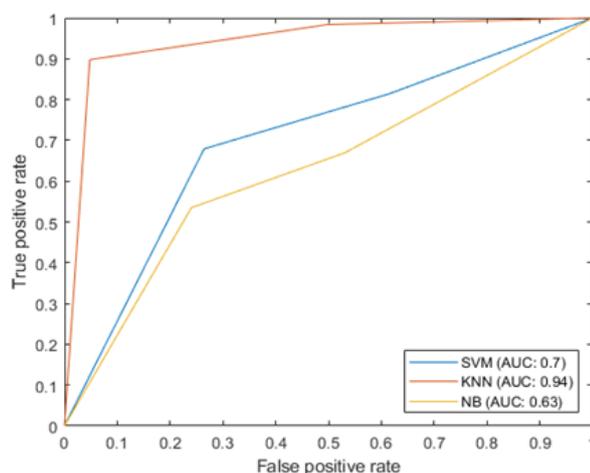
Figure 76. ROC curves per classifier (for combined group-3 classes valence classification).

Active group (2-classes) –

Table **38** show the confusion matrices for the three classifiers per valence class (negative, positive), the out-of-sample accuracies obtained and the corresponding F-scores. The best performing classifier was KNN with 93.11% accuracy and the lowest out-of-sample misclassification rate (k-Loss$_{KNN}$ = 0.06, k-Loss$_{SVM}$ =0.28, k-Loss$_{NB}$ = 0.37). Figure 79 shows the ROC curves per classifier with their corresponding area under curve (AUC).

Table 38. Confusion matrixes (2 classes) of the testing set for the SVM, KNN, NB classifiers for valence, together with their equivalent CV accuracy percentages and F-scores. Values are expressed as percentages.

| Confusion Matrix of testing set - 2 classes					
Classifier	Valence	Negative	Positive	Accuracy	F-Scores
SVM	Negative	83.53	16.47	71.89%	0.74, 0.69
	Positive	39.05	60.95		
KNN	Negative	95.19	4.81	93.62%	0.94, 0.94
	Positive	7.86	92.14		
Naïve Bayes	Negative	63.51	36.49	62.97%	0.62, 0.63
	Positive	37.54	63.49		
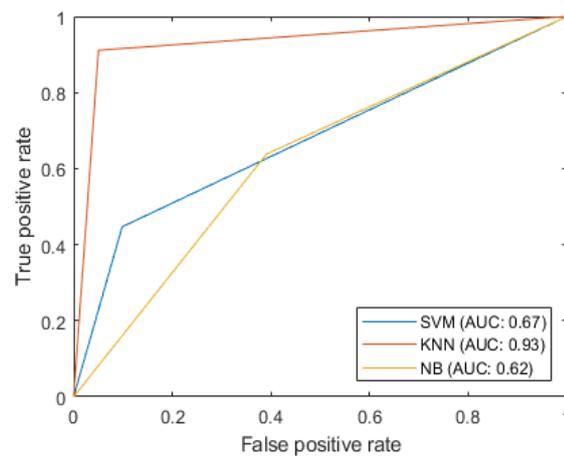
Figure 77. ROC curves per classifier (for Active group-2 classes valence classification).

Active Group (3-classes) – Table 39 shows the confusion matrices, the out-of-sample accuracies and F-scores per optimised classifier for detection of 3 classes of valence, negative, neutral and positive. The best performing classifier was KNN with 88.91% accuracy and the lowest out-of-sample misclassification rate (k-Loss$_{KNN}$ = 0.11, k-Loss$_{SVM}$ =0.38, k-Loss$_{NB}$ = 0.46). Figure 80 shows the ROC curves per classifier with their corresponding area under curve (AUC).

Table 39. Confusion matrixes of the testing set (3 classes) for the SVM, KNN, NB classifiers for valence, together with their equivalent CV accuracy percentages and F-scores. Values are expressed as percentages.

Confusion Matrix -Valence Testing Set - 3 classes						
Classifier	Valence	Negative	Neutral	Positive	Accuracy	F-scores
SVM	Negative	62.75	6.63	30.61	63.65%	0.64, 0.58, 0.65
	Neutral	21.67	50.21	28.13		
	Positive	20.13	6.67	73.20		

|					
	Negative	90.75	8.29	0.97	
KNN	Neutral	4.54	86.58	8.88	88.91%	0.93, 0.82, 0.90
	Positive	1.15	10.09	88.76		
	Negative	61.27	14.56	24.18		
Naïve Bayes	Neutral	30.94	45.15	23.91	53.58%	0.54, 0.50, 0.56
	Positive	33.05	13.46	53.50		
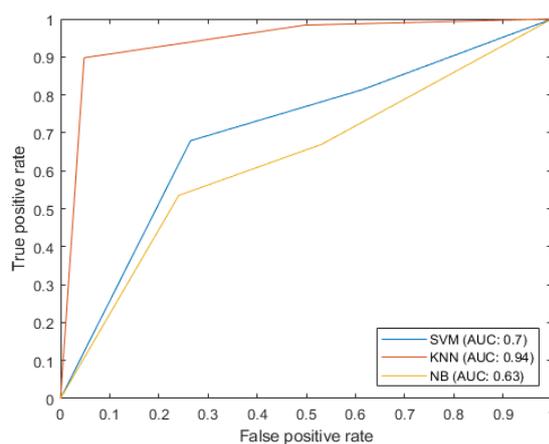
Figure 78. ROC curves per classifier (for Active group - 2 classes valence classification).

Passive group (2-classes) – Table 40 show the confusion matrices for the three classifiers for two valence classes (negative, positive). The best performing classifier was KNN with 93.33% accuracy and the lowest out-of-sample misclassification rate (k-Loss$_{KNN}$ = 0.06, k-Loss$_{SVM}$ =0.19, k-Loss$_{NB}$ = 0.32). Figure 81 shows the ROC curves per classifier with their corresponding areas under curve (AUC).

Table 40. Confusion matrixes of the testing set of passive group for the SVM, KNN, NB classifiers for valence, together with their equivalent CV accuracy percentages and F-scores. Values are expressed as percentages.

| Confusion Matrix of testing set - 2 classes					
Classifier	Valence	Negative	Positive	Accuracy	F-Scores
SVM	Negative	68.43	31.57	80.61	0.77, 0.83
	Positive	8.93	91.07		
KNN	Negative	91.93	8.07	93.33	0.93, 0.94
	Positive	5.47	94.53		
Naïve Bayes	Negative	59.55	49.45	68.49	0.64, 0.72
	Positive	23.82	76.18		
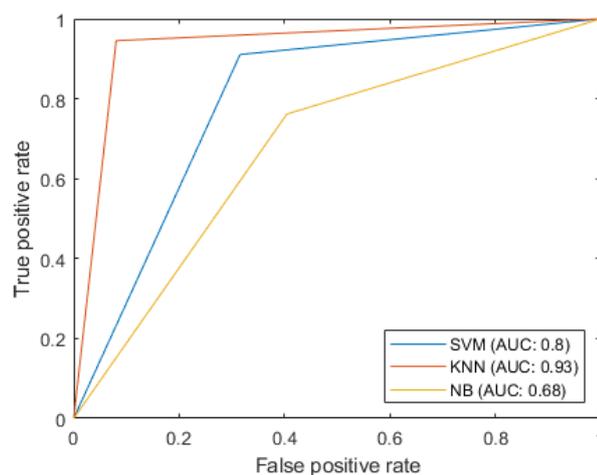
Figure 79. ROC curves per classifier (for Passive group-2 classes valence classification)

Passive group (3-classes) –Table 41 show the confusion matrices for three classes valence (negative, neutral, positive) from data of the passive group., the out-of-sample accuracies obtained and the corresponding F-scores. The best performing classifier was KNN with 88.39% accuracy and the lowest out-of-sample misclassification rate (k-Loss$_{KNN}$ = 0.10, k-Loss$_{SVM}$ =0.24, k-Loss$_{NB}$ = 0.42). Figure 82 shows the ROC curves per classifier with their corresponding area under curve (AUC).

Table 41. Confusion matrixes of the testing set of passive group for the SVM, KNN, NB classifiers for 3 classes of valence, together with their equivalent CV accuracy percentages and F-scores. Values are expressed as percentages.

| Confusion Matrix -Valence Testing Set -3 classes					
Classifier	Valence	Negative	Neutral	Positive	Accuracy	F-scores
	Negative	84.04	11.32	4.64		
SVM	Neutral	19.10	69.01	11.88	75.75	0.79, 0.72, 0.76
	Positive	14.65	12.48	72.87		
	Negative	90.96	8.25	0.79		
KNN	Neutral	5.99	87.76	6.25	90.17	0.92, 0.86, 0.92
	Positive	0.50	7.62	91.88		
Naïve Bayes	Negative	66.90	19.65	13.44		
	Neutral	31.43	51.76	16.81	57.58	0.62, 0.54, 0.56
	Positive	28.22	19.41	52.38		
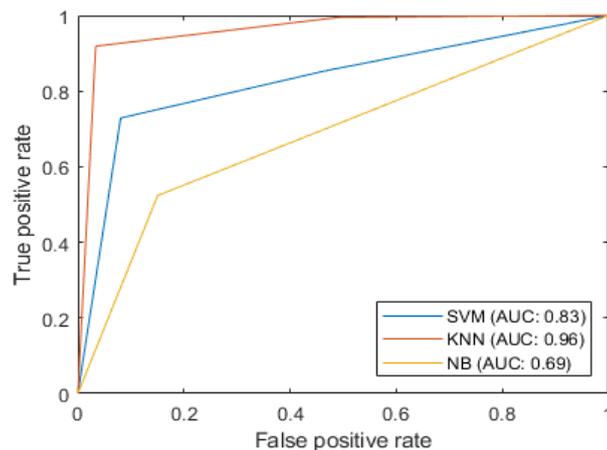
Figure 80. ROC curves per classifier (for combined group-3 classes valence classification).

Valence Classification with fusion of EMG and PPG features – In Table 42 the confusion matrices, accuracies and F-scores and misclassification rates are shown per classifier using the combination of features from the EMG and PPG sensors (fusion). Figure 83 shows the ROC curves per classifier when data from both groups are combined.

