GAME-CALIBRATED AND USER-TAILORED REMOTE DETECTION OF EMOTIONS
A non-intrusive, multifactorial camera-based approach for detecting stress and boredom of players in games
DOCTORAL DISSERTATION

GAME-CALIBRATED AND USER-TAILORED REMOTE DETECTION OF EMOTIONS

A non-intrusive, multifactorial camera-based approach for detecting stress and boredom of players in games

FERNANDO BEVILACQUA
Informatics

UNIVERSITY OF SKÖVDE
Fernando Bevilacqua, 2018

Doctoral Dissertation

*Title:* Game-calibrated and user-tailored remote detection of emotions

A non-intrusive, multifactorial camera-based approach for detecting stress and boredom of players in games

University of Skövde, Sweden
www.his.se

*Printer:* BrandFactory AB, Gothenburg

ISBN 978-91-984187-9-8
Dissertation Series No. 27 (2018)
To Marilia, whose love, dedication, and courage are beyond words.
“A prática é tudo.”
—Edson Arantes “Pelé” do Nascimento
ABSTRACT

Questionnaires and physiological measurements are the most common approach used to obtain data for emotion estimation in the field of human-computer interaction (HCI) and games research. Both approaches interfere with the natural behavior of users. Initiatives based on computer vision and the remote extraction of user signals for emotion estimation exist, however they are limited. Experiments of such initiatives have been performed under extremely controlled situations with few game-related stimuli. Users had a passive role with limited possibilities for interaction or emotional involvement, instead of game-based emotion stimuli, where users take an active role in the process, making decisions and directly interacting with the media. Previous works also focus on predictive models based on a group perspective. As a consequence, a model is usually trained from the data of several users, which in practice describes the average behavior of the group, excluding or diluting key individualities of each user. In that light, there is a lack of initiatives focusing on non-obtrusive, user-tailored emotion detection models, in particular regarding stress and boredom, within the context of games research that is based on emotion data generated from game stimuli. This research aims to fill that gap, providing the HCI and the games research community with an emotion detection process that can be used to remotely study user’s emotions in a non-obtrusive way within the context of games.

The main knowledge contribution of this research is a novel process for emotion detection that is non-obtrusive, user-tailored and game-based. It uses remotely acquired signals, namely, heart rate (HR) and facial actions (FA), to create a user-tailored model, i.e. trained neural network, able to detect the emotional states of boredom and stress of a given subject. The process is automated and relies on computer vision and remote photoplethysmography (rPPG) to acquire user signals, so that specialized equipment, e.g. HR sensors, is not required and only an ordinary camera is needed. The approach comprises two phases: training (or calibration) and testing. In the training phase, a model is trained using a user-tailored approach, i.e. data from a given subject playing calibration games is used to create a model for that given subject. Calibration games are a novel emotion elicitation material introduced by this research. These games are carefully designed to present a difficulty level that constantly and linearly progresses over time without a pre-defined stopping point. They induce emotional states of boredom and stress, accounting for particularities at an individual level. Finally, the testing phase occurs in a game session involving a subject playing any ordinary, non-calibration game, e.g. Super Mario. During the testing phase, the subject’s signals are remotely acquired and fed into the model previously trained for that particular subject. The model subsequently outputs the estimated emotional state of that given subject for that particular testing game.

The method for emotion detection proposed in this thesis has been conceived on the basis of established theories and it has been carefully evaluated in experimental setups. Results show a statistical significance classification of emotional states with a mean accuracy of 61.6%. Finally, this thesis presents a series of systematic evaluations conducted
in order to understand the relation between psychophysiological signals and emotions. Facial behavior and physiological signals, i.e. HR, are analyzed and discussed as indicators of emotional states. This research reveals that individualities can be detected regarding facial activity, e.g. an increased number of facial actions during the stressful part of games. Regarding physiological signals, findings are aligned with and reinforce previous research that indicates higher HR mean during stressful situations in a gaming context. Results also suggest that changes in HR during gaming sessions are a promising indicator of stress. The method for the remote detection of emotions, presented in this thesis, is feasible, but does contain limitations. Nevertheless, it is a solid initiative to move away from questionnaires and physical sensors into a non-obtrusive, remote-based solution for the evaluation of user emotions.
SAMMANFATTNING


Metoden för känslomätning som föreslås i denna avhandling är baserad på tidigare etablerade teorier och har även blivit utvärderad i en serie kontrollerade experiment. Resultat från utvärdering visar att det finns en statistiskt signifikant identifiering av känslotillstånd med en precision på 61,6%. Utöver presentationen av det framtagna verktyget för känslomätningar presenteras även en serie av systematiska utvärderingar av förhållandet mellan psykofysiologiska signaler och känslor. Användning av ansiktsmusklér och fysiologiska signaler (till exempel HR) analyseras och deras roll som indikatorer...
på känslotillstånd diskuteras. Denna forskning visar att individuella egenheter i människors ansiktsuttryck kan identifieras (till exempel ökad mängd och intensitet av olika ansiktsuttryck under stressframkallande spelsegment). Angående fysiologiska signaler är studieresultaten förenliga med, och styrker, tidigare forskning som drar paralleller mellan HR och stresskänslor i spelsituationer. Metoden för fjärmätning av känslotillstånd som presenteras i denna avhandling är användbar, men har vissa begränsningar. Oavsett detta är metoden ett lovande första steg bort från användning av frågeformulär och fysiskt påträngande sensorer och mot fjärrsamlingsbaserade lösningar för utvärdering av användares känslotillstånd.
ACKNOWLEDGMENTS

When I was a kid, I thought every person living on Earth should speak the same language. In my young self’s mind, it made perfect sense, because things would be so much easier for everybody. Communication is key after all. As I grew up, however, I realized that language is one of the many cultural aspects that makes us who we are. Humans are unique creatures that are full of dreams, fears, and experiences. I honestly believe that learning from experiences is essential, it can transform the way we live and work. That idea echoed in my mind for years, until the day I was gifted with the opportunity of doing a PhD while being sponsored. I did everything in my reach to make this happen away from the land I call home, Brazil. My homeland has great PhD programs, but I wanted to see things using new cultural lenses. Lenses that are different from the ones I was given at my birthplace. I wanted to work following a mindset that I never experienced before, using a language that is not my own. The PhD portrayed in this book was a life-changing journey for me. I have learned about many technical topics, but I have certainly learned about different cultures and ideas. Before doing my PhD, I had never set foot in Sweden, but now I will never forget the day I landed in Scandinavia. Finally, I must say that my PhD journey was certainly not an isolated event. It was the culmination of several steps that were laid out throughout years and were influenced by many people. I could fill pages of this thesis with names I would like to thank, even knowing that I incur the danger of leaving someone out. Below is my tentative of saying thank you to all those people, in no particular order.

A heartfelt, warm and special thank you to Marília Landerdahl Abreu, my wife, who stood with me through good and bad moments in life. Your love, dedication and support are always paramount in my life. I will never forget your courage and sacrifices towards this PhD. Thank you for accompanying and helping me in this great journey, which was unknown, far and so different from what we were used to. I am very thankful for having you and your love by my side. My love for you gives meaning to even the smallest things.

I would like to state a huge thank you to two persons that directly guided my work and contributed to its completion: my supervisors Per Backlund and Henrik Engström. A significant part of having a great experience during a PhD is all about good supervision, which I had plenty of. My supervisors not only encouraged me, refined my ideas, respected my opinions and allowed me to become academically independent, they have taught me how to do proper research. They worked beyond the duties of supervision to ensure I had a great staying in Sweden, showing me places to visit, things to do and try. I also need to mention how well they helped me deal with many situations outside my PhD that impacted my work somehow, including my doubts and anxieties.

I would like to express my appreciation and gratitude towards the Federal University of Fronteira Sul (UFFS) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), the two Brazilian government institutions that funded my PhD. UFFS,
my employer, allowed me to pursue a PhD while continue paying my regular salary in full, even if I was not there to fulfill my duties. UFFS is a young institution with less than 10 years of history when I left for my PhD, but it never prevented the investment and trust in me. I also want to thank my UFFS colleagues and faculty members that in any form helped me with my PhD. I extend my thank you to CNPq, which also deposited trust in me by approving my project for a four-year scholarship. The funding paid for all my travel expenses to move in and out of Sweden, as well as complemented my financial support there. Finally, I want to thank the University of Skövde and the EU Interreg ÖKS project Game Hub Scandinavia for sponsoring my activities as well.

Thank you to Björn Berg Marklund for the friendship and for welcoming me into his family. Your efforts helped me find the way in the Swedish system in several matters. I have learned so much about the Swedish culture while having an awesome time barbecuing, game jamming, playing Contra III, petting Digby, or having a nice day by the lake (among many others). Thank you to Adriano Moreira Costa and his family for the friendship and great moments together. I appreciate you filming my spectacular falls while (trying) to ice skate, having dinners or helping me when heavy clouds were upon. As I once said, it is amazing to find the best Brazil has to offer outside Brazil itself. Finally, thank you to Denio Duarte and his family for the friendship and great moments together as well. I bet we will never forget the −35°C experienced in north Sweden, or the walks in the woods of Varola.

I had the pleasure of working and sharing moments with great people from my department and from the university. Thank you all! Your help towards me, my studies and my experiments was invaluable. A special thank you to Torbjörn Svensson for helping with cameras and lights, as well as for welcoming me in such a nice way right after my arrival in Sweden. I will not forget the tour around Varnhem and my very first Swedish fika shared with your lovely family. Thank you also to Marcus Toftedahl for all the help, including finding subjects for my experiment, gearing up my data processing, or teaching me how to cook mushrooms. I also want to thank Katrin Dannberg for all the time you put in helping me with my experiment, in particular spreading the word about my study. I could never have found so many subjects without your help. I enjoyed all the Swedish you have taught me, even if I managed to learn just a few words. Thank you to Mikael Lebram for helping me gear up my experiments and for showing me (and let me try) so many cool projects, like the car simulator. Finally, thank you to Andrea Dião Jonsson for helping me with the paperwork I had to go through to make this PhD happen, for helping me land in Sweden with a place to stay, and for all the help regarding my experiments.

I want to acknowledge and thank the members of my defense committee and opponents who invested their time and expertise to ensure the quality of this work. I also want to note the contributions from those who evaluated my work since its proposal until its final form: a big thank you to David Vernon, Mikael Johannesson, Stefan Seipel, and Veronica Sundstedt for the invaluable feedback regarding my research. Your comments helped me steer my research towards fruitful venues, which led me to produce a better piece of work. I want to thank as well all those teaching in the courses I attended at the University of Skövde. I have learned a great deal in those classes regarding a variety of topics and methods. Particularly, I want to thank Jeremy Rose for the essential help when I was selecting theories to support my thesis, and also for lowering my own positivist and quantitative walls, allowing me to see the bright light of the philosophy
of science. Thank you to Cesar Tadeu Pozzer, who made me aware of the University of Skövde and its PhD program years after being my supervisor during my masters. You provided me with essential information to enroll in the program and to move to Sweden, igniting this whole journey. Finally, I want to thank my professors at UFSM, particularly Andrea Schwertner Charão and Marcos Cordeiro d’Ornellas. You all prepared me for the great challenges ahead of any undergraduate program, which helped me conduct my PhD work several years after we have crossed paths.

Thank you to Antonieta, Delcio, and Gabriele Bevilacqua, my mother, father, and sister. You always surrounded me with love and were there for me in regard to all sorts of matters, including education and personal development. I would never be here without your solid foundation. Life has winds of its own and we sometimes face difficult and heavy decisions, but I am sure we all have learned from those. I appreciate your love and thoughts towards me.

Doing a PhD can be a lonely experience, but thankfully I had an awesome group of fellow PhD students around me. You all helped me advance my own studies in a way or another, be it with a quick talk, a word of advice, time to participate in my experiments, discussing ideas, or by simply sharing common struggles so I could put my own problems into perspective. A huge thank you to all PhD students at Portalen from the second to the fifth floor. A particular thank you to the guys I regularly met (at 3 ± 5) during the fika time on the fifth floor: Elio Ventocilla, Navoda Senavirathne, Niclas Ståhl, András Márki, and Nikolas Huhnstock. Sharing joy, our daily fights, or comparing the size of a sequoia tree to a building were important moments for me to keep things grounded and sane.

Finally, I want to make a special note of appreciation for those who direct or indirectly contributed to my work: all participants of my experiments, Vera Lindroos for proofreading this thesis, Thomas Fischer for creating and helping me with the \LaTeX{} template that typesets this thesis, Radu Dinu, Elin Tomasdottir, Anette Andersson, Melissa Mozifian, the authors of the open source software I used, and my SFI teachers: Marianne Persson, Patrik Fredén de los Rios and Ulf Nilsson. Last, but in absolutely no circumstance least, thank you to Espresso House and its great coffee and friendly staff. Your coffee fueled my work and this thesis to the point I didn’t even have to specify my order to your baristas; they already knew what I wanted. If this thesis smells anything, it smells “a salted caramel latte with cream to take away, please”.

This work has been performed with support from: CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil; University of Skövde; EU Interreg ÖKS project Game Hub Scandinavia; UFFS, Federal University of Fronteira Sul.
PUBLICATIONS

During this research project, a number of manuscripts with varying relevance to the core aims of this thesis were published.

PUBLICATIONS WITH HIGH RELEVANCE

PAPER I

Fernando Bevilacqua, Per Backlund, and Henrik Engström (2015). “Proposal for Non-Contact Analysis of Multimodal Inputs to Measure Stress Level in Serious Games.” In: 2015 7th International Conference on Games and Virtual Worlds for Serious Applications (VS-Games). IEEE. Institute of Electrical & Electronics Engineers (IEEE), pp. 1–4. DOI: 10.1109/vs-games.2015.7295783

Author’s contribution: conception and design of the study, acquisition of data, analysis and interpretation of data, drafting and writing the article.