Table 42. Confusion matrixes and performance metrics per classifier using combined features (fusion), per group (active, passive and all combined). The confusion matrix and accuracy per classifier are expressed in percentages.

Confusion Matrix -Valence– 2 Classes from EMG & PPG features (fusion)							
Group	Classifier	Arousal	Neg.	Pos.	Accur.	F-scores	k-loss
All (combined)	SVM	Negative	74.72	25.28	73.42%	0.75, 0.72	0.27
		Positive	28.03	71.97			
	KNN	Negative	72.87	27.13	80.87%	0.80, 0.82	0.19
		Positive	10.22	89.78			
	NB	Negative	65.37	24.63	59.31%	0.61, 0.57	0.41
		Positive	47.46	52.54			
Active	SVM	Negative	58.25	41.75	69.57%	0.63, 0.74	0.30
		Positive	21.27	78.73			
	KNN	Negative	75.04	24.96	76.88%	0.74, 0.79	0.23
		Positive	21.64	78.36			
	NB	Negative	46.29	53.71	60.38%	0.54, 0.53	0.40
		Positive	28.24	71.76			
Passive	SVM	Negative	81.92	18.08	75.74%	0.79, 0.71	0.24
		Positive	32.15	67.85			

|						
KNN	Negative	83.38	16.62	81.86%	0.84, 0.79	0.18
	Positive	20.07	79.93			
NB	Negative	65.09	34.91	61.86%	0.66, 0.57	0.38
	Positive	42.27	57,73			
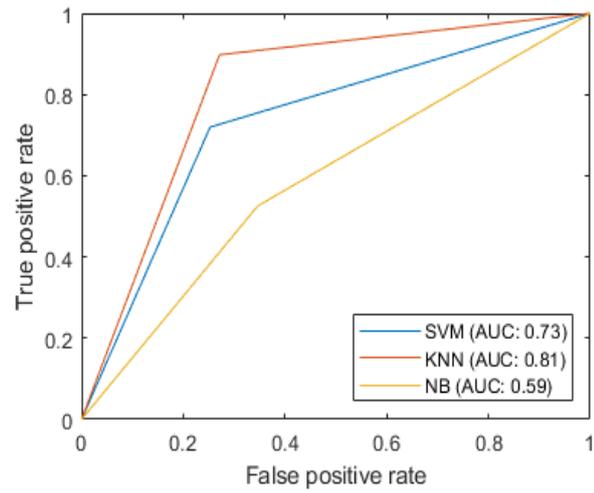
Figure 81. ROC curves per classifier foe valence detection (2 classes) from all participants (both groups combined).

A.2.3 User independent (mixed-subjects): Arousal Classification from PPG features

Combined group (2-classes) –

Table **43** shows the confusion matrixes, accuracies and F-scores of the testing set for three classifiers for two arousal classes, from PPG features from data of both groups combined. The KNN achieved the higher accuracy (83.53%), and the lowest out-of-sample misclassification rate (k-Loss$_{KNN}$ = 0.16, k-Loss$_{SVM}$ = 0.33, k-Loss$_{NB}$ = 0.39). Figure 84 shows the ROC curves per classifier with their corresponding area under curve (AUC).

Table 43. Confusion matrixes per classifier for 2 classes of arousal, together with their equivalent accuracy percentages and F-scores (using data from both groups). Values are expressed as percentages

| Confusion Matrix -Arousal – 2 Classes (Combined groups)					
Classifier	Arousal	Low	High	Accuracy	F-scores
SVM	Low	60.23	39.77	66.79%	0.63, 0.70
	High	27.16	72.84		
KNN	Low	93.93	6.07	83.53%	0.85, 0.82
	High	26.03	73.97		
Naïve Bayes	Low	58.41	41.59	61.36%	0.59, 0.63
	High	35.93	64.07		
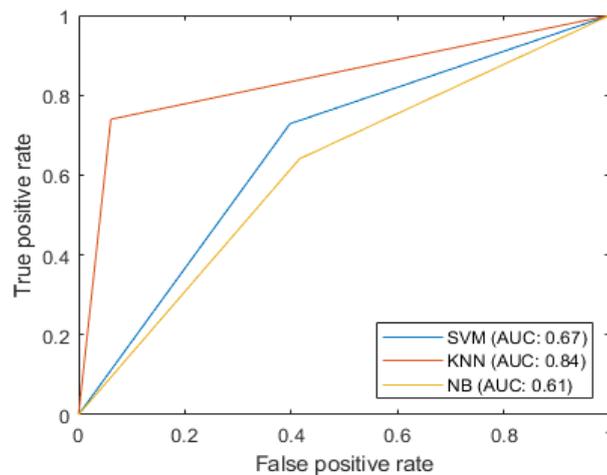
Figure 82. ROC curves per classifier for combined groups (2-classes Arousal classification).

Combined group (3-classes) –

Table **44** shows the confusion matrixes, accuracies and F-scores of the testing set for three classifiers for two arousal classes, from PPG features from data of both groups combined. The KNN achieved the higher accuracy (76.89%), and the lowest out-of-sample misclassification rate (k-Loss$_{KNN}$ = 0.23, k-Loss$_{SVM}$ = 0.48, k-Loss$_{NB}$ = 0.55). Figure 87 shows the ROC curves per classifier with their corresponding area under curve (AUC).

Table 44. Confusion matrixes for 3 classes of arousal, together with their accuracy percentages and F-scores (combined groups). Values are expressed in percentages.

| Confusion Matrix - Arousal -3 classes (Combined groups)						
Classifier	Arousal	Low	Medium	High	Accuracy	F-scores
	Low	56.30	18.90	24.80		
SVM	Medium	26.22	45.96	27.82	54.41%	0.55, 0.48, 0.53
	High	24.43	21.44	54.13		
	Low	90.80	7.29	1.86		
KNN	Medium	25.15	65.55	9.30	76.89%	0.78, 0.71, 0.80
	High	18.84	8.39	72.77		
	Low	43.44	27.25	29.31		
Naïve Bayes	Medium	27.44	44.74	27.82	44.98%	0.45, 0.43, 0.47
	High	26.96	26.30	46.74		
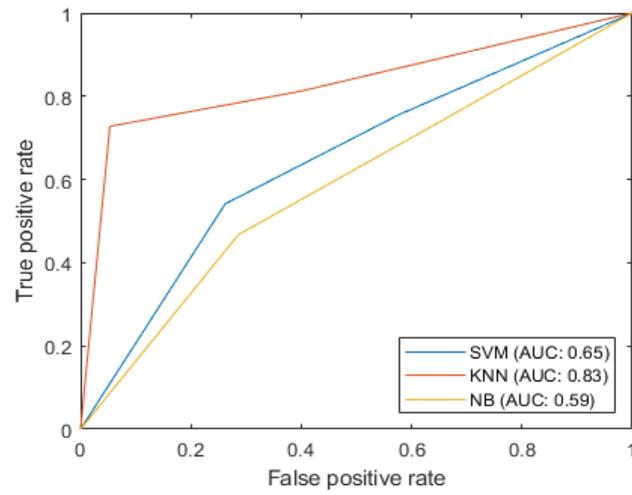
Figure 83. ROC curves per classifier for combined groups (3-classes Arousal classification).

Active group (2-classes) – Table 45 shows the confusion matrix for the three classifiers (2 levels of arousal, low-high), the out-of-sample average accuracies per model tested and the F-Scores distinguishing two arousal classes (low, high), from PPG features of the active group. The KNN achieved the higher accuracy (86.10 %), and the lowest out-of-sample misclassification rate (k-Loss$_{KNN}$ = 0.14, k-Loss$_{SVM}$ =0.30, k-Loss$_{NB}$ = 0.38). Figure 86 shows the ROC curves per classifier with their corresponding area under curve (AUC).

Table 45. Confusion matrices for 2 classes of arousal, together with their equivalent accuracy percentages and F-scores for the active group.

| Confusion Matrix -Arousal -Testing Set – 2 Classes (Active)					
Classifier	Arousal	Low	High	Accuracy	F-scores
SVM	Low	75.38	24.62	69.63 %	0.74, 0.64
	High	37.97	62.03		
KNN	Low	93.61	6.39	86.10 %	0.88, 0.83
	High	23.82	76.18		
Naïve Bayes	Low	66.92	33.08	62.25 %	0.67, 0.56
	High	43.92	56.08		
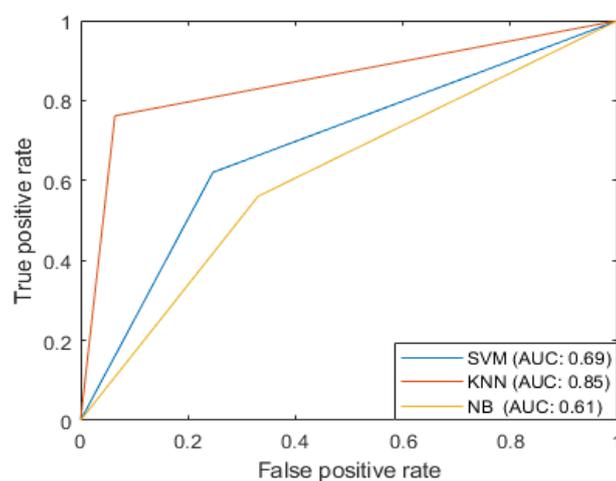
Figure 84. ROC per classifier for Active group (2-classes Arousal classification)

Active group (3-classes) –

Table **46** shows the confusion matrices, the out-of-sample accuracies and F-scores per optimised classifier for detection of 3 classes of arousal (low, average and high), from PPG features of the active group. The KNN achieved the higher accuracy

(76.93%), and the lowest out-of-sample misclassification rate (k-Loss$_{KNN}$ = 0.23, k-Loss$_{SVM}$ = 0.40, k-Loss$_{NB}$ = 0.50). Figure 89 shows the ROC curves per classifier with their corresponding area under curve (AUC).

Table 46. Confusion matrixes for the SVM, KNN, NB classifiers for 3 classes of arousal, together with their equivalent CV accuracy percentages and F-scores (active group) . Values are expressed as percentages.

| Confusion Matrix -Arousal Testing set -3 classes (Active)						
Classifier	Valence	Low	Medium	High	Accuracy	F-scores
SVM	Low	63.57	26.39	10.04		
	Medium	25.56	63.33	11.11	59.63%	0.60 0.60, 0.57
	High	25.45	24.11	50.45		
KNN	Low	92.19	6.32	1.49		
	Medium	27.04	65.19	7.78	76.93%	0.79, 0.73, 0.79
	High	18.30	8.89	72.77		
Naïve Bayes	Low	55.76	28.25	15.98		
	Medium	26.67	53.33	20	50.20%	0.56, 0.50, 0.43
	High	21.43	38.84	39.73		
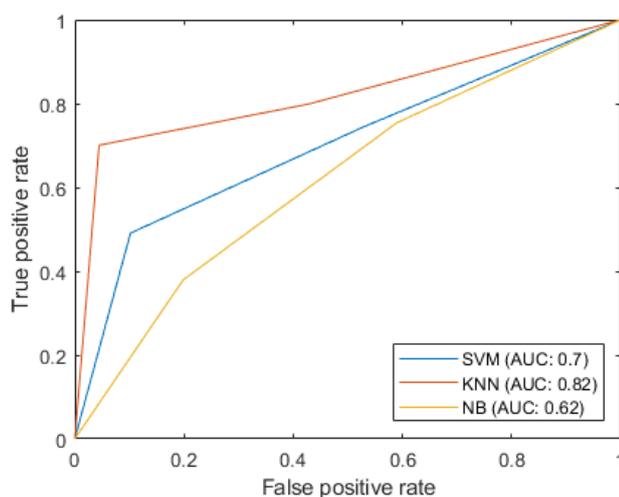
Figure 85. Roc curves per classifier for Active group (3 classes Arousal classification)

Passive group (2-classes) –

Table **47** shows the confusion matrixes, accuracies and F-scores of the testing set for three classifiers for two arousal classes, from PPG features of the passive group. The KNN achieved the higher accuracy (80.54%), and the lowest out-of-sample misclassification rate (k-Loss$_{KNN}$ = 0.19, k-Loss$_{SVM}$ = 0.32, k-Loss$_{NB}$ = 0.38). Figure

88 shows the ROC curves per classifier with their corresponding area under curve (AUC).

Table 47. Confusion matrixes per classifier for 2 classes of arousal, together with their equivalent CV accuracy percentages and F-scores (passive group). Values are expressed as percentages

| Confusion Matrix -Arousal -Testing set – 2 Classes (Passive)					
Classifier	Arousal	Low	High	Accuracy	F-scores
SVM	Low	69.06	30.94	67.70	0.68, 0.67
	High	33.68	66.32		
KNN	Low	91.82	8.18	80.54	0.83, 0.78
	High	30.94	69.06		
Naïve Bayes	Low	65.70	34.30	62.50	0.64, 0.61
	High	40.75	59.25		
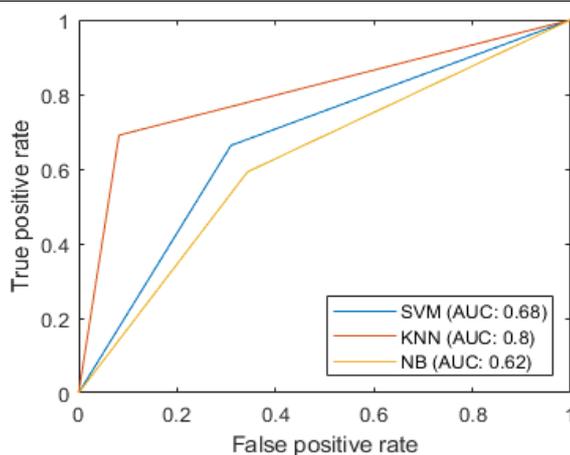
Figure 86. ROC per classifier for Passive group (2-classes Arousal classification).

Passive group (3-classes) – Table 48 show the confusion matrixes, accuracies and F-scores for three classes arousal (low, medium, high) from PPG features of the passive group. The KNN achieved the higher accuracy (77.30%), and the lowest out-of-sample misclassification rate (k-Loss$_{KNN}$ = 0.23, k-Loss$_{SVM}$ = 0.38, k-Loss$_{NB}$ = 0.49). Figure 89 shows the ROC curves per classifier with their corresponding area under curve (AUC).

Table 48. Confusion matrixes for 3 classes of arousal, together with their equivalent CV accuracy percentages and F-scores (passive group). Values are expressed as percentages.

Confusion Matrix - Arousal Testing set -3 classes (Passive)						
Classifier	Arousal	Low	Medium	High	Accuracy	F-scores
SVM	Low	68.28	14.19	17.53		
	Medium	24.86	55.54	19.60	61.81%	0.64, 0.59, 0.62
	High	21.54	17.36	61.10		
KNN	Low	92.82	5.83	1.34		
	Medium	25.70	68.97	5.99	77.30%	0.77, 0.76, 0.79
	High	24.37	6.18	69.45		
Naïve Bayes	Low	53.26	21.87	24.87		[0.53, 0.49, 0.52]
	Medium	25.41	48.46	26.13	51.23%	
	High	25.04	23.21	51.75		
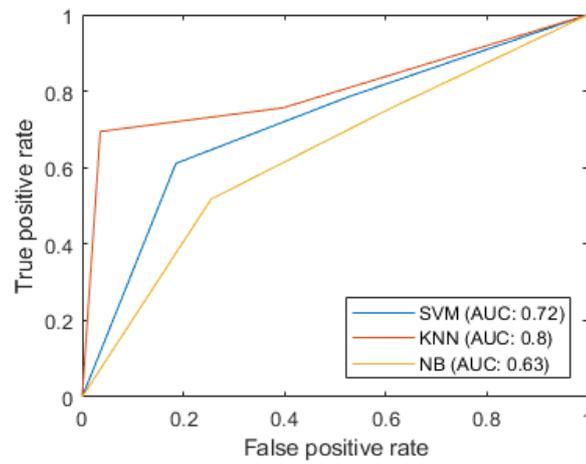
Figure 87. ROC per classifier for Passive group (3-classes Arousal classification).

Arousal classification with fusion of EMG and PPG features – In . The Roc curves are presented in Figure 90.

Table 49 the confusion matrices, accuracies and F-scores and misclassification rates are shown per classifier using the combination of features from the EMG and PPG sensors (fusion). The Roc curves are presented in Figure 90.

Table 49. Confusion matrixes and performance metrics per classifier using combined features (fusion), per group (active, passive and all combined). The confusion matrix and accuracy per classifier are expressed in percentages.

| Confusion Matrix -Arousal – 2 Classes from EMG & PPG features (fusion)							
Group	Classifier	Arousal	Low	High	Accur.	F-scores	k-loss
All (combined)	SVM	Low	71.12	28.88	74.22	0.74, 0.75	0.26
		High	22.62	77.38			
	KNN	Low	73.14	26.86	81.47	0.80, 0.83	0.19
		High	10.04	89.96			
	NB	Low	64.39	35.61	59.23	0.61, 0.57]	0.40
		High	46.03	53.97			
Active	SVM	Low	66.90	30.10	67.98	0.70, 0.66	0.32
		High	34.14	65.86			
	KNN	Low	89.15	10.85	80.87	0.83, 0.78	0.19
		High	28.29	71.71			
	NB	Low	52.07	47.93	53.53	0.54, 0.53	0.47
		High	45.29	54.71			
Passive	SVM	Low	65.84	34.16	75.68	0.72, 0.79	0.24
		High	15.77	84.23			
	KNN	Low	70.38	29.62	80.67	0.77, 0.83	0.19
		High	10.39	89.61			
	NB	Low	66.25	33.75	58.92	0.60, 0.58	0.41
		High	47.46	52.54			
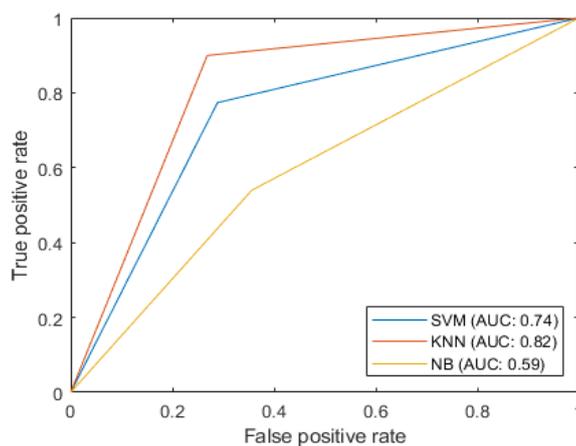
Figure 88. ROC curves per classifier for combined groups using EMG and PPG derived features (2-classes Arousal classification).

Arousal Classification from EMG features

Combined groups (2-classes) –

Table **50** shows the confusion matrices, accuracies and F-scores of the testing set for three classifiers for two arousal classes, from EMG features from data of both groups combined. The KNN achieved the higher accuracy (93.37%), and the lowest out-of-sample misclassification rate (k-Loss$_{KNN}$ = 0.07, k-Loss$_{SVM}$ = 0.17, k-Loss$_{NB}$ = 0.28). Figure 91 shows the ROC curves per classifier with their corresponding area under curve (AUC).

Table 50. Confusion matrixes and performance metrics per classifier using combined features (fusion), for all users combined for arousal detection (2-classes) using only EMG features. The confusion matrix and accuracy per classifier are expressed in percentages.
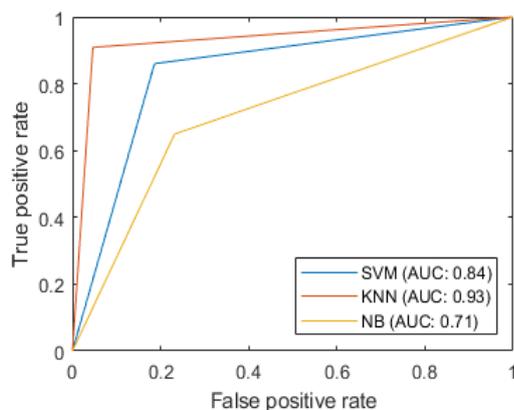

Figure 89. ROC curves per classifier for 2-classes arousal detection using EMG features (all users).