PAPER II

Fernando Bevilacqua, Per Backlund, and Henrik Engström (2016). “Variations of Facial Actions While Playing Games with Inducing Boredom and Stress.” In: 2016 8th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES). IEEE. Institute of Electrical and Electronics Engineers (IEEE), pp. 1–8. DOI: 10.1109/vs-games.2016.7590374

Author’s contribution: conception and design of the study, acquisition of data, analysis and interpretation of data, drafting and writing the article.

PAPER III


Author’s contribution: conception and design of the study, acquisition of data, analysis and interpretation of data, drafting and writing the article.
PAPER IV
Author’s contribution: conception and design of the study, acquisition of data, analysis and interpretation of data, drafting and writing the article.

PAPER V
Author’s contribution: conception and design of the study, acquisition of data, analysis and interpretation of data, drafting and writing the article.

INDIRECTLY RELATED PUBLICATIONS
PAPER VI
Author’s contribution: interpretation of data, drafting the article or revising it critically for important intellectual content.

PAPER VII
Author’s contribution: drafting the article or revising it critically for important intellectual content.
CONTENTS

I. Introduction and methodology

1 Introduction
   1.1 Problem specification ......................................................... 5
   1.2 Research aim ................................................................. 6
   1.3 Knowledge contributions .................................................. 7
   1.4 Thesis overview and structure .......................................... 11
       1.4.1 Text organization ..................................................... 13
       1.4.2 Definitions and scope ............................................... 14

2 Research methodology .................................................. 17
   2.1 Ethics and privacy in this research .................................. 20

II. Literature review .................................................. 23

3 Games and emotions .............................................. 25
   3.1 Emotions theory ............................................................ 26
   3.2 Immersion, engagement and sense of presence .................... 27
   3.3 Instruments for assessment of emotions ............................. 28

4 Emotions and facial analysis .............................................. 31
   4.1 Facial analysis based on sensors .................................... 32
   4.2 Facial analysis based on computer vision ............................ 32
       4.2.1 Constrained Local Model ....................................... 34
       4.2.2 Cascaded Regression ............................................... 35
   4.3 Facial-based emotion detection ....................................... 36
   4.4 Summary ................................................................. 39

5 Emotions and physiological signals .............................................. 41
   5.1 Physiology of heart rate ............................................... 42
   5.2 Heart rate, stress and frustration .................................. 43
10.2.1 Facial behavior and emotions .............................................. 127
10.2.2 Physiological activity and emotions ..................................... 127

11 Software for emotion detection .................................................. 129
  11.1 Overall structure .................................................................. 129
    11.1.1 Face detector ................................................................. 130
    11.1.2 Face analyzer ................................................................. 130
    11.1.3 Signal estimator .............................................................. 131
    11.1.4 Report manager .............................................................. 131
    11.1.5 Emotion model and estimator .......................................... 132

12 Ethics and privacy ...................................................................... 133
  12.1 Ethical use of technology ....................................................... 133
  12.2 Privacy and personal data ...................................................... 134
  12.3 Ethical and privacy implications of this research ...................... 135

13 Limitations and critique ............................................................. 139

V. Closing remarks ......................................................................... 143

14 Conclusion .................................................................................. 145
  14.1 Fulfillment of research objectives .......................................... 145
  14.2 Answering the research question ........................................... 147
  14.3 Closing remarks .................................................................... 148

15 Future work ................................................................................. 149

References ....................................................................................... 155

Publications in the Dissertation Series ............................................. 169
LIST OF FIGURES

1.1 General structure of the proposed method for remote detection of stress and boredom levels of players during their interaction with games. .......................... 9

1.2 Calibration phase composed of emotion elicitation games (calibration games) and remote acquisition of signals from the user. The result of this phase is a user-tailored model applied to detect emotions. ........................................ 9

1.3 Progression of the level of difficulty of a calibration game over time along with the corresponding variations of the emotional states of stress and boredom experienced by the user................................................................. 10

1.4 Emotion estimation phase. Remotely acquired signals from the player are fed into a user-tailored model that outputs the stress/boredom levels of the player during the interaction with an ordinary game........................................... 10

1.5 Overview showing how each chapter covers the work towards the achievement of the research objectives, i.e. O1 to O6 ................................................. 11

1.6 Different fields involved in this research. Main contribution is in the area of Games Research. ...................................................................................... 14

2.1 Design science research process model. Adapted from Vaishnavi and Kuechler (2015). ......................................................................................... 17

2.2 Dependency among the main parts of this research: emotion elicitation, acquisition of user signals and emotion estimation. ............................. 18

2.3 The Generate/Test cycle. Adapted from Hevner et al. (2004) ..................... 19

3.1 Repeating cycle of increasing challenge followed by a reward, keeping the player in the flow zone. Reproduced from Schell (2014). ............... 25

3.2 Eight channel model of flow. Reproduced from Nakamura and Csikszentmihalyi (2014). ............................................................................................ 26

3.3 Representation of the Circumplex Model of Affect. Horizontal axis represents the valence dimension and the vertical axis represents the arousal or activation dimension. Reproduced from Posner, Russell, and Peterson (2005). .................. 27


3.5 Visual representation of the Affective Slider. Reproduced from Betella and Verschure (2016). .................................................................................. 29
4.1 Facial muscles. (a) Corrugator supercili. (b) Orbicularis oculi. (c) Zygomaticus minor. (d) Zygomaticus major. Adapted from "Sobotta’s Atlas and Text-book of Human Anatomy", by Dr. Johannes Sobotta (Illustration: K. Hajek and A. Schmitson), 1909. Reproduced from (Wikimedia Commons, 2013). 33

4.2 Example of face alignment. From left to right: input image, detected face, and aligned face. 34

4.3 Shape models of CLM. (a) Configuration of the shape models with different variations. (b) Shape models and their respective entries in the texture model. Reproduced from Yu (2010). 34


4.7 Distance-based facial feature descriptors. Left: detected facial landmarks. Right: highlight of the distance between facial landmarks. Reproduced from Samara et al. (2016). 38

5.1 Two cardiac cycles with highlights on each of its electrical waves. The X axis represents time in milliseconds and the Y axis represents the wave amplitude. A heartbeat is connected to the Q, R and S waves, known as the QRS complex. Reproduced from Ahmed, Begum, and Islam (2010). 42

6.1 General structure of a physical photoplethysmographic system. Adapted from Chwyl (2016). 45

6.2 General algorithm framework common to all rPPG techniques. Adapted from Rouast et al. (2016). 46

6.3 Overall structure of rPPG approach based on ICA. Adapted from Poh, McDuff, and Picard (2010). 49

7.1 Non-obtrusive and remote approach based on multifactorial analysis to identify user emotion. Reproduced from D. Zhou et al. (2015). 54

7.2 Highlight of distance calculation regarding eye aperture. Reproduced from Giannakakis et al. (2017). 56

8.1 Experiment setup. On the left, an image illustrating the position of the equipment, including the angle of the external light source. On the right, an image highlighting the position and angle of the video camera. 60

8.2 Mushroom (left), Platformer (center) and Tetris (right). In Mushroom, player has to drag and drop the correct mushrooms into the character, discarding the wrong ones into the trash. In Platformer, the player has to jump over or slide under obstacles while collecting hearts. In our version of Tetris, there are no hints about the next piece to be added to the screen. 63
8.3 Annotated facial actions (FA). (a) Smile not showing teeth; (b) Smile showing teeth; (c) Lip pucker; (d) Lip stretcher; (e) Lip suck; (f) Lip pressor; (g) Lips parted; (h) Tongue touching lips; (i) Mouth movement right; (j) Mouth movement left; (k) Lower lip bite; (l) Frown; (m) Brow raiser; (n) Lid tightener; (o) Brow lowering.

8.4 Statistical correlation of $HR_{gt}$ and $HR_{video}$ applied to the video segments of each game, as well as to the video segments of all games.

8.5 Distribution of values of $M_e$ for all games. The x-axis represents intervals of values of $M_e$ while the y-axis represents the percentage of subjects that presented an estimation error within the interval informed in the x-axis.

8.6 Distribution of values of RMSE and $M_eRate$ for all games. The x-axis represents intervals of values of RMSE or $M_eRate$ while the y-axis represents the percentage of subjects that presented an estimation error within the interval informed in the x-axis.

8.7 Variations of distance of the ROI central position for subjects 17 (low estimation errors), 3 (moderate to high estimation errors) and 1 (high estimation errors) during their gaming sessions. Values were subtracted from session mean to facilitate analysis and comparison among different games/subjects.

8.8 Variations of the ROI diagonal length for subjects 17 (low estimation errors), 3 (moderate to high estimation errors) and 1 (high estimation errors) during their gaming sessions. Values were subtracted from the session mean to facilitate analysis and comparison among different games/subjects.

8.9 Examples of body movement and facial activity during gaming sessions. (a) Partial face occlusion by subject’s hand; (b) Head tilt and movement during laugh action.

8.10 Facial landmarks and features. (a) Highlight of 68 detected facial landmarks. (b) Visual representation of the facial features.

8.11 Extraction of video segments $H_0$ and $H_1$ containing boring and stressful game interactions, respectively. Initial $D$ seconds of any video $V_{s,i}$ are ignored and the remaining is divided into three pieces, from which the first and the last ones are selected. Stripes highlight discarded video segments.

8.12 Iteration of a 3-fold Leave-One-Session-Out Cross-Validation performed on the gaming session of a given subject with 3 games, i.e. A, B and C. Data of two calibration games, e.g. A and B, are used to train the machine learning model, while data of the third calibration game, e.g. C, are used to evaluate the model.

9.1 Experiment setup. (a) Position of equipment, showing computer, camera, and external light source. (b) Highlight of the video camera and its angle.

9.2 Two (uninterrupted) parts of the experiment. (a) Calibration part; (b) Testing part. G: calibration game, Q: questionnaire about game/level, R: resting, ABC: levels of Infinite Mario, E: demographic questionnaire.

9.3 Screenshots from Infinite Mario. From left to right, level types Overground, Underground and Castle, respectively.

9.4 Extraction of video segments $H_0$ and $H_1$ containing boring and stressful interactions, respectively, in the calibration games. Initial $D = 45$ seconds of any video $C_{i,g}$ are ignored and the remainder is divided into three segments, from which the first and the last ones are selected. Stripes highlight discarded video segments.
9.5 Training and evaluation of a user-tailored emotion classifier. (a) Training of the emotion classifier; (b) Constructions of a testing dataset; (c) Evaluation of the emotion classifier.

9.6 Distribution of accuracy values at a subject level. (a) Histogram showing the number of subjects and the accuracy rate obtained in their emotion classification. (b) Density curve regarding the distribution of accuracy values.

11.1 Software developed as a partial instantiation of the proposed method working on a video file. Detected facial points and eye gaze information are superimposed on each frame of the video.

11.2 Overall structure of the software and its components.

11.3 Visual representation of the 68 points detected as facial landmarks. Blue circle represents the average position (center of mass) of all detected facial landmarks.

11.4 Visualization of data provided by the face analyzer. Available information: FACS AUs, detected face, motion and instability of the face, and variations of eye/mouth areas over time.

11.5 Visualization of data provided by the signal estimator. Available information: most prominent frequencies detected, photoplethysmographic signal, estimated HR, and the ROI used in the estimation.
LIST OF TABLES

4.1 Categorization of facial elements connected with stress and anxiety according to Giannakakis et al. (2017) ................................................................. 32

5.1 Most common psychophysiological measurements used in human interaction studies (Jerritta et al., 2011) ................................................................. 41

6.1 Frequency resolution $\Delta f$ (in Hz) as a function of window size (in samples) and frame rate (in FPS) (Roald, 2013) ................................................................. 51

8.1 Number of FA annotations made for all subjects during periods $H_0$ and $H_1$ of the games ........................................................................................................ 65

8.2 Subject-based frequency of FA that appeared in the same period of all three games 67

8.3 Values of $V_{g,t}^{g,t}$, the relativized HR mean coefficient, for all subjects ($s$) in a given game $g$ (M is for Mushroom, P for Platformer and T for Tetris), grouped by intervals ($t$) of 1 minute ........................................................................................................ 70

8.4 Mean of the differences of $V_{g,t}^{g,t}$ at the periods $t = 1$ (second minute of gameplay) and $t = n$ (last minute of gameplay), for all subjects in each game ($g$). Values in bpm (beats per minute). Significance was tested with a one-tailed paired t-test 71

8.5 Mean of the differences of $V_{g,t}^{g,t}$ at key periods, for all subjects in a given game $g$. Values in bpm (beats per minute) ................................................................. 71

8.6 Performance of the rPPG technique applied to the testing set ................. 75

8.7 Accuracy measurements of the rPPG technique when applied to the video segments of a given game and of all games ................................................................. 76

8.8 Information regarding calculated facial features ........................................ 84

8.9 Mean of differences ($\pm$SD) of features between periods $H_0$ and $H_1$ ($N = 59$). Units expressed in normalized pixels ................................................................. 90

8.10 Percentage of change of features from period $H_0$ to $H_1$ in the Mushroom game ($N = 20$). ................................................................. 92

8.11 Percentage of change of features from period $H_0$ to $H_1$ in the Platformer game ($N = 19$). ................................................................. 93

8.12 Percentage of change of features from period $H_0$ to $H_1$ in the Tetris game ($N = 20$). ................................................................. 94

8.13 Description of features used for classification ........................................ 96

8.14 Tests and their respective feature sets ..................................................... 99
8.15 Mean values of resulting classification metrics ............................... 100
8.16 Minimum and maximum mean values of resulting classification metrics ...... 100