Confusion Matrix -Arousal – 2 Classes (Combined groups)					
Classifier	Arousal	Low	High	Accuracy	F-scores
SVM	Low	81.38	18.62	83.45%	0.84, 0.82
	High	13.98	86.02		
KNN	Low	95.37	4.63	93.37%	0.94, 0.92
	High	9.11	90.89		
Naïve Bayes	Low	76.81	23.19	71.52%	0.75, 0.67
	High	35.04	64.96		

Active group (2-classes) **–**

Table **51** shows the confusion matrixes, accuracies and F-scores of the testing set for three classifiers for two arousal classes, from EMG features from the Active group. The KNN achieved the higher accuracy (90.93%), and the lowest out-of-sample misclassification rate (k-Loss$_{KNN}$ = 0.09, k-Loss$_{SVM}$ = 0.28, k-Loss$_{NB}$ = 0.37). Figure 92 shows the ROC curves per classifier with their corresponding area under curve (AUC).

Table 51. Arousal classification using EMG features (Active group)

| Confusion Matrix -Arousal – 2 Classes (Active group)					
Classifier	Arousal	Low	High	Accuracy	F-scores
SVM	Low	58.75	41.25	72.28	0.89, 0.75
	High	13.27	86.73		
KNN	Low	89.69	10.31	90.93	0.91, 0.91
	High	7.74	92.26		
Naïve Bayes	Low	65.11	34.89	62.84	0.64, 0.61
	High	39.58	60.42		
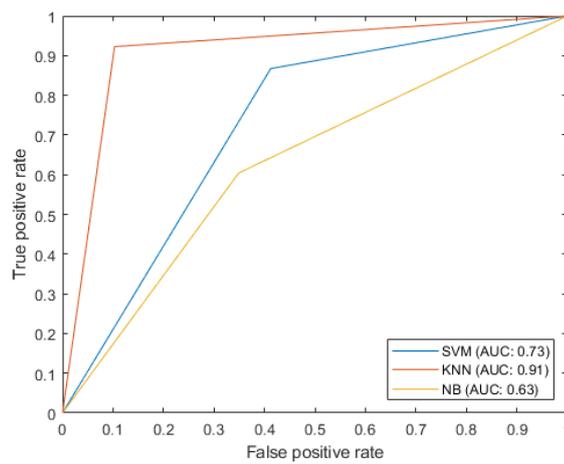
Figure 90. ROC curves per classifier for 2 -classes arousal detection using EMG features. (active group)

Passive group (2-classes) –

Table **52** shows the confusion matrixes, accuracies and F-scores of the testing set for three classifiers for two arousal classes, from EMG features from the Active group. The KNN achieved the higher accuracy (91.71%), and the lowest out-of-sample misclassification rate (k-Loss$_{KNN}$ = 0.08, k-Loss$_{SVM}$ = 0.20, k-Loss$_{NB}$ = 0.36). Figure 93 shows the ROC curves per classifier with their corresponding area under curve (AUC).

Table 52. Arousal classification using EMG features (Active group)

| Confusion Matrix -Arousal – 2 Classes (Active group)					
Classifier	Arousal	Low	High	Accuracy	F-scores
SVM	Low	76.21	23.79	80.09	0.80, 0.81
	High	15.85	84.15		
KNN	Low	92.93	7.07	91.71	0.92, 0.91
	High	9.57	90.43		
Naïve Bayes	Low	74.98	25.02	63.82	0.67, 0.58
	High	47.87	52.13		
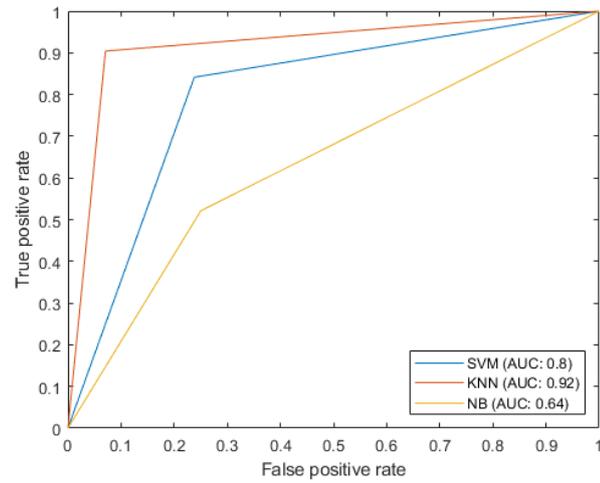
Figure 91. ROC curves per classifier for 2-classes arousal detection using EMG features (passive group)

B.   Appendix: Glossary

Accuracy: The extent to which a system's detection performance corresponds to the given 'ground-truth' given by the participants in an experimental task.

Affect: A general property of experience that has at least two features: pleasantness or unpleasantness (valence) and degree of arousal. Affect is part of every waking moment of life and is not specific to instances of emotion, although all emotional experiences have affect at their core.

Agreement: The extent to which two people provide consistent responses; high agreement produces high inter-subject consistency.

Appraisal: A psychological feature of experience (e.g., experiencing an event as unexpected). The word appraisal is commonly used in research as the cognitive assessment of a feature of an experience (e.g., the judgement of whether the event was unexpected or novel).

Approach/avoidance: A fundamental dimension of motivated behaviour towards an event or stimulus.

Arousal: The intensity of the physiological component of one's emotions. Arousal is one dimension of affect, and it is independent of valence. For example, someone can be happy and serene (high in valence but low in arousal) while watching the sea waves and they can be happy and excited (same level of valence but high in arousal) when receiving a good surprise gift.

Consistency: An outcome that does not vary greatly across time, context, and different individuals.

Emotional expression: A facial configuration, bodily movement, or vocal expression that reliability and specifically communicates an emotional state.

Facial action coding system (FACS): A system to describe and quantify visible human facial movements.

Facial expression: A facial configuration caused by contractions of the facial muscles that is often inferred to express an internal state.

Generalize/generalizability: High venerability means that findings can be replicated or a detection model can be applied to new users, across different settings.

Multimodal: Combining information from more than one of the senses (e.g., vision and audition). Similarly, with sensor multimodality researchers refer to the integration of various types of sensors to collect data for a certain task.

Perceiver-dependent: An observation that depend on human judgment.

Prototype: A new, preliminary version of a model, for example of a type of technology. Prototypes are often used to present feasibility of one's idea that the creation or performance of such model is possible.

Reliability: An observation that is repeatable across time, context, and individuals.

Sensitivity - Refers to the ability of a system or a being in detecting changes around them. High sensitivity can provide with finer-grained detections. In binary classification test, sensitivity is the ratio of correctly detected true positives.

Specificity: A measure of evaluating how well the system can correctly detect true negatives, or 'false alarms'. A balanced, detection system should have a good a level of specificity as sensitivity (see sensitivity).

Universal: Something that is common or shared by all humans, across cultures.

C. Appendix: Study Materials

Chapter 4: Video validation study (Section 4.2.3).

Table 53. Videos clips selected from the film database by Samson et al. [247]. The video clips are sorted by targeted affective category, with their rescaled mean valence and arousal scores per video (scaled from 1-6 to 1-9). The video names are extracted from the original list.

Category	ID	Video Titles	V_{FL}	A_{FL}
PL	1	Babybitesbrosfingers	7.11	2.89
	2	Babydoesn'tlovehis daddy	6.26	2.41
	3	Babyshiccupandjauph	6.65	2.81
	4	Cookiebaby	6.57	2.68
	5	Smartbabywithpacifier	6.81	2.33
	6	Excalatorspinning	6.36	3.05
	7	Beatboxbabydance	6.70	2.82
	8	Catsucklesair •	7.16	3.00
PH	1	Babydancebeyonce	7.37	3.82
	2	Babyfailshulahoop	7.45	3.83
	3	Babycontrolscheers	7.91	3.80
	4	Bridelauphingduringvows	6.84	3.59
	5	Girlthrownintobasketballhoop	5.75	3.93
	6	Pandasneezealot	6.89	3.80
	7	Singingdog	7.30	3.93
	8	Weedingphtographerfails	6.30	3.45
NL	1	Armbentfromskateboard	2.76	2.74
	2	Bmxfaceplant	2.87	2.49
	3	Boyfaceplants	3.02	2.60
	4	Bullwrongtarget	3.26	3.03
	5	Snowboardercrashes	3.72	2.81
	6	Tablebackflip	3.13	2.92
	7	Skaterfallsbreakwrist	2.31	3.08
	8	Bikeintowall	2.82	2.95
NH	1	Bikefalloffcliff	2.58	4.06
	2	Bullthrownandtrample	2.65	4.07
	3	Carhitsskater	2.57	4.02
	4	Breakdancerkickskid	2.87	3.90
	5	Crocbitesman	2.23	4.18
	6	Fatboyrollercoaster	3.58	3.99
	7	Horribleskiaccident	2.74	3.90
	8	Motorcyclejumperfriuiflesshort	2.90	4.01
NEUTRAL	1	Bart	4.44	1.51
	2	Boydrinkingtea	4.22	1.34
	3	Assembly	4.26	1.45
	4	Airport2	4.30	1.50
	5	Cablecar	4.54	1.67
	6	Café	4.66	1.35

7	Cityinthenight	4.34	1.48
8	Denvertrain	4.65	1.83
9	Eatingpizza	4.52	1.61
10	Eatingwithchopstics	4.66	1.83
11	Gilsbruthingtheirteeth	4.57	1.62
12	Hairwashing	4.47	1.56
13	Hikinginthewood	4.65	1.69
14	Museum	4.30	1.69
15	Nystreet	4.28	1.88
16	Ridingthetube1	4.31	1.56
17	Ridingthetube2	4.50	1.56
18	Sanfran	4.06	1.54
19	Sittingonthesofa	4.02	1.50
20	Snow	4.57	1.70
21	Swimlaps	4.57	1.48
22	Tea	4.33	1.37
23	Treadmill	4.47	1.51
24	Vangoghmuseum	4.65	1.83
25	Pillow	4.81	1.91

Chapter 5: Construction of the Virtual Environments.

Photos of office room were used to recreate the VR office replicas. The following pictures were taken from the physical room in order to design details used for the development of the virtual replicas.

Table 54. This table contains pictures of the targeted stimuli used in the VR scenes. All other static objects and events of the scenes were also tracked but were omitted from this table.

Tracked objects/events per scene	Scene / VR_Object/EventID	Picture
The bookcase	All / 15	
The clock	All /24	
The green folder	All / 7	
The grey notebook	All / 8	
The guitar	All/10	

The window	All /17	
The newsboard	All /16	
The cup	All / 0	
The mouse	All / 5	
The desks	All / 37	
The bin	Neutral / 9	

The mirror	All / 22	
The carpet	All /18	
The fire alarm	Negative /34	
The documents	Negative / 42	
The lightening /shadow on the window	Negative /39	
The glitch in viewpoint	Negative /36	
The fire/smoke	Negative /32	

The overflowing rubbish	Negative /9,53	
The flickering light	Negative / 33	
The mirror	Negative /39	
The spiders	Negative / 8-18	
The light bulb fusing	Negative /33	
The spilt drink (coffee-cup)	Negative / 28	

The rat	Negative /100	
The spider attack	Negative /37	
The green plant	Positive / 2	
Poster of a baby	Positive / 0, 201(sound)	
The light explosion (Changed Event "Light Explosion" to "Light and Laughter".)	Positive /9	
The mirror reflection	Positive /4	
The dog poster	Positive / 39	

The butterflies	Positive /6	
The robot	Positive /1	
The monitor message	Positive /205	
Star dust in the room	Positive /7	
The guitar	Positive /25	
The growing flower on the desk	Positive /3	
The birds	Positive /203	

The amplifier	C-10 Positive /7	
The ball	Positive / 34	

Correlation of Coefficient per stimulus used in survey (Section 5.3)

Table 55. List of video and object stimuli used in survey prior to main Fedem3 study. For each stimulus the mean (M), standard deviation (Std.) and Coefficient of Variation (CV) are reported for valence (V) and Arousal (A).

	Neutral Condition				Negative Condition				Positive Condition		
	M	**Std**	**CV**		**M**	**Std**	**CV**		**M**	**Std**	**CV**
Video_V	4.806	1.468	30.555	**Video_V**	3.119	1.644	52.697	**Video_V**	6.179	1.486	24.041
Video_A	2.552	1.669	65.382	**Video_A**	6.134	1.811	29.526	**Video_A**	4.552	2.208	48.498
Neu_1_V	4.750	1.199	25.241	**Neg_1_V**	3.209	1.570	48.915	**Pos_1_V**	5.388	1.516	28.130
Neu_1_A	2.373	1.534	64.642	**Neg_1_A**	5.851	2.053	35.097	**Pos_1_A**	3.343	1.974	59.047
Neu_2_V	4.478	1.353	30.225	**Neg_2_V**	4.075	1.696	41.613	**Pos_2_V**	5.746	1.624	28.260
Neu_2_A	2.179	1.424	65.345	**Neg_2_A**	3.776	1.991	52.731	**Pos_2_A**	3.358	1.906	56.757
Neu_3_V	4.537	1.605	35.380	**Neg_3_V**	3.254	1.713	52.658	**Pos_3_V**	5.403	1.924	35.617
Neu_3_A	2.388	1.545	64.693	**Neg_3_A**	6.090	1.991	32.687	**Pos_3_A**	5.030	2.178	43.309
Neu_4_V	4.343	1.531	35.256	**Neg_4_V**	3.612	1.476	40.857	**Pos_4_V**	5.657	1.881	33.256
Neu_4_A	2.493	1.705	68.400	**Neg_4_A**	4.821	1.977	41.007	**Pos_4_A**	4.373	2.101	48.037
Neu_5_V	4.612	1.666	36.119	**Neg_5_V**	3.299	1.612	48.860	**Pos_5_V**	6.537	1.765	26.995
Neu_5_A	2.940	1.803	61.314	**Neg_5_A**	5.776	1.786	30.915	**Pos_5_A**	4.448	2.469	55.520
Neu_6_V	4.776	1.544	32.319	**Neg_6_V**	4.000	1.574	39.351	**Pos_6_V**	6.881	1.881	27.337
Neu_6_A	3.119	1.825	58.491	**Neg_6_A**	3.433	1.941	56.545	**Pos_6_A**	4.761	2.419	50.815
Neu_7_V	4.418	1.468	33.218	**Neg_7_V**	3.522	1.587	45.051	**Pos_7_V**	5.776	2.043	35.369
Neu_7_A	2.642	1.562	59.115	**Neg_7_A**	5.328	2.003	37.589	**Pos_7_A**	4.925	2.377	48.265
Neu_8_V	4.612	1.564	33.914	**Neg_8_V**	2.761	1.845	66.835	**Pos_8_V**	6.015	1.732	28.795
Neu_8_A	2.418	1.575	65.158	**Neg_8_A**	6.433	2.228	34.628	**Pos_8_A**	4.522	2.160	47.772
Neu_9_V	4.701	1.526	32.459	**Neg_9_V**	3.104	1.854	59.716	**Pos_9_V**	6.269	1.833	29.244
Neu_9_A	2.567	1.538	59.903	**Neg_9_A**	5.896	2.395	40.622	**Pos_9_A**	4.896	2.293	46.838
Neu_10_V	4.761	1.436	30.163	**Neg_10_V**	3.776	1.794	47.510	**Pos_10_V**	4.955	1.559	31.464
Neu_10_A	2.463	1.587	64.425	**Neg_10_A**	5.030	2.246	44.651	**Pos_10_A**	3.149	1.910	60.660
Neu_11_V	4.597	1.649	35.865	**Neg_11_V**	4.075	1.558	38.236	**Pos_11_V**	5.537	1.765	31.870
Neu_11_A	2.388	1.555	65.096	**Neg_11_A**	3.343	1.936	57.905	**Pos_11_A**	3.627	2.218	61.161
Neu_12_V	4.791	1.626	33.932	**Neg_12_V**	3.761	1.754	46.640	**Pos_12_V**	5.910	1.690	28.601
Neu_12_A	3.507	2.083	59.392	**Neg_12_A**	4.463	2.146	48.095	**Pos_12_A**	3.821	2.137	55.918
Neu_13_V	4.597	1.394	30.315	**Neg_13_V**	2.627	1.953	74.365	**Pos_13_V**	4.731	1.589	33.584
Neu_13_A	2.373	1.601	67.451	**Neg_13_A**	6.657	2.328	34.972	**Pos_13_A**	3.284	1.802	54.885
Neu_14_V	4.881	1.399	28.655	**Neg_14_V**	3.209	1.579	49.210	**Pos_14_V**	5.299	1.820	34.357
Neu_14_A	2.761	1.763	63.839	**Neg_14_A**	5.567	2.132	38.290	**Pos_14_A**	3.343	1.951	58.364
M_CV_V			32.361				49.987				30.920
Std.			3.018				10.104				2.967
M_CV_A			63.376				41.838				53.382
Std.			3.056				8.538				5.556

Table 56. Mean self-rating valence and arousal scores of static objects per VE scenario. ID is for the identity number of the object, M for mean, Sd for standard deviation, CV (%) for coefficient of variation

A/V		Positive VE				Negative VE				Neutral VE		
	ID	M.	SD.	CV.	ID	M	SD	CV	ID	M.	SD.	CV.
V.	1	0.96	0.27	27.91	1	0.81	0.34	41.97	1	0.73	0.25	34.01
A.		0.93	0.35	37.41		0.51	0.43	84.13		0.41	0.31	76.64
V.	2	0.95	0.27	28.56	6	0.96	0.29	30.52	5	1.00	0.29	29.50
A.		0.93	0.35	37.28		0.88	0.31	35.69		0.86	0.27	31.72
V.	3	0.96	0.27	27.92	8	0.96	0.27	28.17	6	0.99	0.27	27.19
A.		0.93	0.34	36.65		0.92	0.40	43.33		0.98	0.42	43.16
V.	4	0.95	0.27	28.54	9	0.98	0.25	25.16	7	1.01	0.25	25.20
A.		0.93	0.35	37.29		0.88	0.35	39.70		0.89	0.33	37.34
V.	5	0.96	0.27	28.58	10	0.97	0.26	26.93	8	1.01	0.26	25.56
A.		0.92	0.35	37.78		0.89	0.38	42.62		0.94	0.40	42.72
V.	6	0.97	0.26	27.23	11	0.93	0.24	25.92	9	0.95	0.23	23.97
A.		0.90	0.31	34.04		0.90	0.41	45.08		0.96	0.44	45.72
V.	7	0.96	0.27	27.92	12	0.95	0.29	30.54	10	1.01	0.28	28.29
A.		0.93	0.34	36.65		0.89	0.33	37.41		0.89	0.31	34.91
V.	8	0.97	0.27	27.85	13	0.97	0.26	26.93	11	1.01	0.26	25.56
A.		0.90	0.31	34.74		0.89	0.38	42.62		0.94	0.40	42.72
V.	9	0.96	0.27	27.92	14	0.97	0.26	26.93	12	1.01	0.26	25.56
A.		0.93	0.34	36.65		0.89	0.38	42.62		0.94	0.40	42.72
V.	10	0.92	0.27	28.92	17	0.97	0.26	27.24	15	1.00	0.26	25.89
A.		0.90	0.33	36.56		0.89	0.39	43.26		0.94	0.41	43.26
V.	11	0.95	0.27	28.56	18	1.00	0.24	24.15	16	1.02	0.25	24.13
A.		0.93	0.35	37.28		0.93	0.36	39.20		0.96	0.39	41.03
V.	12	0.94	0.30	31.60	20	0.97	0.26	26.93	17	1.01	0.26	25.56
A.		0.92	0.38	41.44		0.89	0.38	42.62		0.94	0.40	42.72
V.	13	0.97	0.26	26.74	22	0.96	0.26	27.03	18	1.00	0.26	25.65
A.		0.95	0.34	35.53		0.90	0.39	42.84		0.94	0.40	42.77
V.	15	0.94	0.27	29.02	24	0.99	0.26	26.33	20	1.03	0.25	24.24
A.		0.92	0.35	37.87		0.89	0.37	42.08		0.93	0.39	42.09
V.	16	0.95	0.28	29.23	23	0.97	0.27	27.36	21	1.01	0.26	25.95
A.		0.93	0.36	38.24		0.89	0.39	43.32		0.94	0.41	43.33
V.	17	0.97	0.27	28.13	25	0.97	0.26	26.93	22	1.01	0.26	25.56
A.		0.95	0.35	36.82		0.89	0.38	42.62		0.94	0.40	42.72
V.	18	0.94	0.26	28.02	26	1.00	0.27	27.51	23	1.02	0.29	28.20
A.		0.94	0.35	36.72		0.84	0.34	40.45		0.81	0.32	39.23
V.	19	0.93	0.23	25.09	27	0.99	0.28	27.89	24	1.03	0.27	26.24