9.1 Levels of Infinite Mario and adjustments made to induce a particular emotional state. ........................................................................................................ 108
9.2 Mean value of the answers given in the self-reported emotional state questionnaire after levels of Infinite Mario ........................................ 114
9.3 Mean values of resulting classification metrics ................................. 115
**LIST OF ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAM</td>
<td>Active Appearance Model</td>
</tr>
<tr>
<td>ANS</td>
<td>Autonomic Nervous System</td>
</tr>
<tr>
<td>AUC</td>
<td>Area Under the Curve</td>
</tr>
<tr>
<td>BP</td>
<td>Blood Pressure</td>
</tr>
<tr>
<td>bpm</td>
<td>beats per minute</td>
</tr>
<tr>
<td>BSS</td>
<td>Blind Source Separation</td>
</tr>
<tr>
<td>BVP</td>
<td>Blood Volume Pulse</td>
</tr>
<tr>
<td>CLM</td>
<td>Constrained Local Model</td>
</tr>
<tr>
<td>CLNF</td>
<td>Constrained Local Neural Fields</td>
</tr>
<tr>
<td>CMA</td>
<td>Circumplex Model of Affect</td>
</tr>
<tr>
<td>COM</td>
<td>Center of Mass</td>
</tr>
<tr>
<td>COTS</td>
<td>Commercial Off-the-shelf</td>
</tr>
<tr>
<td>DSR</td>
<td>Design Science Research</td>
</tr>
<tr>
<td>ECG</td>
<td>Electrocardiogram</td>
</tr>
<tr>
<td>EEG</td>
<td>Electroencephalogram</td>
</tr>
<tr>
<td>EMG</td>
<td>Electromyography</td>
</tr>
<tr>
<td>ERT</td>
<td>Ensemble of Regression Trees</td>
</tr>
<tr>
<td>FA</td>
<td>Facial Actions</td>
</tr>
<tr>
<td>FACS</td>
<td>Facial Action Coding System</td>
</tr>
<tr>
<td>FPS</td>
<td>Frames Per Second</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>GSR</td>
<td>Galvanic Skin Response</td>
</tr>
<tr>
<td>HR</td>
<td>Heart Rate</td>
</tr>
<tr>
<td>HRV</td>
<td>Hear Rate Variability</td>
</tr>
<tr>
<td>IBI</td>
<td>Inter Beat Interval</td>
</tr>
<tr>
<td>ICA</td>
<td>Independent Component Analysis</td>
</tr>
<tr>
<td>LOOCV</td>
<td>Leave-One-Out Cross-Validation</td>
</tr>
<tr>
<td>LOSOCV</td>
<td>Leave-One-Session-Out Cross-Validation</td>
</tr>
<tr>
<td>PNS</td>
<td>Parasympathetic Nervous System</td>
</tr>
<tr>
<td>PPG</td>
<td>Photoplethysmography</td>
</tr>
<tr>
<td>RGB</td>
<td>Red Green Blue</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root Mean Square Error</td>
</tr>
<tr>
<td>ROI</td>
<td>Region of Interest</td>
</tr>
<tr>
<td>rPPG</td>
<td>Remote Photoplethysmography</td>
</tr>
<tr>
<td>RR</td>
<td>Respiratory Rate</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-noise Ratio</td>
</tr>
<tr>
<td>SNS</td>
<td>Sympathetic Nervous System</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
</tbody>
</table>
PART I

INTRODUCTION AND METHODOLOGY
CHAPTER 1
INTRODUCTION

In human-computer interaction (HCI) research, the study of the relation between users and systems is of interest. Within the context of games research in particular, the relation between player and game is an important topic. Such a relation comprehends concepts such as engagement and immersion (Boyle et al., 2012) and the investigation of the elements that influence those concepts. Researchers and practitioners benefit from tools applied to perform the aforementioned investigations, such as the ones illustrated in the following hypothetical scenarios.

**Scenario 1:** a games researcher wants to investigate the stress level of a user during a training session with a serious game. The researcher points an ordinary camera at the user’s face and asks him/her to play a few games for 15 minutes for calibration purposes. Thereafter, the researcher records a video of the user’s face while he/she plays the serious game to be investigated. The user experience is not disturbed by inconvenient sensors attached to the body, nor is the user constantly interrupted during the game play to answer questionnaires. After the user finishes playing, a software shows a report informing the researcher about the stress levels throughout the session. On another occasion, the researcher asks the same user to play a different serious game being investigated. This time the researcher skips the calibration phase because the profile of that user is already known (no re-calibration phase is needed). Once again the researcher points a camera at the user’s face, films the gaming session and, at the end, the software reports the stress levels.

**Scenario 2:** a small, game developer company wants to check if a new title to be released is well balanced, i.e. not too difficult nor too easy to play. The small company has several hours of video recordings of users who have play-tested the game, however there is no budget or the time to manually inspect the material in order to find useful information. A representative of the company then invites the users involved in the play-testing sessions to visit the company again. A new video of each user playing a few calibration games for approximately 15 minutes is recorded. Subsequently, the company representative feeds a computer software with the newly created user videos and the already existing videos with hours of gameplay. In minutes, all the material is analyzed and the software indicates the points in time where the stress level of the users was higher than their usual behavior. The company then inspects the problems and adjusts the game accordingly, increasing its chances of success.

These scenarios illustrate the investigation work-flow that researchers and practitioners apply when a novel process which is the research aim of this thesis is used. Currently, researchers perform such investigations by relying on obtrusive and cumbersome methods in order to be able to capture the user’s emotional state. The most common techniques used to obtain data regarding the emotional states of players in a game are self-reports (questionnaires) and physiological measurements (Mekler et al., 2014). Although questionnaires are practical and easy to use tools, they require a shift in attention, hence breaking or affecting the level of engagement/immersion of the user. Physiological signals, on the other hand, have been used to obtain information from users without caus-
ing interruptions (Bousefsaf, Maouei, and Pruski, 2013b; Yun et al., 2009; Rani et al., 2006; Tijs, Brokken, and IJsselsteijn, 2008). Such tools as sensors, despite avoiding interruptions, are usually perceived as uncomfortable and intrusive, since they need to be properly attached to various parts of the user’s body. Additionally, sensors might restrict a player’s motion abilities, e.g. a sensor attached to a finger prevents the use of that finger. Sensors also increase a user’s awareness of being monitored (Yamakoshi et al., 2007; Yamaguchi, Wakasugi, and Sakakima, 2006; Healey and Picard, 2005), which affects the results of an investigation.

Despite these problems, sensors continue to be used because there is a significant amount of information that can be read from the human body, such as heart rate (HR), respiratory rate (RR), facial expressions, among others. Such information in the human body can be regarded as input signals for emotion estimation. A number of studies (Kukolja et al., 2014) suggest that the analysis of a combination of different input signals, known as multimodal or multifactorial analysis, is more likely to produce accurate results when mapping emotional states. Physiological signals, e.g. HR, are considered reliable sources since they are hard to fake (because of their link to the central nervous system), as opposed to facial expressions (Landowska, 2014), for instance. When combined in the same analysis, however, such signals can complement each other and provide more information about emotional states. The process of mapping such signals to an emotional state, however, is a significant task. It involves testing/defining what are the possible emotional states a person can experience (Mandryk, Atkins, and Inkpen, 2006), as well as comparing which signals are better predictors of such states (Jerritta et al., 2011). A common approach used to perform the mapping between input signals and emotional states is the application of machine learning models.

The use of a machine learning model commonly starts by exposing a group of users to some emotion elicitation material, e.g. images and videos with known emotional labels such as stress and boredom. Signals from those users, e.g. HR and facial expressions, are measured during the interaction and used to train the machine learning model according to the labeled elicitation material. Ideally, the trained model can be generalized and used to detect the emotional state of different users, based on the analysis of their signals. This approach, however, fails to learn individual nuances since it assumes all users behave similarly. In practice, this approach is limited to detecting the average behavior of the training group. The great variability between individuals regarding physiological signals and emotional states does influence the process. Studies have shown that the correlation between facial analysis and the emotional states of the training population significantly differs from the expected correlation described in the literature for other populations (Grafsgaard et al., 2013). Additionally there are indications that a machine learning model presents higher prediction rates for users with the highest self-reported emotional levels during the training phase, as well as the lowest prediction rates for participants with the lowest self-reported emotional levels during the training phase (McDuff, Hernandez, et al., 2016). It emphasizes the individualities of each user and the need of a user-tailored approach that preserves such characteristics.

Investigations regarding a user-tailored approach can be found in the literature. It has been proven that a model created from a group of users is less effective at detecting emotions than a model created from a single person which is used to analyze that same person in the future (Bailenson et al., 2008). This user-tailored approach is more likely to learn individual characteristics, not the average features of the training population. Additionally, some works show a migration from physical, obtrusive approaches for signal acquisition from users in favor of remote-based approaches. Advances in areas such as computer vision allow the remote acquisition of input signals, including HR infor-
mation, based on the analysis of videos of users. The remote detection of HR, for instance, proved a promising approach applied to infer boredom/stress levels (Kukolja et al., 2014) or cognitive stress (McDuff, Gontarek, and Picard, 2014b) of a person. Such a remote and non-obtrusive approach, combined with a user-tailored machine learning model, allows the development of new tools for emotion detection.

This thesis presents an approach built on the previously mentioned studies applied to the context of games. The main contribution is the detection of emotional states of users during gaming sessions using remote acquisition of signals via computer vision, a user-tailored model and emotion elicitation based on a novel game-based calibration phase. The approach is automated and implemented as a software without the need of specialized equipment, e.g. sensors, only a regular video camera.

The following sections describe how the proposed approach can be achieved, showing the problem specification, the research aim and its contributions.

1.1 PROBLEM SPECIFICATION

As previously described, questionnaires and physiological measurements using intrusive sensors are the most common approaches used to obtain data for emotion estimation. Both approaches interfere with the natural behavior of users, which affects any research procedure. Improvements to such approaches have been proposed in the literature, including the use of computer vision for remote extraction of user signals and a user-tailored machine learning model to map those signals into emotional states. The material used for emotion elicitation is also an important component of the process to accurately capture the singularities of each user.

One of the problems with previous work is directly connected to the emotion elicitation material used in the process. In the majority of the existing studies, subjects had limited interaction with the content being presented: they performed tasks mentally (e.g. counting), watched videos/images or performed gamified cognitive tests for a short period of time. These are artificial situations that are unlikely to happen in a context involving games. The models trained from such emotion elicitation sources are less likely to cover the range of emotional activity featured by users during gaming sessions, especially those with a challenging game lasting several minutes. There is a lack of investigations regarding the use of games as emotion elicitation sources, which is of interest to the games research community. The process of detecting the emotions of users while they play games is more likely to succeed with a model trained from game-based emotion elicitation materials instead of images and videos. With game-based emotion stimuli, users take an active role in the process, making decisions and directly interacting with the content. It results in more genuine emotional manifestations. When images, videos or gamified tests are employed, users take a passive role with limited possibilities for interaction or emotional involvement, resulting in less significant emotional manifestations.

Regarding the initiatives based on computer vision and emotion estimation, the remote detection of HR, for instance, proved a promising approach used to infer boredom/stress levels (Kukolja et al., 2014) or the cognitive stress (McDuff, Gontarek, and Picard, 2014b) of a person. The application of such techniques, however, has not been proposed in a context involving games and the natural behavior of users. Experiments regarding the use of computer vision and signal extraction were performed under extremely controlled situations with few game-related stimuli. A significant limitation with such approaches was that subjects were asked to remain still during the experiment. This is uncommon user behavior during the interaction with emotional stimulation which hinders the real
efficiency of such remote detection techniques. In particular, when game-based emotion elicitation is employed, users are likely to behave in a more natural way, e.g. featuring facial expressions and moving the body (Bevilacqua, Backlund, and Engström, 2016), which directly affects the remote measurements of physiological signals. The use of such computer vision techniques within the context of games and natural behavior must have its reliability confirmed. Additionally, the techniques must be adapted to overcome the challenges associated with its usage in the context where users behave naturally while playing games instead of being oriented to remain still.

Finally, previous works focus on predictive models based on a collective perspective. As a consequence, a model is usually trained from the data of several users, which in practice describes the average behavior of the group and excludes the key individualities of each user. Such individualities are the main characteristics that define a person, since people are different in many aspects, including expectations regarding culture and personal belief (Goldberg, 1993). It has been proven that a user-tailored approach is more likely to produce better emotional estimations (Bailenson et al., 2008), however no previous work has focused on game-based emotion elicitation combined with a user-tailored model. It is reasonable to believe that those individual characteristics might be better observed with more personalized and complex emotion elicitation materials such as games. Furthermore, a user-tailored approach is likely to preserve and better account for individual characteristics in a method for emotion detection, as opposed to a group model to detect emotions. Additionally, models created from a group are highly affected by ethnical and gender bias, since it is significantly difficult to obtain data from a group that accurately represents the world population. Such limitation is non-existent in a user-tailored model, since the approach is, by design, based on the data of a single person who is already a perfect representation of him/herself.

In summary, previous works focus on models trained from the data of a population instead of a user-tailored approach. As such they dilute the peculiarities of each user and tend to predict the average behavior of a group. When a model is used, it is trained with emotion elicitation materials based on images (Giannakakis et al., 2017; Anttonen and Surakka, 2005), videos (Bailenson et al., 2008; Grundlehner et al., 2009) and gamified cognitive tests (McDuff, Gontarek, and Picard, 2014b; McDuff, Hernandez, et al., 2016). The use of games as emotion elicitation sources is not fully explored. The acquisition of user signals, e.g. HR and facial expression, is performed remotely via computer vision, however, its applicability in a context involving games and natural behavior lacks further investigation. In that light, there is a lack of initiatives focusing on non-obtrusive, user-tailored emotion detection models, in particular, regarding stress and boredom, within the context of games research that is based on emotion data generated from game stimuli. This thesis presents research that aims to fill that gap, providing the games research community with a process to remotely detect the emotional state of users in a non-obtrusive way, based on a model trained from a novel game-based calibration phase, which directly relates to the context of games research.