| | | | | | | | | | | | |
|----|----|------|------|-------|------|------|-------|------|------|-------|
| A. | | 0.93 | 0.38 | 40.44 | 0.86 | 0.39 | 45.05 | 0.91 | 0.41 | 45.83 |
| V. | 20 | 0.96 | 0.28 | 28.55 | | | | | | |
| A. | | 0.88 | 0.31 | 34.63 | | | | | | |
| V. | 21 | 0.95 | 0.28 | 29.71 | | | | | | |
| A. | | 0.94 | 0.37 | 39.91 | | | | | | |
| V. | 22 | 0.96 | 0.27 | 27.91 | | | | | | |
| A. | | 0.93 | 0.35 | 37.41 | | | | | | |
| V. | 23 | 0.96 | 0.27 | 27.92 | | | | | | |
| A. | | 0.93 | 0.34 | 36.65 | | | | | | |
| V. | 24 | 0.96 | 0.27 | 28.46 | | | | | | |
| A. | | 0.91 | 0.33 | 36.31 | | | | | | |
| V. | 25 | 0.96 | 0.27 | 27.92 | | | | | | |
| A. | | 0.93 | 0.34 | 36.65 | | | | | | |
| V. | 31 | 0.99 | 0.27 | 27.59 | | | | | | |
| A. | | 0.97 | 0.36 | 36.99 | | | | | | |
| V. | 34 | 0.95 | 0.27 | 28.56 | | | | | | |
| A. | | 0.93 | 0.35 | 37.28 | | | | | | |
| V. | 35 | 0.96 | 0.27 | 28.58 | | | | | | |
| A. | | 0.92 | 0.35 | 37.78 | | | | | | |
| V. | 37 | 0.96 | 0.27 | 27.92 | | | | | | |
| A. | | 0.93 | 0.34 | 36.65 | | | | | | |
| V. | 38 | 0.95 | 0.27 | 28.54 | | | | | | |
| A. | | 0.93 | 0.35 | 37.29 | | | | | | |
| V. | 39 | 0.95 | 0.27 | 28.57 | | | | | | |
| A. | | 0.93 | 0.35 | 37.45 | | | | | | |
| V. | 40 | 0.96 | 0.27 | 27.92 | | | | | | |
| A. | | 0.93 | 0.34 | 36.65 | | | | | | |
| V. | 41 | 0.94 | 0.29 | 30.39 | | | | | | |
| A. | | 0.88 | 0.34 | 39.01 | | | | | | |
| V. | 42 | 0.97 | 0.27 | 27.78 | | | | | | |
| A. | | 0.93 | 0.35 | 37.46 | | | | | | |
| V. | 43 | 0.96 | 0.27 | 27.92 | | | | | | |
| A. | | 0.93 | 0.34 | 36.65 | | | | | | |
| V. | 44 | 0.96 | 0.27 | 27.92 | | | | | | |
| A. | | 0.93 | 0.34 | 36.65 | | | | | | |
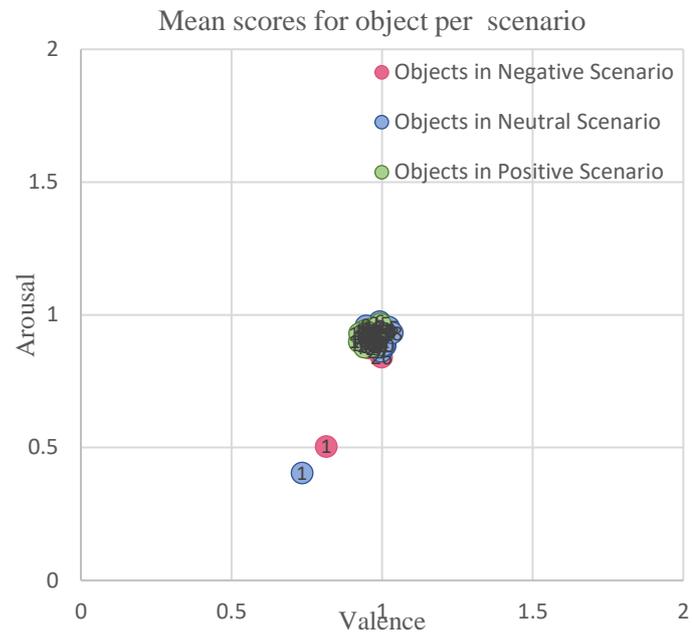
**Figure 92.** Mean scores of objects per VE scenario on the AV cartesian space.
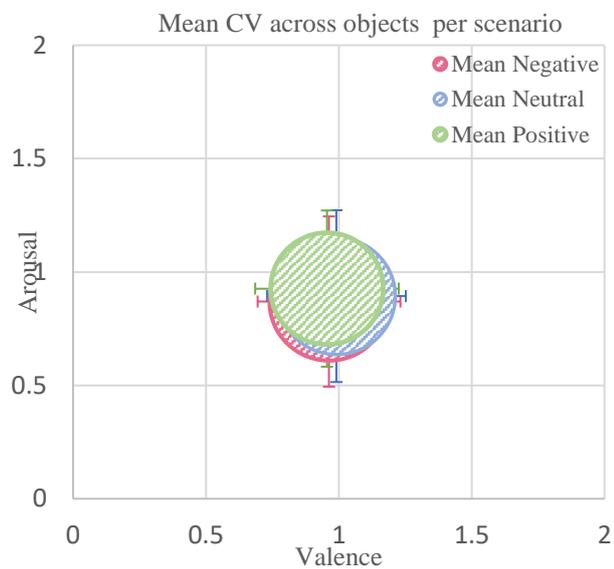
Figure 93. Mean scores of objects per scenario and their standard deviations. The mean CV score is represented by the scale of the bubble.

Chapter 5-6: Design amendments in VEs

The results and feedback obtained from this survey were used to improve and finalise the design of the VEs, including the objects/events that were present in the VEs. In particular, we decided to change the timing between certain events (additional 6 seconds were added between time-controlled events) to allow additional time for the user to inspect the room and to further differentiate the stimuli activation times. We changed the design and the animation of the 'rat' object, to make it more realistic and slightly more negative, so that it can attract the user's attention. The event could now be triggered four times instead of one, to ensure visibility by the user. The light explosion camera-effect event in the positive scenario was replaced by an event where the window opened and laughing sounds from outside the room were activated. Small lighting touches were applied in all three rooms to improve the visibility of the stimuli and to improve the shadow displays. This was achieved by added an office lamp on the desks of all three VEs and by using performance effective light baking when appropriate [445]. In the case of the negative scenario, in order to keep the dark atmosphere of the room, an additional candle in front of skull-looking sculpture was added on the second right shelf of the bookcase.

As the underwater baseline environment elicited higher levels of valence and arousal than expected, we replaced it with an office-based room, from which all potentially stimulating objects and events were excluded. That room was called 'baseline VE', containing only the bookcase the walls and the window, which we used in the main study to relax our participants and record baseline data.

To improve the positive valence ratings of the positive scene, we added a few objects. Firstly, an audio soundtrack of a baby laughing was added to the 'baby poster' object. The sound is only triggered when the participant looked directly towards the poster. The sound was implemented to be 3-D, with a centre at the baby poster when the sound is stronger, reaching towards the window and the end of the room. Its sound volume was decreasing with distance from the poster. Additionally, a picture of a goat was added on the second right self of the bookcase, which when triggered by direct gaze of the user, it would automatically play a short video of a goat laughing from an online repository. The window view was also improved by using tree images and 3D plants. A small second ball, 'pokemon ball' was added on the second desk and the initial '(beach) ball' was slightly enlarged. The remaining stimuli were kept as initially designed.

D. Appendix: Questionnaires

Demographics' Questionnaire

No. _____ Time : _____ Date: _____

Please answer to the following questions and select by ticking or circling your response where appropriate.

1. How old are you?

2. With what gender do you identify?

Prefer not to say   ☐

3. Do you speak English fluently?

Yes

No

4. What is the highest level of education you have completed?

Secondary School

Technical School (2 years)

College

Bachelor's degree

Master's degree

Doctoral degree

Professional degree (MD, JD, etc.)

Other:

5. Are you left or right-handed?

Left-handed

Right-handed

I use both hands equally well

6. Have you participated in a similar experiment before?

Yes

No

7. **On a scale from 1-10, please rate your experience with virtual reality?**

(Not at all experienced)   (Very experienced)

1   2   3   4   5   6   7   8   9   10

8. Do you suffer from any mental/psychological disorders (e.g. clinically diagnosed anxiety and depression)?

Yes   No

9. **Do you suffer from any cardiovascular conditions?**

 Yes   No

10.   **Are you aware of having any conditions that are affecting facial movements such as facial palsy or stroke?**

Yes   No

11. **Do you suffer from any of the following?**
○   excessive fear of snakes (ophidiophobia),
○   excessive fear of insects (entomophobia) or excessive fear of spiders (arachnophobia),
○   excessive fear of fire (pyrophobia),
○   excessive fear of enclosed spaces (claustrophobia),
○   excessive fear of the dark (nyctophobia)
○   intense motion sickness (nausea, sweating, vertigo, dizziness etc.) while traveling or during a Virtual reality experience – only for VR task participants

         Yes   No

If you selected "Yes" in any of the above four questions (8-11) we strongly recommend avoiding taking part in this study. *(*Please do not continue with the survey (for online survey).)*

VR Environment Survey

Note, the SAM figures, the videos, figures and skip/show question logics were omitted from this version of the questionnaire.

Q217   Participant   Agreement   Form

Title of Project: Online Survey - Evaluation of emotional responses in Virtual Reality
In this online survey you will be asked to rate how different events and/or objects that appear in three Virtual Reality environments' videos make you feel. The total time required to fill the questionnaire can range from 20-30 minutes.

In this form we ask you to confirm whether you agree to take part in the Project. You should only agree to take part in the project if you understand what this will mean for you. If you complete the rest of this form, you will be confirming to us that:
1.   1. You have read and understood the Project Participant Information Sheet and have been given access the BU Research Participant Privacy Notice which sets out how we collect and use personal information (https://www1.bournemouth.ac.uk/about/governance/access-information/data-protection-privacy)
2.   You have had the opportunity to ask questions;
3.   You understand that:
a.   Your participation is voluntary. You can stop participating in research activities at any time without giving a reason, and you are free to decline to answer any particular question(s).
b.   If you withdraw from participating in the Project, you may not always be able to withdraw all of your data from further use within the Project, particularly once we have anonymised your data and we can no longer identify you.
c.   Data you provide may be included in an anonymised form within a dataset to be archived at BU's Online Research Data Repository.
d.   Data you provide may be used in an anonymised form by the research team to support other research projects in the future, including future publications, reports or presentations.

Q218 I agree to take part in the project on the basis set out above

○   Yes (1)

○   No (2)

Q13 How old are you?
_____
Q15 How do you identify yourself?

Female (1)

Male (2)

Other (3)

Q17 Which hand do you predominantly use?

Left (1)

Right (2)

Either (ambidextrous) (3)

Q19 Do you speak English fluently?

Yes (1)

No (2)

Q21 What level of Education do you have?

Less than high school (1)

High school graduate (2)

College (3)

Degree (4)

Masters (5)

Doctorate (6)

If other, please specify (7) _____

Q23 Please rate your experience in any immersive experiences, particularly virtual reality.

	No experience (1)	Novice (2)	Average (3)	Expert (4)
Experience in immersive reality (1)				

Q25 Please rate your experience with video games

	No experience (1)	Novice (2)	Average (3)	Expert (4)
Experience in gaming (1)				

Q27 Do you suffer from any mental/psychological disorders (such as anxiety/depression)?

○　　　　Yes (1)

○　　　　No (2)

Q29 Do you suffer from chronic fatigue syndrome (extreme fatigue that doesn't go away with rest or sleep)?

○ Yes (1)

○ No (2)

Q241 Do you suffer from encephalopathy (a disease that affects the structure or function of your brain)?

○ Yes (6)

○ No (7)

Q242 Do you suffer from any cardiovascular disorders (any problems with the heart or blood vessels)?

○ No (1)

○ If yes, please be more specific: (2)

_____

Q35 Do you suffer from any of the following psychological phobias?

If yes, please give a number from 1 to 10 to indicate how much this phobia affects you (above 5 meaning a severe phobia). If no, then leave the space blank.

	Yes (1-10)
Spider phobia (Arachnophobia) (1)	
Entomophobia (excessive fear of insects) (2)	
Pyrophobia (excessive fear of fire) (3)	
Claustrophobia (excessive fear of enclosed spaces) (4)	
Nyctophobia (excessive fear of the dark) (5)	

Q37 **If you have scored higher than 5 for any of the stated phobias then please don't continue with the study.**

End of Block: demographics and phobias

Q345 **Instructions**

 Here we will add an example of what the survey looks like.

You will see a video window and underneath that, you will find 1 question.
Once you have watched the video and answered the question, you will be able to click the "-->"
or " next" button to view the next page. In the next page, you will find 2 sets of 5 figures. We
call them SAM figures.
SAM shows different kinds of feelings: Happy vs. Unhappy and Calm/Bored vs.
Excited/Aroused.
You will be asked to rate the emotion you felt using these two scales/sets of figures for the
video you just watched.

There are no right or wrong answers, so simply respond as honestly as you can.

Q347 (EXAMPLE) Please watch the entire video and report how you felt using the sliders
below.

Q348 *Please rate how you felt watching this video*

Each SAM figure varies along each side. In this illustration the first figure is unhappy
/frowning (annoyed, unsatisfied, melancholic, bored) and the last one is smiling being very
happy (*pleased, satisfied, contented, hopeful*).

If what you saw made you feel very unhappy, you can select 1 or if you feel very happy you can
select 9.

Q349 Example: Please select valence (*you can change the rating by clicking and dragging the
slider to your preference)*

		from happy/smiling				to unhappy/frowning			
	1	2	3	4	5	6	7	8	9
()									
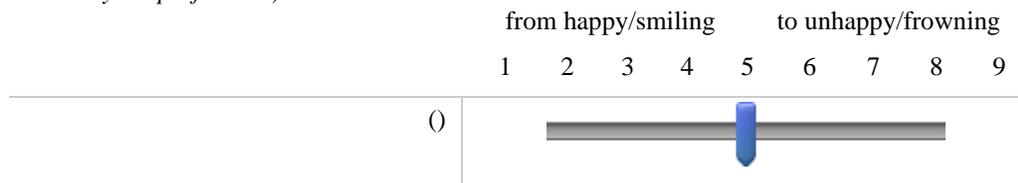
Q350 *Please rate the level of excitement*

Now you can select the excitement/intensity of the emotion you felt during the video.

This scale ranges from **calm** (**relaxed, sluggish, dull, sleepy, unaroused**) to **aroused**
(**stimulated, excited, frenzied, jittery, wide-awake**d).

The descriptions on the scales will remain for the duration of the survey.

Q351 Please select arousal

                from calm/unaroused   to excited/aroused

()

Q352 You will be asked to watch a video (4 videos in total) and rate how specific objects/events appearing in the video made you feel. Some of the videos may prompt emotional experiences; others may seem relatively neutral.

We want to remind you your rating of each picture should reflect your immediate personal experience, and no more. **Please rate each one as you actually felt while watching the video.** It is important for the completion of the survey that you <u>watch and rate all the videos</u> included in this survey. **Important events may be found in the middle or even at the end of a video clip** so please watch the whole duration of each video.

Feel free to take breaks before each video throughout the survey, however if possible please avoid taking breaks until all questions per video stimulus are answered (14 in total per video).

Page
Break

Q353 **Part 2: Main Survey**

Please take a short break before the start of the main survey to ensure you are feeling relaxed and that your computer/phone has enough battery, and your headphones are connected.

When you are ready to begin, click "yes" in the following question.

Q354 Would you like to begin?

◯   Yes (1)

Q356 Important notice: We want to remind you that the sliders for the video rating will be available after you have watched the entire video**.** Simply press the video to play and once it has ended click the "next" or "->" button at the bottom of the page to continue.

Q243 Before starting to watch the videos, we would like to adjust the audio volume for the rest of this survey to a level that it is comfortable for you. The following sound check is required in order to set an optimal volume level for you regarding the overall audio tracks included in the videos   of   this   survey.

Instructions:

1) Set your Audio volume to ~50%

2) Play the following video and adjust the volume so as to barely hear the synchronous tones.

If you cannot hear the audio of this video, please ensure that the mute switch is not enabled, the headphones or speakers are connected, and that the volume level of the YouTube player is set to 100%. We recommend a volume range between 28 - 58%.

When you complete this step, please press the "next" button to continue with the survey.

Q245 How many times did the bell ring?

○   0 (6)

○   1 (7)

○   2 (8)

○   3 (9)

○   4 (10)

○   5 (11)

Q88 Please watch the **entire video** and report how you felt using the sliders in the next page.

(BASELINE VIDEO)

Q214   Please   rate   how   you   felt   watching   this   video,   in   terms   of   valence

	from unhappy/frowning	to happy/smiling
	1 2 3 4 5	6 7 8 9
1 ()		

Q94 Please rate how you felt in terms of the level of excitement.

	from calm/relaxed	to excited/aroused
	1 2 3 4 5	6 7 8 9
1 ()		

Q206 Please watch the entire video and rate how you felt using the sliders in the next page.

*(NEUTRAL VIDEO) *This video does not contain audio*

Q208 Please rate how you felt watching this video, in terms of valence

	from uphappy/frowning	to happy/smiling
	1 2 3 4 5	6 7 8 9

1 ()

Q210 Please rate how you felt watching this video, in terms of the level of excitement

from calm/relaxed   to excited/ aroused

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |

1 ()

Page Break

(Example of questions presented for each stimulus after the video of a VE.)

Q95 The bookcase:

 (IMAGE for static objects/ SHORT VIDEO for interactive events)

Q378 Do you remember this object?

○   Yes (5)

○   No (6)

Q426   Please   rate   how   you   felt   during   this   stimulus   in   terms   of   valence:

	from unhappy/frowning	to happy/smiling
	1 2 3 4 5 6 7 8 9	
()	▬▬▬▬▬▬▼▬▬▬▬▬	

Q97   In   terms   of   the   level   of   excitement

	from calm/relaxed	to excited/aroused
	1 2 3 4 5 6 7 8 9	
1 ()	▬▬▬▬▬▬▼▬▬▬▬▬	
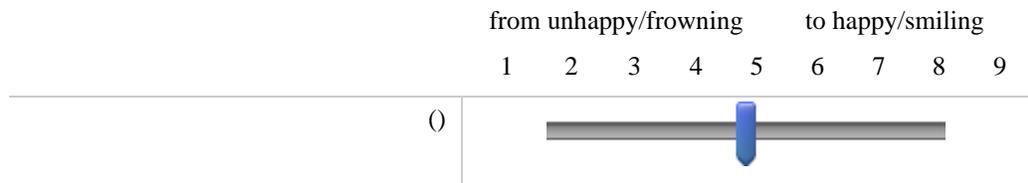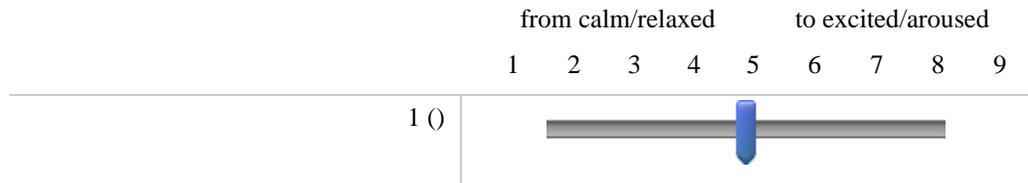
Page Break

The next section was presented after all stimuli from the affective category were rated by the participant.