1.2 RESEARCH AIM

The aim of this research is to produce an emotion detection process that relies on computer vision to remotely acquire psychophysiological signals from a person, in order to detect his/her emotional state regarding stress and boredom. The emotion detection is based on data obtained in a game-based calibration phase.

The overall research question is the following:
How can the emotional state of players during the interaction with games be remotely detected on a user-tailored basis with the utilization of an ordinary camera and games as emotion elicitation sources for calibration?

The following research objectives (O) to support the overall aim of this project have been identified:

**O1**: identification of the main concepts, theories and signals associated with the psychophysiological profile of users and their emotions within the field of HCI, particularly regarding games research. The outcome of this objective is a definition of stress and boredom within the context of this research, as well as the identification of the psychophysiological signals that are commonly applied to emotion detection.

**O2**: identification of existing computer vision techniques that can be employed to remotely extract the identified psychophysiological signals of users via the analysis of videos. The investigation includes the analysis of how existing techniques are being applied to emotion detection. The set of signals to be remotely extracted is based on the results of objective O1.

**O3**: investigation of the feasibility, accuracy and challenges of applying the identified computer vision techniques, regarding the extraction of the signals, within the context of computer games. This objective also encompasses the analysis of the behavior of players during gaming sessions and how it affects the technique.

**O4**: investigation and validation of the concept of a game-based calibration phase as an emotion elicitation source able to provide data to fit a user-tailored predictive model. The result of this objective is to design and validate a set of calibration games that can trigger the emotional responses required for the analysis of the remotely obtained signals and detection of boredom and stress levels by the model.

**O5**: proposal of a user-tailored, multifactorial model that uses the identified physiological and non-physiological signals, the computer vision technique and the calibration data to detect the current stress/boredom levels of a person while he/she plays any video game.

**O6**: experimental validation of the proposed emotion detection process through an experiment involving a commercial off-the-shelf game.

1.3 KNOWLEDGE CONTRIBUTIONS

The result of this research adds to the body of knowledge of HCI and games research. Information regarding concepts, models and theories involving games, emotions and computer vision has been identified, evaluated and orchestrated to work in combination. The main knowledge contribution of this research is a novel process for emotion detection that is remote, non-intrusive and constructed from a game-based, multifactorial, user-tailored calibration phase. The process is able to detect the stress/boredom levels of users during their interaction with games.

The proposed emotion detection process per se is the main contribution; however, it is composed of different parts that possess individual contributions on their own. A list of all those individual contributions and their relation to the research objectives mentioned in Section 1.2 follows:
The concept of calibration games, which are games with specific characteristics used as emotional elicitation sources to identify an emotional profile of users (O1 and O4).

Game-based calibration process that uses calibration games to train a machine learning model for emotion detection regarding stress and boredom (O4 and O5).

Identification and adaptation of computer vision techniques suitable for the remote extraction of user signals in a context involving games and natural behavior (O2 and O3).

A multifactorial, user-tailored machine learning model that maps a set of remotely acquired user signals, e.g. HR and facial actions, onto emotional states related to stress and boredom (O1, O4 and O5).

Validation of the proposed emotion detection process in an experimental setup (O6).

The purely remote-based approach proposed by this research enhances the available methods for the investigation of the emotional states of stress and boredom. The approach, which is based on a novel user-tailored, game-based calibration phase, maps a set of variations of signals onto two specific emotional states, i.e. stress and boredom. This information can be used by other researchers to identify important moments during the interaction of players with games, such as when the recognized pattern is closer to stress. In game design research, for instance, that instrumentation can be used as another way of obtaining information from a user during a game session. The use of questionnaires, which shift the player’s focus away from the game, can be enhanced and/or replaced by the use of the proposed method, making the process less obtrusive. By remotely reading information regarding stress and boredom, a researcher can use such information to better understand concepts such as engagement, frustration, immersion and flow in games, for instance. Additionally it can be used in any activity that relies on stress/boredom as an important measurement, for example, usability tests in software and games. Another contribution is a better understanding of how the selected signals are related to stress/boredom. Other researchers might use that information in contexts outside the area of games research, such as the measurement of customer satisfaction or interest in stores.

The general structure of the proposed process is illustrated in Figure 1.1. The process contains two main phases: a calibration and an emotion estimation phase. In the calibration phase, the user plays a set of carefully designed games, named calibration games, that act as emotion elicitation sources. During this phase, the user signals elicited from the interaction with the games, e.g. HR and facial expressions, are remotely acquired and used to train a user-tailored model. This model is the foundation to the detection of the stress and boredom levels of that particular user in any other game. In the emotion estimation phase, the user interacts with any ordinary game, e.g. a serious game, while his/her signals are remotely acquired and fed into the previously trained user-tailored model. The model then outputs the current levels of stress and boredom for that user in that game.

The process is based on a non-contact, multifactorial analysis of user signals obtained from a video stream via computer vision. The principal of the emotion detection phase is based on a user-tailored machine learning model which is trained with information obtained from the user while he/she played a set of games in the calibration phase. The
A user-tailored machine learning model is trained according to the process presented in Figure 1.2. Each user plays a set of calibration games while being recorded by a camera. Computer vision is used to process the video feed and remotely extract signals from the user, such as HR and facial actions. Those signals are used as input to train the machine learning model for that particular user (user-tailored model).

The games used in the calibration phase act as emotion elicitation sources. Each of those games is casual-themed and carefully designed to trigger two distinct emotions, i.e., boredom and stress, featuring a progressive transition between them, as illustrated by Figure 1.3. At the beginning of the game, the difficulty level (green line) is low and the user is required to perform few or no actions. The games are designed in a way that does not enable the user to increase the pace of the gameplay or make it faster based on personal skills. As a consequence, the user is forced to play a low-paced gameplay, which leads to an emotional state of boredom (blue curve). As time progresses, the pace of the gameplay and its difficulty level increase linearly. The increase happens at fixed time intervals, e.g., every 60 seconds. At some point in time, which is different for each user depending on gaming skills and personal preferences, the pace of the gameplay and the difficulty level will be overwhelming, leading the user to an emotional state of stress (red curve). As the difficulty level continues to increase, the stress level of the user will also increase. Finally, the difficulty level will increase to the point at which the user is unable to cope with the game. This will lead to consecutive mistakes in the game and eventually terminate it, e.g., health bar of the main character reaches zero.

The above mentioned calibration games are designed to trigger specific emotions and vary them over time. Consequently the remotely collected information from the user during the calibration phase contains a detailed variation profile of the person being analyzed, including changes of each signal and the theoretically known emotional state.
of the user at that moment. If a person has a better response to a certain physiological signal instead of another, e.g. HR over facial expressions, then the variation of that signal accounts for more weight in the training of the model. Since the training process is completely based on the signals of a single user, nuances and individual behavior are likely to be registered and learned. The calibration phase needs to be performed once per person.

After the calibration phase, the person can play any other ordinary game and be monitored in an emotion estimation phase, as illustrated by Figure 1.4. As the user plays the game, signals are remotely acquired via computer vision. These signals are then used as input to the trained user-tailored model of that particular person, which, as a result, produces an estimation of the emotional state regarding stress and boredom for that person in that game. The process relies on the same remotely acquired signals with the addition of the predictions of the model according to the training performed during the calibration phase.

Given that the user has a trained user-tailored model, the emotion estimation phase can be performed for any game as many times as desired. The model uses the remotely obtained signals from the user in conjunction with the calibration data to detect the player’s changes regarding stress and boredom levels in any other game.
1.4 THESIS OVERVIEW AND STRUCTURE

This thesis is divided in parts whose chapters present and explain in detail the scientific methodology, the theoretical background and the studies conducted to achieve the research aim. Figure 1.5 illustrates how each chapter covers the work towards achieving the research objectives described in Section 1.2. Overall, the solution to remotely detecting user emotions proposed by this thesis relies on three main elements: emotion elicitation, i.e. calibration games, acquisition of users' signals, i.e. computer vision techniques, and emotion estimation, i.e. user-tailored machine learning model. These three elements are significantly intertwined and directly affect each other. In summary, when users interact with game-based emotion elicitation material, they tend to move, laugh and occlude the face, which add noise to (or impede) the estimations performed by the computer vision techniques. This then affects or even prevents the creation of an emotion detection model, requiring the games to be re-worked or adapted to ensure they continue to serve as emotion elicitation material. Consequentially, this can lead to a chain reaction that has cyclical effect over the previously mentioned elements.

Figure 1.5: Overview showing how each chapter covers the work towards the achievement of the research objectives, i.e. O1 to O6.

According to the literature review conducted for this research, the use of those three main elements combined has never been tried before. As a result, it was not possible to determine beforehand whether they could actually work in combination, in order to detect user emotions remotely. A set of iterations would be required to test, evaluate and learn about those elements working together. Therefore, Design Science, an iterative problem solving research method, was deemed the best approach to conduct the investigation, as explained in Chapter 2 (page 17). Using such research method, firstly the literature is consulted and a tentative solution is proposed and evaluated. The results of such an evaluation provide new insights about the problem, which are then interpreted against the literature, leading to a new tentative solution. The cycle is repeated until a solid contribution is formed. For the research conducted in this thesis, the literature review made it clear that a more fruitful way of detecting emotions would be to interpret them as a manifestation of psychophysiological signals, e.g. HR and facial activity, instead of interpreting them according to the premises of psychology. It steers the research towards human physiology instead of psychology, whose interpretation is likely less sub-
jective and more quantitative oriented. The literature review also focused on identifying which psychophysiological signals have been used in previous work that concentrated on emotion detection, along with identification of computer vision techniques that can remotely extract those signals. After the set of signals and computer vision techniques to be used in this research were selected, two experiments were conducted, in which six iterations of the design science cycle were performed: five iterations in Experiment 1 (studies 1 to 5), and one iteration in Experiment 2 (final evaluation). Experiment 1, detailed in Chapter 8 (page 59), aimed to evaluate the feasibility of using the previously mentioned three main elements of this research in conjunction. It was mainly designed to test whether the concept of calibration games, a novel aspect of this thesis, would cause an emotional reaction in the subjects that could be remotely detected from psychophysiological signals. Data gathered in experiment 1 was analyzed in five studies, i.e. studies 1 to 5, which can be regarded as five iterations in the design science cycle.

**Study 1:** detailed in Section 8.5 (page 63) was an exploratory evaluation of facial activity as well as perceived boredom and stress levels. The evaluation is connected to previous works and emotion/game theories presented in Chapters 3 and 4 (pages 25 and 31, respectively). The analysis of the subjects’ self-reported emotional states statistically confirmed that they perceived the games as being boring at the beginning and stressful at the end. It supported the idea that calibration games, a corner stone of this thesis, could be used as emotion elicitation material. Additionally, the manual and empirical analysis of the video recordings indicated that the subjects presented more facial actions during stressful periods of the games compared to boring periods. Finally, there were indications that subjects featured a neutral face most of the time, which implies that it is not trivial to estimate emotions based purely on facial analysis without a context.

**Study 2:** detailed in Section 8.6 (page 67) evaluated whether calibration games could produce variations in physiological signals, namely HR, as described by the theories and previous work presented in Chapter 5 (on page 41). The analysis of the HR collected with a physical sensor, i.e. watch, during the experiment statistically confirmed that the subject’s HR was different during stressful and boring periods of the game. This information confirmed that calibration games could be used as emotion elicitation material, effectively inducing variations in physiological signals in subjects exposed to them, which could be used to detect emotions.

**Study 3:** detailed in Section 8.7 (page 73) evaluated the feasibility of remotely detecting the variations of HR that were confirmed in Study 2. Remote photoplethysmography (rPPG), as described by the theories and works presented in Chapter 6 (on page 45), was the technique used to remotely estimate HR information from videos of subjects interacting with games. Estimations of HR obtained with rPPG were compared to HR measurements collected with the physical sensor. This highlighted how the technique is affected by the natural behavior of subjects, e.g. movement and laughter. The study also provided insights about the mean estimation error of the technique when it is affected by the introduced noise of the subjects’ natural behavior.

**Study 4:** detailed in Section 8.8 (page 82) evaluated the facial activity of subjects similarly to Study 1, however, it used a completely automated process relying on computer vision. A method to automatically track facial muscles connected to emotional reactions, which were reported by previous works detailed in Chapter 4 (page 31), was developed and evaluated in the study. The results of the automated facial analysis conducted on all the subjects statistically confirmed the findings of the manual analysis performed in Study 1, suggesting that the subjects presented more facial activity during the stressful rather than boring periods of the games.
Study 5: detailed in Section 8.9 (page 94) was the first iteration in the design science cycle where all the three previously mentioned main elements of this research were in place and working together to detect the emotional state of subjects. Based on previous works presented in Chapter 7 (page 53), a neural network was trained and used to remotely detect emotions. For each subject, two calibration games were used to train the model, i.e. user-tailored neural network, while one calibration game was left out for use in evaluating the accuracy of remotely estimating the emotional state of the subject. Permutations were employed to ensure all calibration games were used in the 2-training-and-1-testing configuration. Additionally, different sets of signals, i.e. facial activity and HR, only facial activity, only HR, an so on, were evaluated in relation to the accuracy of emotion detection. The test was motivated by findings described in Chapter 7 (page 53), which suggest that a multifactorial analysis, when more than one signal is used in the emotion detection process, yields better results than using a single signal in isolation. The results regarding the accuracy of the emotion estimation indicated that a multifactorial approach is indeed better suited for the process. Additionally, the results indicated that the proposed method is able to perform emotion estimations better than chance-level classification, which confirmed the feasibility of the method proposed in this thesis. Finally, the achieved results indicated that the joint use of calibration games and the remote acquisition of signals to train a user-tailored machine learning model can detect the emotional states of users.