Q276
We would like to evaluate the feeling of 'presence' (the feeling of actually being in the virtual space) when walking through or watching a pre-recorded video of a virtual reality environment. Please rate the following questions honestly about how you felt when watching the videos. You will notice that some questions are very similar to each other. This is necessary for statistical reasons.

Q277 Please feel free to watch the video again if you need to re-jog your memory.

(VIDEO)

(Presence questionnaires follows. Here we show an example question.)

Q247-INV1 How aware were you of the real world surroundings while watching the navigation of the virtual environment in the video (i.e. sounds, room temperature, other people, etc.)?

	Extremely aware	Moderately aware	Not aware at all

	-3	-2	-1	0	1	2	3
1 ()							

Page Break

End of Block: neutral

Stage 2 VR study Questionnaire, Science Museum - Memory, presence and sickness per VE.

Q213 Participant number:

_____

Q267 Which computer are you using? (e.g. 1, 2, 3, 4)

_____

Q37

Forest scene:

These questions will be asked after each scene to help us understand how you felt while walking through the virtual reality environment.

All questions will be on a scale of 1-9, where:

1 = Not at all

5 = Kind of

9 = Very much

Q38

Did you enjoy the VR experience?

|	Not at all					Very much			
	1	2	3	4	5	6	7	8	9
1 ()									
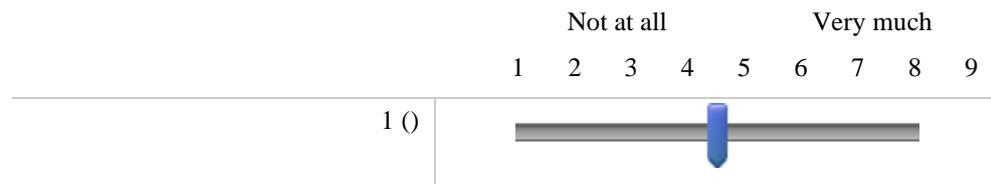
Q39

Did you have a sense of "being there"? (To which extend do you feel present in the virtual environment, as if you were really there)

Not at all					Very much			
1	2	3	4	5	6	7	8	9

1 ()
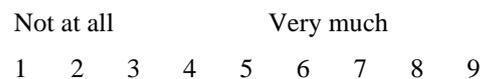
Q40

Did you feel completely captivated by the Virtual Environment?

Not at all   Very much

1   2   3   4   5   6   7   8   9

1 ()

Q41

Did you think of the virtual environment more as images you saw rather than somewhere that you visited?

Images   Place you visited

1   2   3   4   5   6   7   8   9

1 ()

Q42

Did you experience any discomfort during your experience with the system?

Not at all   Very much

1   2   3   4   5   6   7   8   9

1 ()

Q43

Did you experience dizziness, nausea or disorientation?

Not at all   Very much

1   2   3   4   5   6   7   8   9

1 ()

Q44

How difficult did you find the VR task (walking in the environment while rating)?

Not at all   Very much

1   2   3   4   5   6   7   8   9

1 ()

End of Block: forest presence + sickness

For experimenter:

Q266 During Baseline (please write if the user moved around, how intense the movement was, if they spoke, and flag any other issues related to the recording or data quality)

☐   Baseline 1 (1)

_____

☐   Baseline 2 (2)

_____

☐   Baseline 3 (3)

_____

Q212 Select Next Scene (Use randomizer per Participant)

○ Neutral (1)

○ Positive (2)

○ Negative (3)

Repeated three times, once per VE experienced. The Memory table contained the stimuli corresponding to each VE. In this example we showed the stimuli in the Negative VE.

Q214 Negative VE

1. "Could you tell me what you remember most from that scene?"
- Tick off every object as yes- very memorable

2. Ask about all other objects they did not refer to and how confident
- Tick off as slightly memorable or not memorable

	No definitely didn't see it (1)	Kind of confident (after asking) (2)	Yes very confident (on their own) (3)
Spooky mirror/skull (1)			
Fire (2)			
Fire alarm (3)			
Spider attack (4)			
Spiders in room (5)			
Flickering light (6)			
Lightening/Man outside (8)			
Glitch in viewpoint (9)			
Documents (11)			
Spilt coffee cup (12)			
Overflowing bin (13)			
Light bulb explodes (15)			
Spooky music (16)			
Rat (17)			
candle/skull (27)			
Office room (28)			
Any other mentioned (24)			
Any other mentioned (25)			
Any other mentioned (26)			

Q261 *For things not added in the list above:*

_____

Q238

**Negative**   **scene:**

Q262 Valence ratings: Could you rate how this experience made you feel on a scale of 1 - 9:

1 = completely unhappy, annoyed, unsatisfied, melancholic, bored

5 = neutral

9 = happy, pleased, satisfied, contented, hopeful

	negative	positive
	1 2 3 4 5 6 7 8 9
()	
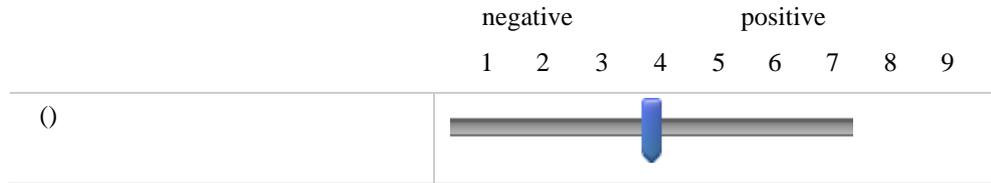
Q264 Arousal ratings: How intense was this emotion

1 = Relaxed, calm, uninterested

9 = Excited, stimulated, interested

| | 1 2 3 4 5 6 7 8 9 |
| () | |
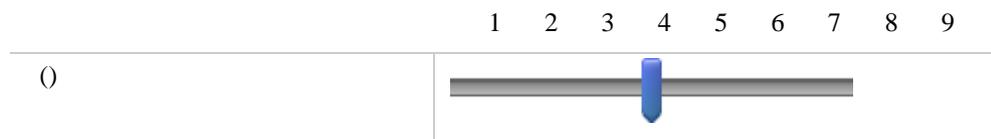
Q239 On a scale of 1-9 did you enjoy the VR experience?

	Not at all	Very much
	1 2 3 4 5 6 7 8 9
1 ()	

Q240 Presence rating: Did you have a sense of "being there"?

| | Not at all | Very much |
| | 1 2 3 4 5 6 7 8 9 |
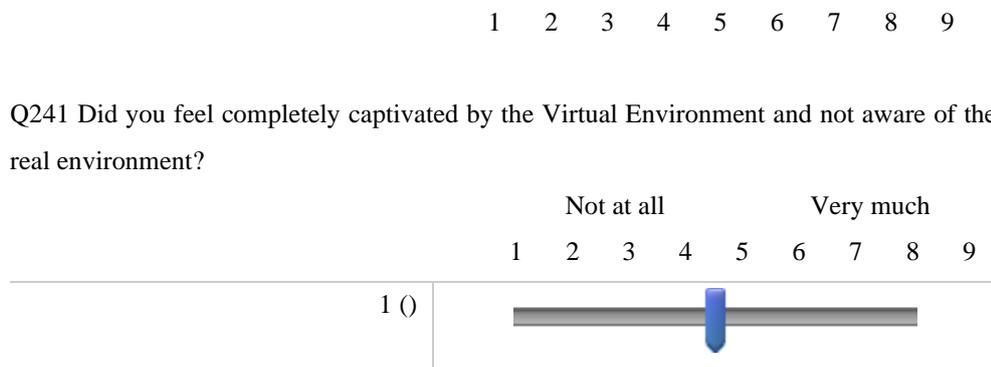
Q241 Did you feel completely captivated by the Virtual Environment and not aware of the real environment?

| | Not at all | Very much |
| | 1 2 3 4 5 6 7 8 9 |
| 1 () | |
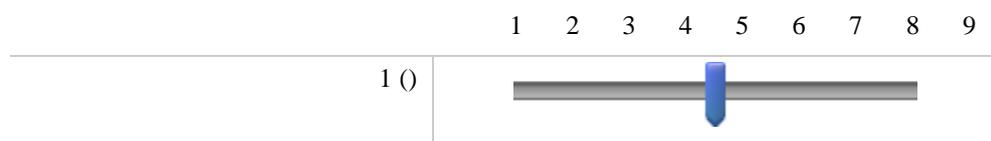
Q242 Did you think of the virtual environment more as images you saw rather than somewhere that you visited?

| | More like images | More like a place/space |
| | 1 2 3 4 5 6 7 8 9 |

1 ()	
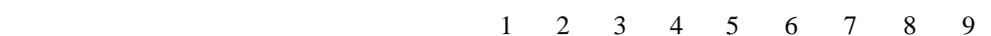
Q243 Did you experience any discomfort during your experience with the system?

	Not at all					Very much			
	1	2	3	4	5	6	7	8	9

1 ()	

Q244 Did you experience dizziness, nausea or disorientation?

	Not at all					Very much			
	1	2	3	4	5	6	7	8	9

1 ()	

Q245 How difficult did you find the VR task (walking in the environment while rating and remembering the objects/events around you)?

	Not at all					Very much			
	1	2	3	4	5	6	7	8	9

1 ()	

End of Block: new presence + sickness negative