Following the confirmation of feasibility provided by Study 5, a new experiment was designed and conducted to validate the proposed method. At this point in the research project, all the main elements of the proposed method were in place and working together, which allowed an evaluation on a larger scale, compared to the first experiment. Experiment 2, detailed in Chapter 9 (page 103), aimed to evaluate the accuracy of the final method, proposed in this thesis, to remotely detect the emotions of users interacting with a game. In experiment 2, the subjects played the same calibration games of experiment 1, however they also played 7 levels of Infinite Mario, detailed in Section 9.2.3 (page 106), which is a clone of the commercial off-the-shelf (COTS) game, Super Mario. Data from the calibration games were used to train a user-tailored neural network, which was employed to estimate the emotional state of each subject during the interaction with the levels of Infinite Mario. The results confirmed on a larger scale the findings of the previously conducted Study 5, supporting the idea that the method proposed in this thesis for the remote estimation of emotions is feasible.

1.4.1 TEXT ORGANIZATION

The thesis structure first introduces the methodology used to conduct the research, as presented in Chapter 2 (page 17). Subsequently, a literature review presents the theoretical background that supports this research. Concepts, theories and techniques are presented and contextualized within the aims of this thesis, including emotions and games (Chapter 3, page 25), emotions and facial analysis (Chapter 4, page 31), emotions and physiological signals (Chapter 5, page 41), remote extraction of physiological signals (Chapter 6, page 45), and multifactorial emotion estimation (Chapter 7, page 53). The following includes a part that contains chapters related to the two experiments conducted to investigate, evaluate and validate the elements proposed by this research. Experiment 1 (Chapter 8, page 59) and its five studies show how the method proposed in this thesis was constructed. Experiment 2 (Chapter 9, page 103) shows how the proposed method was evaluated in a larger scale in a scenario that is more similar to a real-use case. Thereafter, the results achieved in this research project, along with a discussion
of its implications in the field of games research and other areas are presented (Chapter 10, page 121). This part also includes chapters that discuss the ethical and privacy considerations of this research (Chapter 12, page 133), as well as limitations and critiques related to the proposed method (Chapter 13, page 139). Finally, closing remarks, a conclusion (Chapter 14, page 145) and suggestions of future work (Chapter 15, page 149) are presented.

1.4.2 DEFINITIONS AND SCOPE

The research presented in this thesis is a multidisciplinary work that involves theories and concepts from different fields. Some of those concepts, particularly regarding emotions, are shared among the fields, however they have different definitions and interpretations. As a consequence, it is important to establish what the fields involved in this research are, as well as the understanding of concepts in the context of this work, particularly regarding emotions. Figure 1.6 illustrates the fields related to this research together with the community that the work contributes to.

![Figure 1.6: Different fields involved in this research. Main contribution is in the area of Games Research.](image)

The present research mainly involves and contributes to the field of games research, particularly the branch related to variations of emotions during the interaction between users and games. The process of monitoring user emotions in human-computer interaction, known as affective computing (Picard, 2000), is a challenging endeavor and a recurrent research topic. This thesis brings concepts of computer vision into the field of games research, to enhance the process of monitoring user emotions, making it non-obtrusive by proposing a method to remotely acquire and analyze player’s signals in order to detect stress and boredom levels. As detailed in Chapter 3 (page 25), many different theories can be used to model emotions and consequentially define stress and boredom in different contexts, including games, for instance. Despite being connected to the topic of human emotions, this research does not focus on Psychology or Cognitive Science, whose definition of stress and boredom might carry a different meaning and correlation with games. In the context of this research, the definition of emotions is less related to Psychology or Cognitive Science; instead, it is based on Biology with focus on a quantitative analysis of the human physiology. The research presented in this thesis relies heavily on the fact that physiological arousal is connected to emotion regulation (Appelhans and Luecken, 2006; Schubert et al., 2009), in particular the interactions of...
the Autonomic Nervous System (ANS). The systems that compose the ANS interact to produce variations in the physiological signals to maintain a lower or higher degree of physiological arousal, e.g. reacting to a threatening situation triggers the body to alertness, which is commonly referred to as the “fight or flight” response. As a result, both stress and boredom are defined in the scope of this thesis as emotional states that are manifestations of psychophysiological signals in the body caused by such interactions of the ANS. It is out of the scope of this thesis to create a (new) definition of stress and boredom based on the analysis of psychophysiological signals. On the contrary, the aim is to identify those emotional states on the basis of the analysis of the previously mentioned signals and their variations. Those signals are the source of information used to classify the emotional state of subjects.
CHAPTER 2
RESEARCH METHODOLOGY

The aim of this research was to produce a technology-based solution to the problem of non-contact emotion detection within the context of games research. The solution is a method comprising a user-tailored model trained from a game-based calibration phase and able to infer the emotional state of a player regarding stress and boredom via the analysis of remotely acquired user signals. While the constructs, models and methods involved in such an aim have been individually studied in previous research, the combination of all these elements in a single solution within the context of games research is novel. Since the utilization of these elements in combination had not yet been demonstrated to work, it was necessary to conduct an iterative and incremental process to identify challenges, problems and solutions to achieve the desired goal.

A research methodology that fits such an iterative process is Design Science Research (DSR). Typically, DSR is a problem solving process focused on developing new artifacts. Hevner et al. (2004) define design science in the context of Information Systems as a process that explores a relevant problem within an environment, iteratively measuring and refining the proposed solution according to the existing body of knowledge. The progress is made iteratively as the scope of the design of the artifact is expanded based on the discovery of available means, ends and constraints. Similarly, Johannesson and Perjons (2014) define design science as the scientific study and creation of artifacts as
they are developed and used by people, with the goal of solving practical problems of general interest. The outcome of DSR includes the contextual knowledge about the artifacts.

Vaishnavi and Kuechler (2015) structure the mentioned iterative process in five steps, illustrated in Figure 2.1: awareness, suggestion, development, evaluation and conclusion. These steps constitute the DSR process model, or DSR cycle. The awareness step is the recognition and articulation of a problem from an environment, which originate from studying, for example, the existing literature. The suggestion step presents a tentative design of how the problem might be addressed, which envisions a new functional artifact with a novel configuration of existing and/or new elements. The development and the evaluation steps comprehend the implementation of the tentative design and its analysis with well defined metrics and measurements. The evaluation either confirms or contradicts the hypothesis about the behavior of the object, leading to new awareness (and other iterations of the process) or to a conclusion. Finally, the conclusion step determines why and how the artifact worked or did not work within its environment. Vaishnavi and Kuechler (2015) categorize the knowledge that is gained during the research process and presented in the conclusion step as firm and loose ends. In the former, the conclusion shows facts that can be repeatably applied or behavior that can be repeatably invoked, while in the later, the conclusion shows anomalous behavior that defies explanation and may serve as further research topics.

The types of artifacts resulting from DSR are constructs, models, methods and/or instantiations (Oates, 2005; Johannesson and Perjons, 2014). Constructs are the terms, concepts, definitions and notations required to formulate and represent the problem. Models are a combination of constructs related to each other to represent possible solutions to a problem. Methods provide guidance for the models to be produced and the process to solve the problem. Finally, instantiation is a working system which demonstrates that theories and artifacts, i.e constructs, models and methods, can be implemented in a computer-based system.

The present research involved the use and orchestration of three main components, illustrated in Figure 2.2. These components are a game-based emotion elicitation part, composed of calibration games, the part involving the remote acquisition of user signals via computer vision, and an emotion estimation part, composed of a machine learning model. The use of game-based material in a calibration phase in this research influences how users behave during the emotion elicitation process, e.g. body movement and facial expression. The movement of users directly impacts the techniques for the remote extraction of user signals during both the calibration and the emotion estimation phases,
since these techniques are highly influenced by motion. The accuracy of the techniques regarding the remotely acquired signals is affected as well, which might invalidate the feasibility of remotely reading determined physiological and non-physiological signals required by the emotion estimation model.

The interaction among the mentioned components, i.e. emotion elicitation, acquisition of user signals and emotion estimation, must be continuously investigated and adapted to overcome the previously described challenges. As a consequence, an iterative cycle of development and research is required, as illustrated by Figure 2.3. In each iteration, a possible solution for the current set of problems is generated, rigorously tested and evaluated, producing information to guide the next iterations in the cycle. The set of design alternatives in this research was related to the following: the identification of physiological and non-physiological signals to be used in the emotion estimation process, how they can be elicited with games in a calibration phase, which computer vision techniques can be employed to remotely acquire the signals, and which machine learning model is able to map the information into emotional states. The set of constraints involves problems associated with users behaving naturally, e.g. laughing and moving the body during the procedure, the use of non-specialized hardware, e.g. ordinary camera, accuracy and efficiency of the solution, among others.

Figure 2.3: The Generate/Test cycle. Adapted from Hevner et al. (2004)

Design-science research requires the application of rigorous methods in both the construction and evaluation of the designed artifacts (Hevner et al., 2004; Johannesson and Perjons, 2014; Oates, 2005). One of such evaluation methods is experimental research, which is the strategy used to build and validate the knowledge in this thesis. Such an approach is composed of a set of research designs that use controlled testing and manipulation of variables in order to understand causal processes (Robson and McCartan, 2016). The foundation of an experiment is to manipulate a variable (or a set of them) and measure any changes in other variables. It establishes the effect on a dependent variable, which is the focus of the research. The method constructed in this research links the variations of user signals to emotional states of stress and boredom in the context of games. Consequently, there is a causal effect in the process, since identified variations (cause) will precede changes in stress/boredom levels (effect). It progresses to the construction of a hypothesis where the cause will consistently lead to the same effect, so that the link between the variations of signals and emotional levels can be inferred or predicted.

The preferred experimental design for the present research was based on a within-subject approach (Lane, 2015). In such an approach, all participants perform at all levels of the treatment and there are no control groups. Such a design could be criticized for having low internal validity, since it is not possible to unambiguously attribute cause and effect (Kirk, 1982). A two-group approach could be suggested as having stronger internal validity, since it contains a control group and allows a less ambiguous conclusion. In the context of the present research, however, any multiple group design implies the comparison of physiological signals and emotional perceptions among different people.
Given the social and cultural background of the participants, it is virtually impossible to compare two groups of people regarding stress/boredom. People have different preferences, culture and expectations, which cause maturation and history threats to internal validity (Trochim and Donnelly, 2001). In that light, the within-subject approach relies on a one-group experimental design to increase internal validity, since subjects are compared with themselves, which removes inter-subject differences.

Design science research was the approach deemed adequate for the research presented in this thesis. The iterative nature of the methodology allowed the investigation, validation and better understanding of the interactions among the components involved in the proposed research aim. The research presented in this thesis produced one main artifact, i.e. a game-based, user-tailored method for the estimation of emotions remotely. The artifact was produced from six iterations performed following the DSR cycle illustrated in Figure 2.1. Chapter 8 documents five of those iterations, which were focused on proposing a solution for the dependency among the main parts of this research: emotion elicitation, acquisition of user signals and emotion estimation, as previously mentioned and illustrated in Figure 2.2. Section 8.9, in particular, details the iteration that produced the final tentative design that is the result of this thesis. The main artifact was composed of several different components, e.g. facial analysis and rPPG estimation of HR, that were orchestrated together to produce a game-based method for emotion estimation. Finally, Chapter 9 presents the final iteration in the DSR cycle. In that iteration a complete evaluation of the tentative design was conducted. It produced several performance measurements that are presented and discussed in Section 9.3 and Chapter 10.

2.1 ETHICS AND PRIVACY IN THIS RESEARCH

The research presented in this thesis involved several elements that are sensitive to ethics and privacy, including how the research was conducted. This research involved two experiments and several studies conducted on the data gathered from these experiments. All the experiments, as well the aforementioned analysis and data collection, were ethically conducted and handled according to the Principles for Research Ethics in the Humanities and Social Science. Particularly all materials and procedures were designed according to recommendations of the CODEX, the Swedish Research Council, which provides guidelines, ethics codes and laws that regulate and place ethical demands on the research process.

Following such guidelines, participants of the aforementioned experiments had given a written consent stating that they understood what was happening and that they voluntarily wanted to participate in the study being conducted. Participants were informed that they may choose not to participate and they may withdraw their consent to participate at any time. In such a case, they would not be asked for any explanations regarding their decision not to participate or to withdraw from any activity. Additionally, subjects were informed regarding the description of the research, including the overall research plan and the aim of the research being conducted. A main component of this research is the idea of calibration games, i.e. games whose difficulty level continually progresses over time to induce stress. It is important to state and clarify that such an induced emotional state of stress comes from the addition or speed up of game elements that the subject needs to handle on screen, e.g. more cards to sort or faster falling pieces. Both the

1 Forskningsetiska Principer inom Humanistisk-Samhällsvetenskaplig Forskning
2 http://www.codex.vr.se
content and theme of all the games used in this research are casual, family friendly and completely based on cartoon graphics. No severe emotional or traumatizing stress was induced, only the expected level of stress resulted from an ordinary interaction with a game whose elements change, i.e. intensity in the activity on screen. In no circumstance was offensive, violent, culturally or morally disturbing content used in the games, e.g. guns and gore, to induce any emotional state. The games used in this research did not differ from any existing commercial game whose difficulty level also increases to induce stress, in order to produce a feeling of reward when the player completes a challenge. As described in Chapters 8 and 9 (pages 59 and 103, respectively), before starting any experiment, the subjects were instructed to not give up in the middle of the games. Such instruction was contextualized with the games to be played and never superseded or prevented the subjects from interrupting the activity at hand and leaving the study, as they were informed they could do, even during the time they were playing the games. As previously mentioned, subjects were informed on more than one occasion that they could stop any activity and leave the study at any time. Therefore, it was deemed that no further actions regarding ethical considerations were needed, since the experiments only included normal interaction with standard technology.

Data privacy issues were also made clear to all participants, e.g. that sessions would be video-recorded, and all efforts were put in place to ensure the participants’ privacy. For example, the participants were informed beforehand about what kind of information would be collected from them, e.g. HR, video, in-game actions and answers to questionnaires. They were informed that the collected data would only be used for non-commercial research purposes, that all collected data would be stored off-line on a hard-drive that only the principal investigator could access. Data protection included restricted access to such a hard-drive, which was handled and physically kept in the principal investigator’s possession at the facilities of the University of Skövde. This helped to ensure that unauthorized persons would not have access to the data. All the participants were also informed beforehand that all the data would be anonymously collected and their identity would not be revealed, including in any publication resulting from this work. It is not possible to trace any published results or the information in the present thesis back to the subjects. The participants were assured that their data would never be publicly disclosed, i.e. videos, HR information, pictures, and answers to questionnaires, without their written consent. Some of the publications of this research include pictures of participants, whose consent had been given upon prior request regarding the use of such pictures.

Finally, in order to ensure subjects fully understood the research being conducted and how their data would be used, in both experiments each subject attended a debriefing session. In this session, the researcher explained to the subjects how the games they had just played were designed, including the linear progression of boredom and stress. Additionally, the subjects were informed about how the data would be analyzed, e.g. to find a relation between HR and stressful moments, or that remote estimations of their HR would be performed on the video recordings and tested against the data collected by the watch the participants had used. Participants had the opportunity to ask questions about the experiment, the study, the data collection or any matter related to the research they deemed important. It is also important to highlight that a wearable sport device, i.e. watch (TomTom Cardio Runner), has been used in this research. This device was selected instead of a medical grade equipment to reduce privacy and ethical concerns regarding the data gathering of personal, biological information.
PART II

LITERATURE REVIEW
CHAPTER 3
GAMES AND EMOTIONS

Research of games is a broad topic that involves different disciplines and definitions. In the context of this thesis, a game is defined as a system in which players engage in an artificial conflict, defined by rules, that results in a goal (Salen and Zimmerman, 2004). An artificial conflict is a set of challenges, e.g. sort elements within a time constraint, the player must overcome in order to achieve the goal of the game.

![Figure 3.1: Repeating cycle of increasing challenge followed by a reward, keeping the player in the flow zone. Reproduced from Schell (2014).](image)

Schell (2014) mentions that, from a game design perspective, the difficulty level of such challenges affects the emotional state of players, e.g. moments of boredom or anxiety/stress. Every time a reward is given to the player, which is usually a tool to increase the player’s power, the game’s challenge level is lowered because the player becomes more skilled. After a period, this increase in the player’s skill level causes the game to become boring because the challenges become easier to overcome. At that point, the game increases the difficulty level again, raising the challenge levels for the player once more, causing anxiety. The anxiety and stress period lasts until the player is rewarded again, whereupon the newly obtained power eventually lowers the challenge levels again (resulting in boredom), causing the cycle to repeat itself.

Chen (2007) used the theory of flow or the “theory of optimal experience”, originally established by Mihaly Csikszentmihalyi, to describe how player experience, in large part, depends on the relationship between a game’s challenge and its player’s level of skill. A challenge beyond the player’s skill to address and overcome causes anxiety, while the opposite results in disinterest, leading to boredom (Chen, 2007). An ideal challenge/skill balance in a repeating cycle of increasing challenge followed by a reward keeps the player in a state of optimal experience and concentration, i.e. flow zone, as illustrated by Figure 3.1.
The following sections describe in more detail the theoretical foundation of emotions, connecting them to the context of games.

3.1 EMOTIONS THEORY

In the field of games research, one of the most mentioned theories regarding emotions is the theory of flow. It has been used as the foundation for several concepts, including engagement and immersion (Brown and Cairns, 2004), sense of presence (Weibel and Wissmath, 2011) and applicability in game design (Sweetser and Wyeth, 2005; Chen, 2007; Cruz and Uresti, 2017). Flow was originally defined as a phenomenon in which a person experiences a subjective state characterized by an intense level of attention during the execution of an intrinsically motivated activity (Nakamura and Csikszentmihalyi, 2014). As previously mentioned, an ideal challenge/skill balance in a game leads players to an optimal state of experience and concentration, e.g. flow, therefore, flow constructs are of interest to the games community. This peculiar state of flow, however, is not limited to activities involving games; it can also be experienced in a variety of other activities, e.g. dancing and climbing. The connection of the theory of flow to contexts other than games is outside the scope of this thesis.

Further research (Nakamura and Csikszentmihalyi, 2014) refined the original definition of the flow state, culminating in the eight channel model of flow, illustrated in Figure 3.2. This model better describes the emotional state of users according to the challenge/skill balance in relation to the subject mean, since it is more detailed and refined than the original flow model. In the eight channel model of flow, for instance, a person performing a low skill and low challenge task experiences apathy, while in the original model the classification would indicate a flow state. Emotional states such as stress and boredom can be described as a function of the current player’s skill level and the level of challenge he/she is facing.

Figure 3.2: Eight channel model of flow. Reproduced from Nakamura and Csikszentmihalyi (2014).

Another model commonly mentioned in the literature is the basic emotions proposed by Ekman and Friesen (1971). Constructed from an experiment involving cultural differences, it suggests that particular facial muscular patterns and discrete emotions are
The six emotions mentioned in the theory are happiness, surprise, sadness, fear, anger and disgust, which are strictly basic emotion models of affective state. A contrary definition is presented by Russell (1978), who defined another model of emotions named Circumplex Model of Affect (CMA). Commonly referred to as Russell’s Arousal-Valence (AV) space, the model is contrary to strictly basic emotion models of affective state, where each emotion emerge from independent neural systems (Posner, Russell, and Peterson, 2005). The model proposes a dimensional approach where all affective states arise from the activation of two fundamental neurophysiological systems: arousal (or alertness) and valence (a pleasure-displeasure continuum).

![Figure 3.3: Representation of the Circumplex Model of Affect. Horizontal axis represents the valence dimension and the vertical axis represents the arousal or activation dimension. Reproduced from Posner, Russell, and Peterson (2005).](image)

Figure 3.3 illustrates the AV space. The horizontal axis represents the valence dimension, which varies from the negative, unpleasant spectrum to the positive, pleasant spectrum. The vertical axis represents the arousal or activation dimension, which varies from low (bottom) to high (top). Each emotion is the result of a linear combination of these two dimensions. An emotional state of excitement, for instance, is conceptualized as the product of a positive activation in the neural system associated with valence along with a high activation in the neural system associated with arousal. The different scales of activation of each of those two dimensions produces different emotional states.

### 3.2 IMMERSION, ENGAGEMENT AND SENSE OF PRESENCE

As previously mentioned, the theory of flow is used considerably to explain emotional states and concepts in the field of games research. The definition of flow, however, requires a more sophisticated interpretation, as investigated by further research. Different elements are also connected to flow, such as engagement, immersion and sense of presence.

Immersion, in the concept defined by Brown and Cairns (2004), refers to the degree of involvement of players with a computer game. In that light, the authors theorize that a
player overcomes barriers that limit his/her degree of involvement in immersion. After each barrier is broken, the sense of immersion deepens. The first barrier, for instance, is named engagement and refers to the player’s willingness to invest attention and energy to learn how to play the game. The concept of flow, as defined by Nakamura and Csikszentmihalyi (2014), is an extreme state, which is only achieved when the player has overcome all previously mentioned barriers and is in a “total immersion” state. This condition rarely happens, since it requires the player’s highest level of attention. As a consequence, engagement is more plausible and common during gaming experiences than flow.

The fact that there are several works (Boyle et al., 2012) related to understanding and defining the concepts of engagement and immersion demonstrates the interest of researchers to broaden the view beyond flow alone. Presence, for instance, which describes the player’s feeling of actually being in the game, is reported as an important aspect of engagement and immersion (Weibel and Wissmath, 2011). Studies connected to training simulation (Engström et al., 2016), for instance, also indicate that contextualization (increasing the sense of presence) might affect immersion positively. As well as the efforts to understand and define engagement and immersion, there are also studies that try to measure these concepts. Most of the approaches used in those studies are based on questionnaires, which by nature players answer subjectively. Additionally, that kind of approach usually interferes with any sense of presence the player has, since it requires a shift in attention away from the game, hence breaking or affecting the level of engagement/immersion as well.

Quantitative approaches to measure engagement have also been investigated, such as the use of physiological signals, for example, heart rate (Ravaja et al., 2006) and eye movement (Jennett et al., 2008). The complexity in defining engagement and immersion is also reflected in the task of measuring them. Ravaja et al. (2006), for instance, claim that the significant variation of physiological signals is an obstacle. Signals increase during emotional arousal, but decrease in response to attention engagement, which makes the measurement of engagement a non-trivial process. It highlights the difficulties in correlating qualitative data to more abstract concepts such as engagement, immersion and flow.

3.3 INSTRUMENTS FOR ASSESSMENT OF EMOTIONS

Questionnaires are a common tool used for the assessment of emotional states of users during experiments involving games and emotions. Different formats of questionnaires are found in the literature. The simplest approach is ad-hoc questionnaires, which contain a set of questions designed by the researcher for a particular experiment. A common type of emotion scale in such an approach is Likert scales related to an emotional state, e.g. stress level.

Another instrument for the assessment of emotions is the Self-Assessment Manikin (SAM) (Morris, 1995), which is an efficient cross-cultural measurement of emotional response regarding pleasure, arousal and dominance. Figure 3.4 illustrates the SAM mechanism. SAM employs a nonverbal, graphic depiction of three affective dimensions. In the top row of the questionnaire the user can express the level of pleasure, ranging from a figure with a smile to a figure with an unhappy face. The middle row presents figures ranging from an agitated (high arousal) state to a relaxed, sleepy state. Finally the bottom row represents the level of dominance (control) of the user in the situation at hand.

A similar approach is the Affective Slider (AS) (Betella and Verschure, 2016), a digital
Figure 3.4: Visual representation of the Self-Assessment Manikin. Reproduced from Morris (1995).

self-reporting tool composed of two slider controls for the assessment of pleasure and arousal. Figure 3.5 illustrates the AS self-assessment mechanisms. AS also employs a nonverbal, graphic depiction of the emotional state of the user, however, the focus is on arousal and valence. A user can adjust the sliders to show his/her level of arousal (top slider) and valence (bottom slider).

Figure 3.5: Visual representation of the Affective Slider. Reproduced from Betella and Verschure (2016).

The selection of an emotion assessment instrument is connected to the research context. Both SAM and AS, for instance, are established and proven emotion measurement instruments, which would strengthen the theoretical foundations of the emotion measurement process. However, a disadvantage is that they require the researcher to instruct users on how to properly answer the questionnaire.
CHAPTER 4

EMOTIONS AND FACIAL ANALYSIS

The human face is a source of information and an important part of communication. Several elements connect this channel of information to emotional states, such as facial expressions and the activity of eyes and head (Akakın and Sankur, 2010). The analysis of such elements can convey information regarding emotional states, e.g. facial expressions are considered one of the most relevant features that can provide an indication about emotional states (Cowie et al., 2001).

Facial analysis is a promising approach to detecting the emotional state of players unobtrusively and without interruptions (Cohn and De la Torre, 2014). The use of computer vision for player experience detection is feasible and visual inspection of gaming sessions has shown that the automated analysis of facial expressions is sufficient to infer the emotional state of players (Tan, Rosser, et al., 2012; Tan, Bakkes, and Pisan, 2014b). Automatically detected facial expressions have been correlated with dimensions of game experience (Tan, Bakkes, and Pisan, 2014a) and used to enhance player’s experience in online games (X. Zhou, X. Huang, and Y. Wang, 2004; Zhan et al., 2008). Although automated facial analysis has become mature enough for affective computing, there are several challenges associated with the process. Facial actions are inherently subtle, making them difficult to model, and individual differences in face shape and appearance undermine generalization across subjects (Cohn and De la Torre, 2014). Schemes such as the Facial Action Coding System (FACS) (Ekman and Friesen, 1977; Cohn, Ambadar, and Ekman, 2007) aim to overcome these challenges through standardizing the measurements of facial expression by defining highly regulated procedural techniques to detect facial Action Units (AU). In the FACS scheme, facial actions/movements are decomposed into 46 different AUs anatomically related to facial muscles. When analyzed in defined contexts, such mapped actions can present a correlation between determined facial features and emotional states, e.g. stress and boredom.

Giannakakis et al. (2017) present a literature review focused on facial elements with value for detection of anxiety and stress, including the involvement of eyes (pupil size variations, gaze distribution, blinking rate), mouth (lips deformation, mouth activity) and head (head movement and velocity). Table 4.1 lists all the identified facial elements. There are indications that blinking increases with emotional arousal, including stress and anxiety levels (Dinges et al., 2005). Gaze direction, gaze congruency (agreement between eye and head orientation) and the size of the gaze-cuing effect (facilitation of reaction time towards visual clues) are also influenced by the level of anxiety or stress (Staab, 2014). Similarly, mouth activity is influenced by conditions of stress, particularly lip movement (Dinges et al., 2005) and asymmetric lip deformation (Metaxas, Venkataraman, and Vogler, 2004). Finally, the frequency of mouth openings has been measured as inversely proportional to the stress level under high cognitive load (Liao et al., 2005).

Different approaches have been used to connect facial analysis to the emotional states of users. Initiatives include manual or automated face detection, use of machine learning models to map facial features to emotions and so on. The following sections present,
Table 4.1: Categorization of facial elements connected with stress and anxiety according to Giannakakis et al. (2017)

<table>
<thead>
<tr>
<th>Head</th>
<th>Eyes</th>
<th>Mouth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head movement</td>
<td>Blink rate</td>
<td>Mouth shape</td>
</tr>
<tr>
<td>Skin color</td>
<td>Eyelid response</td>
<td>Lip deformation</td>
</tr>
<tr>
<td>Heart rate (facial PPG)</td>
<td>Eye aperture</td>
<td>Lip corner puller</td>
</tr>
<tr>
<td></td>
<td>Eyebrow movements</td>
<td>Lip corner depressor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lip pressor</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gaze</th>
<th>Pupil</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saccadic eye movements</td>
<td>Pupil size variation</td>
</tr>
<tr>
<td>Gaze spacial distribution</td>
<td>Pupil ratio variation</td>
</tr>
<tr>
<td>Gaze direction</td>
<td></td>
</tr>
</tbody>
</table>

in more detail, works related to face detection, including the approach used for analysis and connection to emotional states.

4.1 FACIAL ANALYSIS BASED ON SENSORS

The analysis of facial behavior commonly relies on data obtained from physical sensors, e.g. electromyography (EMG), or from the application of visual methods to assess the face, e.g. feature extraction via computer vision (Schrader et al., 2017). The approach based on EMG data uses physical sensors attached to subjects to measure the electrical activity of facial muscles, such as the zygomaticus, the orbicularis oculi and the corrugator supercili muscles (Figure 4.1), associated with smiling, eyelids control and frowning, respectively.

Hazlett (2006) presents evidence of more frequent corrugator activity when positive game events occur. Tijs, Brokken, and IJsselsteijn (2008) show that an increased activity of the zygomatic muscle is associated with self-reported positive emotions. Similarly, Ravaja et al. (2006) show that positive and rewarding game events are connected to an increase in zygomatic and orbicularis oculi EMG activity.

Approaches based on EMG are more resilient to variations of lighting conditions and facial occlusion; however, they are obtrusive, since it is necessary to attach physical sensors to the subject’s face.

4.2 FACIAL ANALYSIS BASED ON COMPUTER VISION

Contrary to the obtrusiveness of EMG-based approaches, the analysis of facial behavior using automated visual methods can be performed remotely and without physical contact. The process usually involves the use of computer vision to perform face detection, the localization of facial features (also known as landmarks or fiducial points) and the classification of such information into facial expressions (Salah, Sebe, Gevers, et al., 2010).

Computer vision systems usually rely on image processing, artificial intelligence, e.g.
Figure 4.1: Facial muscles. (a) Corrugator supercilii. (b) Orbicularis oculi. (c) Zygomaticus minor. (d) Zygomaticus major. Adapted from “Sobotta’s Atlas and Textbook of Human Anatomy”, by Dr. Johannes Sobotta (Illustration: K. Hajek and A. Schmitson), 1909. Reproduced from (Wikimedia Commons, 2013).

machine learning, and decision-making techniques to detect and classify objects from images or videos. A class of such objects is the human face, detected and analyzed by a process called facial alignment.

Facial alignment consists of identifying the position of specific features of the face, e.g. eye and nose, after the face has been detected in an image/video. Figure 4.2 demonstrates the process. This procedure is relevant in many different scenarios, for example, facial/expression recognition and pose estimation. Research has been conducted to create accurate and fast methods that can be used to perform face alignment under an ever growing set of challenging conditions, e.g. face movement combined with different lighting configurations.

Many methods have been proposed and a literature review reveals that two basic approaches are widely used in alignment techniques: constrained local models and cascaded regression methods. Both approaches work on an image of a face contained within a rectangle obtained by a face detection algorithm, such as Viola & Jones (Viola and Jones, 2004). The following sections describe each one and mention the most relevant techniques for facial alignment that are based on the approach being described.
4.2.1 CONSTRAINED LOCAL MODEL

The Constrained Local Model (CLM) approach consists of locating a set of points on a target image, then applying a constraint to them. The constraint is usually based on a statistical shape model, which is obtained via training in a set of images featuring manually inserted landmarks. Since the shape model is statistical, the position of the points (landmarks) that it describes will always resemble a face, i.e., the proportions of lines and/or the distance between points will not be so different from a human face (at least not different from the ones found in the training set). Figure 4.3(a) illustrates the configuration of the shape model with different variations.

Figure 4.3: Shape models of CLM. (a) Configuration of the shape models with different variations. (b) Shape models and their respective entries in the texture model. Reproduced from Yu (2010).

In addition to the shape model, there is a texture model that contains a set of patches (images) extracted from the training images by selecting the areas around the inserted landmarks. These patches are used to guide the search procedure in the alignment process, which allows the technique to correctly identify the right model to properly align...
the face being analyzed. Figure 4.3(b) shows different shape models and their respective entries in the texture model.

The process of aligning a face is iterative and starts by sampling points that are placed in the face image according to the current shape estimation of the face. In the first try, this estimation is usually the average face obtained from all training images. The area around the sampled points are extracted and used in a search to locate a set of similar patches in the texture model. The current shape estimation and the texture patches it locates are evaluated according to a cost function. As soon as the shape variation with the minimal cost is found, the process is repeated: new patches are sampled and searched against the texture model, the current estimation is adjusted and so on. Eventually, the cost function will not produce a significantly different value from one iteration to another, which means the current estimation is the best match found.

Figure 4.4 demonstrates the evolution of the technique as it iterates in an image. First the mean shape is placed into the image and the patches are sampled around the (mistakenly) positioned landmarks. As the technique iterates, searched patches progressively induce changes in the current shape model, sampling more accurate patches. Eventually the technique converges to the aligned face.

Figure 4.4: Iteration of CLM during the alignment of an image. Reproduced from Cristinacce and Cootes (2006).

Two techniques that represent the CLM approach are the feature detection and tracking with constrained local models (Cristinacce and Cootes, 2006) and its 3D variation (Baltrusaitis, Robinson, and Morency, 2012), which uses 3D depth data to improve the process.

4.2.2 CASCADED REGRESSION

The cascaded regression methods approach consists of using an initial guess shape that is progressively refined into the final answer (identification of key features in the image). This refinement is performed in a stage-by-stage manner (cascade) and the result of the current stage is used as the input for the next one. In each stage, the adjustment of the current shape (into the aligned result) is performed by a regression function, learnt via training. Early regressors in the cascade handle large variations in the shape, as opposed to the late ones, which focus on specific details. Each regressor extracts features from the image, which are then worked to produce variations in the current guess shape. The extracted features depend on the current shape and they are commonly referred as shape-indexed features.

The shape-indexed features are differences in pixel intensities. The calculation of a shape-indexed feature involves the selection of a few pixels and the subtraction of their
intensities. The way these pixels are selected is usually different for each of the cascaded-regression techniques. Figure 4.5 illustrates an example of the selection of a few pixels for three particular landmarks. The landmark on the top-right (gray circle) highlights the selected pixels that will be used. The difference in intensities among these pixels will define this particular shape-indexed local feature. The shape-indexed local features are used in a decision process in each step (cascade), as illustrated by Figure 4.6. Usually the initial guess shape is the mean shape of the training set. This guess is used to calculate the current set of shape-indexed features to be extracted, which then guides the variation applied to the current shape. The variation to be applied is usually chosen on the basis of the result of a cost function, which selects a variation that minimizes the distance between the current guess and the supposed aligned face. As the process repeats itself, different shape-indexed features are selected, a new variation is calculated and so on. Eventually, the current shape will converge and will represent the alignment for the face being analyzed (final shape estimation).

Figure 4.5: Pixels used in the calculation of a shape-indexed feature. Reproduced from Maris (2015).

Different variations are used to handle the training, the extraction of features and the way the regression is performed. The technique of face alignment with Ensemble of Regression Trees (ERT) (Kazemi and Sullivan, 2014), for instance, estimates the face’s landmarks by inputting the regressors with a sparse subset of pixels’ intensities, which is calculated with a prior probability on the distance of the pixels.

4.3 FACIAL-BASED EMOTION DETECTION

Facial-based emotion detection techniques try to estimate the emotional state of a subject, based on the analysis of facial features. Works involving such an approach commonly focus on detecting or classifying emotional states, based on the six basic emotions proposed by Ekman and Friesen (1971), i.e. happiness, surprise, sadness, fear, anger and disgust. Some visual methods rely on manual or automated FACS-based analysis as a standard for categorizing and measuring emotional expressions (Bartlett et al., 1999). The feasibility of that approach is demonstrated by Kaiser, Wehrle, and P. Edwards
Figure 4.6: Estimation of face shape with regressors in a set of stages. Reproduced from Maris (2015).

(1994), showing that more AU were reported by manual FACS coders during the analysis of video recordings of subjects playing the stressful part of a game compared to its neutral part. Additionally, authors report lip pull corner and inner/outer brow raise as more frequent AUs during gaming sessions. Wehrle and Kaiser (2000) also support the approach of using an automated, FACS-based facial analysis together with data from game events to provide an appraisal analysis of subjects' emotional state.

Similarly, Grafsgaard et al. (2013) present an experiment where facial expression information is used to investigate emotional states. The experiment consists of an analysis of facial AUs during computer mediated, tutoring sessions among students. Subjects and tutors interact through a tutoring software related to computer programming, while subjects are recorded. After each session, subjects answer a questionnaire related to the measurements of cognitive load and engagement. The recordings are analyzed in an automated way with manual verification of the results. A predictive model is then constructed using the questionnaire’s answers and the recording analysis, which results in correlations of facial AU and emotional states. Finally, the authors compare their findings against other research, which differ significantly. For instance, brow lowering has been correlated with confusion in previous work; however, the authors found that it was a positive predictor of student frustration in the context of their experiment.

Heylen et al. (2005) also present a similar investigation in a pilot experiment of a tutoring session related to the application of a subcutaneous injection. Students interact with a virtual patient while using a physical haptic device to administrate an injection. The recordings of the students are analyzed by the researchers to make annotations of the expressions based on their own interpretation of the context. The researchers use a compilation of literature components to guide the evaluation of the collected data. As the authors point out, a variety of expressions occur, but most of the time students maintain a neutral facial expression. The annotated features are (ordered from most to less frequent): smile (total 22), raise eyebrows (11), pull down mouth corners (2) and frown
In contrast to the previously mentioned works, other initiatives aim to detect emotions based on the analysis of facial landmarks and their distances, movement or angles. Joho et al. (2009) use facial analysis for affective video summarisation. The authors use an automated face tracking approach to obtain a vector of motion features of certain regions of the face, named Motion Units (MU). MUs are classified by a Bayesian network, which is trained from labeled data. The results indicate a promising correlation between the manually annotated content of the videos and the automatically classified one.

Somewhat similarly Akakın and Sankur (2010) detect facial landmarks over consecutive frames of videos, whose trajectories (time series) during head gestures and facial expressions are organized in a spatiotemporal matrix. Discriminative features are extracted from the trajectory matrix and used to train machine learning models, i.e. Adaboost and SVM. Classification accuracy is reported to be around 90% for the detection of 7 face and head gestures in a dataset composed of 210 videos of 4 subjects. The detection of emotions, however, is limited to only two states, i.e. happiness and sadness.

Finally, Samara et al. (2016) present the sensing of affective states based on the analysis of the distances of facial landmarks. Automated face detection is employed to detect facial landmarks, however no coding scheme is used to identify the detected points. Instead, the authors use the Euclidian distance between facial landmarks, represented as distance vectors, to train a SVM model to detect expressions. Figure 4.7 illustrates the process. Detection accuracy is improved by a two-state SVM classification model, which the authors name Hierarchical Parallelised Binary Support Machines. Accuracy rates of about 96% were achieved on two facial expression datasets. Samara et al. (2016) also use the Euclidean distance between face points to train a Support Vector Machine (SVM) model to detect expressions. Similarly, Chang et al. (2009) use 12 distances calculated from 14 landmarks to detect fear, love, joy and surprise. Hammal et al. (2007) use 5 facial distances calculated from lines in key regions of the face derived from the MPEG-4 animation standard (Abrantes and Pereira, 1999), e.g. eyebrows, for classification of expressions. Tang and T. S. Huang (2008a) and Tang and T. S. Huang (2008b) use up to 30 Euclidean distances between facial landmarks also obtained from MPEG-4 based 3D face models to recognize the 6 universal facial expressions. Similarly Hupont, Baldas-
sarri, and Cerezo (2013) classify the same emotions by using a correlation-based feature selection technique to select the most significant distances and angles of facial points. The use of facial expressions as a single source of information, however, is contested in the literature. Blom et al. (2014) claim that subjects present a neutral face during most of the time of gameplay and frustration is not captured by facial expressions, but by head movements, talking and hand gestures instead. In a similar conclusion, Shaker, Asteriadis, et al. (2011) show that head expressivity, i.e. movement and velocity, is an indicator of players’ game skills. Additionally, a high frequency level and velocity of head movements is indicative of failing in the game.

4.4 SUMMARY

Facial analyses based on physical sensors, e.g. EMG, provide continuous monitoring of subjects and are not affected by lighting conditions or pose occlusion by a subject’s movement. However, they are obtrusive and the use of sensors increases the user’s awareness of being monitored (Yamakoshi et al., 2007; Yamaguchi, Wakasugi, and Sakakima, 2006; Healey and Picard, 2005). Approaches based on video analysis, e.g. FACS and computer vision, are less intrusive. Despite the fact that FACS has proven to be a useful and quantitative approach for measuring facial expressions (Bartlett et al., 1999), its manual application is laborious, time-consuming and requires certified coders to inspect the video recordings. The application of FACS also has downsides, including different facial expression decoding caused by misinterpretation in specific cultures (Jack, 2013).

Facial analyses from visual methods, such as the previously mentioned feature-based approaches relying on computer vision, are quicker and easier to deploy. Automated facial analysis as a mean to detect the emotional state of players has been proven to work at creating emotionally adapted games (Saari and Turpeinen, 2004) or tools for unobtrusive game research. When automated facial analysis is used, it is often tested on contexts not related to games, or facial cues are derived from models not designed for the analysis of emotional interactions in games, such as the MPEG-4 standard (Abrantes and Pereira, 1999). Such a standard specifies representations for 3D facial animations, not emotional interactions in games. Automated facial analysis is also commonly performed on images or videos whose subjects are acting to produce facial expressions, which are likely to be exaggerated in nature and not genuine emotional manifestations. These are artificial reactions that are unlikely to happen in a context involving subjects interacting with real games, where emotional involvement between subject and game is stronger. Another limitation of previous work is the common focus on detecting facial expressions per se, e.g. 6 universal facial expressions (Ekman and Friesen, 1971), not necessarily detecting isolated facial actions, e.g. frowning, associated with emotional reactions in games. Finally, people are different and elements such as age and familiarity with a game influence the outcome of automated facial analysis of behavioral cues (Asteriadis et al., 2012). Furthermore different games might induce different bindings of facial expressions (Tan, Bakkes, and Pisan, 2014a). Empirical results of manual annotations of facial behavior in gaming sessions have indicated more annotations during stressful than during boring (see Section 8.5) or neutral (Kaiser, Wehrle, and P. Edwards, 1994) parts of games. As a consequence, a more user-tailored contextualization is essential for any study involving facial analysis, particularly involving games.
CHAPTER 5
EMOTIONS AND PHYSIOLOGICAL SIGNALS

The empowerment of computers for the understanding of emotions is one of the focus of human-computer interaction research. Employing the physiological signals of users in the process of detecting emotions involves different input signals, such as electrocardiogram data, skin temperature and electro-dermal activity, among others. A significant number of works explore such inputs and the challenges associated with them regarding emotion recognition using physiological signals (Jerritta et al., 2011).

Table 5.1: Most common psychophysiological measurements used in human interaction studies (Jerritta et al., 2011)

<table>
<thead>
<tr>
<th>Respiratory system</th>
<th>Breaths per minute</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Respiration volume</td>
</tr>
<tr>
<td>Electrodermal activity</td>
<td>Skin conductance (SC)</td>
</tr>
<tr>
<td></td>
<td>Galvanic skin response (GSR)</td>
</tr>
<tr>
<td>Brain activity</td>
<td>Electroencephalography (EEG)</td>
</tr>
<tr>
<td></td>
<td>Brain imaging methods</td>
</tr>
<tr>
<td>Muscular system</td>
<td>Electromyography</td>
</tr>
<tr>
<td>Cardiovascular system</td>
<td>Heart rate variability (HRV)</td>
</tr>
<tr>
<td></td>
<td>Respiratory Sinus Arrhythmia (RSA)</td>
</tr>
<tr>
<td></td>
<td>Cardiac output</td>
</tr>
<tr>
<td></td>
<td>Inter beat interval (IBI)</td>
</tr>
<tr>
<td></td>
<td>Blood pressure (BP)</td>
</tr>
</tbody>
</table>

One of the reasons physiological signals are interesting for emotion recognition is because suppressing emotions or social masking through physiological signals is impossible (K. H. Kim, Bang, and S. R. Kim, 2004). In the games research field, there have been studies aimed at using such theory and psychophysiological methods (Kivikangas et al., 2011). It has been demonstrated that it is possible to use physiological signals to automatically assess different emotional states in the context of games by using a variety of signals (Bousefsaf, Maaoui, and Pruski, 2013b; Yun et al., 2009; Rani et al., 2006; Tijs, Brokken, and Ijsselsteijn, 2008).

Different approaches exist for emotion elicitation stimuli, feature extraction and classification methodologies regarding emotion recognition. Table 5.1 presents the most common psychophysiological measurements used in human interaction studies. The aim of this thesis is to remotely obtain physiological information from users, in order to detect emotional states. As a consequence, the input signals presented in Table 5.1 were selected on the basis of their potential to classify emotional states of users and the existence of techniques that are able to remotely and accurately measure them in a
context involving games and natural behavior. HR was identified as a reliable physiological signal for emotion detection which can be remotely estimated via video of subjects. The application of HR has been confirmed by previous work, including its use for the measurement of emotional states (Kivikangas et al., 2011), and the detection of emotions such as stress (Choi and Gutierrez-Osuna, 2009) and boredom (Yamakoshi et al., 2007). Additionally, computer games were proven to provoke change in the mean HR of players at stressful periods of gameplay (Sharma et al., 2006; Rodriguez et al., 2015), which makes HR an ideal signal for the method presented in this thesis.

The following sections present more details regarding the physiology of HR and its use in emotion detection.

5.1 PHYSIOLOGY OF HEART RATE

The cardiac cycle is periodic and contains a set of waves that reflect depolarization and repolarization events, which are all connected to the functioning mechanism of the heart (Yanowitz, 2012). Figure 5.1 illustrates two cardiac cycles with highlights on each of such waves. The x-axis represents time in milliseconds and the y-axis represents the wave amplitude. A heartbeat is a contraction and relaxation of the heart, which is connected to the Q, R and S waves, known as the QRS complex.

Figure 5.1: Two cardiac cycles with highlights on each of its electrical waves. The X axis represents time in milliseconds and the Y axis represents the wave amplitude. A heartbeat is connected to the Q, R and S waves, known as the QRS complex. Reproduced from Ahmed, Begum, and Islam (2010).

The tracing of the QRS complex allows the monitoring of the heart rate. The peak of the R wave, referred to as R peak, marks a specific time in the cardiac cycle. The difference between two R wave peaks is the RR interval, also known as the inter-beat interval (IBI). The RR interval reflects the entire duration of each heart beat. As a consequence, the amount of R waves within one minute is used to calculate the HR. The variation in time between R peaks, which is the beat-to-beat variability of HR, is referred to as the heart rate variability (HRV).
The HR is connected to the central nervous system and can be influenced by a series of different elements, including modifiable and non-modifiable ones (Valentini and Parati, 2009). Modifiable elements are those related to external influence, such as mental stress or physical activity. Non-modifiable elements are related to physiology, such as age, gender and race. A healthy human adult has a HR within the interval of [45 bpm, 240 bpm], the equivalent of [0.75 Hz, 4 Hz] (Li et al., 2014).

5.2 HEART RATE, STRESS AND FRUSTRATION

The foundation of HR-based approaches for emotional estimation draws on the theory that physiological signals are linked to emotion regulation (Appelhans and Luecken, 2006; Fenton-O’Creevy et al., 2012; Schubert et al., 2009). The automatic nervous system (ANS), a key system in the generation of physiological arousal (Appelhans and Luecken, 2006), is subdivided into the sympathetic nervous system (SNS) and parasympathetic nervous system (PNS). These systems interact (often antagonistically) to produce variations in the physiological signals, to prepare a person to react to a situation. The SNS is dominant during physical and psychological stress, triggering the body to alertness, by increasing the HR, for instance. This is commonly referred to as the “fight or flight” response. The PNS, on the other hand, is dominant in periods of stability and relative safety, maintaining physiological signals at a lower degree of arousal, e.g. by decreasing the HR. The continuous changes between the SNS and PNS impulses cause variations of HR and HRV (Schubert et al., 2009), which refers to beat-to-beat alternations in HR intervals.

Physiological signals, such as HR, are hard to fake due to their link to the ANS, in contrast to facial expressions (Landowska, 2014), for instance. As a consequence, HR and its derivatives, such as HRV, have been used as reliable sources of information in different emotion estimation methods (Kukolja et al., 2014). Additionally, it has been used as an indication of perceived interest and confusion in mobile applications (Xiao and J. Wang, 2015), as player input for games (Stockhausen, Smyzek, and Krömker, 2013), and as triangulation of psychophysiological emotional reactions to digital media stimuli (Nogueira et al., 2015). The significant variation of physiological signals, however, is an obstacle to their use in emotion estimation. Signals increase during emotional arousal, but decrease in response to attention engagement, which makes the measurement of engagement, for instance, a non-trivial process (Ravaja et al., 2006). Despite such challenges, the use of HR and HRV has been demonstrated in a continuous arousal monitor (Grundlehner et al., 2009) as well as a detection mechanism for mental and physical stress based on physical and mental tasks (Vandeput et al., 2009; Garde et al., 2002). The results indicate a higher HR during the mentally demanding task, compared to the rest period. The HR mean alone has also been shown to be a measurement of frustration in a game (Rodriguez et al., 2015).

Vandeput et al. (2009) demonstrate the use of HR and HRV to detect mental and physical stress. Three demanding activities (a postural task, a mental task and a task that is a combination of both) are used and for almost all the HR measures obtained, the demanding activities can be distinguished from the rest period. The authors also point out that mental stress decreased high frequency components of the HRV interval, i.e. $HRV_{HF}$, while increasing low frequency ones, i.e. $HRV_{LF}$. A similar experiment conducted by Garde et al. (2002) involved two tasks: one mental and physically demanding, i.e. digital version of the Stroop color word test (Golden, 1978), while the other was only physically demanding. The authors confirmed the findings of Vandeput et al. (2009) by showing higher HR, increased $HRV_{LF}$ and decreased $HRV_{HF}$ during the mentally demanding task.
demanding task, compared to the rest period.

Bousefsaf, Maaoui, and Pruski (2013b) also use a digital version of the Stroop test in a similar experiment. A combination of HR, HRV and $HRV_{HF}$ is used to estimate the stress state of subjects. The resulting stress state curve tends to decrease during the rest period and increase during stress sessions (as did the HR), in accordance with the previously mentioned works; additionally, the self-reported answers to questionnaires present significant differences in stress level in the rest and the stress sessions as well. McDuff, Gontarek, and Picard (2014b) and McDuff, Hernandez, et al. (2016) also use HR and its variants in order to measure cognitive stress during computer tasks. According to the authors, the average heart rate and breathing rate are not significantly different in any case, which differs from the findings of previously mentioned work. The variations of $HRV_{LF}$ and $HRV_{HF}$, however, are significantly different during the cognitive tasks compared to the rest period; higher $HRV_{LF}$ and lower $HRV_{HF}$ power are found in both cognitive tasks compared to the rest period, which aligns with the findings of previous work. Stress predictions made by the model are consistent with the self-reported answers provided by subjects.

Grundlehner et al. (2009) present a real-time, continuous arousal monitor. Using a wireless sensor network for signal acquisition, the authors record and use four signals from subjects to estimate arousal: electrocardiogram (ECG), respiration, skin conductance and skin temperature. The ECG is applied to calculate HRV, which is then used in the estimation. A regression analysis is performed to identify the importance of the features in the estimation of arousal. $HRV_{LF}$ and $HRV_{HF}$ are not significant compared to the other signals, e.g. skin conductance, while the standard deviation of HRV presents a significant weight. The arousal prediction matches the hypothesized arousal events marked by the authors in each of the experiment parts, e.g. start of noise in the audio part.
CHAPTER 6
REMOTE PHOTOPLETHYSMOGRAPHY: NON-CONTACT HEART RATE MEASUREMENT

Photoplethysmography (PPG) is a technique commonly used to measure HR on the basis of the variations of light absorption in the human skin. The pressure of the cardiac activity causes the blood vessels to change volume and light absorption rate due to the levels of oxygen in the blood flow. Such differences make the light absorption on the skin surface change accordingly. PPG is a time-varying signal resulting from such differences in the light absorption in live human tissue, which can be processed to calculate the HR. The process is illustrated in Figure 6.1. The employment of PPG requires a physical sensor, e.g. finger pulse oximeter, in order to be performed.

Further research on PPG (McDuff, Estepp, et al., 2015) evolved the technique to allow it to be performed remotely, by basing it on the analysis of a video of a person. The remote approach is commonly referred to in the literature as remote photoplethysmography (rPPG) (Allen, 2007). This improvement removed the requirement of a physical sensor or any physical contact for the estimation of HR and its derivatives. A literature review shows the existence of a variety of different rPPG approaches, including thermal-, image-, and movement-based ones (Kranjec et al., 2014; Sereevoravitgul and Kondo, 2014). The following sections present the common structure of an rPPG technique, a survey of existing rPPG techniques and information regarding the accuracy and limitations of such technology.

6.1 STRUCTURE OF THE TECHNIQUE

There are initiatives to classify (Rouast et al., 2016; McDuff, Estepp, et al., 2015) and formally model the algorithmic principles (W. Wang, Brinker, et al., 2016) of rPPG tech-
6.1.1 SIGNAL EXTRACTION

The signal extraction phase extracts a set of raw signals from a video. Step 1 relates to the detection of a region of interest (ROI), which is an area of the video that usually contains the face of the subject. Approaches commonly used for this step are the VJ\(^1\) algorithm of Viola and Jones (2004), a machine learning-based approach to classifying

\(^1\)VJ is the abbreviation of Viola & Jones, which is how the algorithm is commonly mentioned in the literature.
faces, Active Appearance Models (AMM) (G. J. Edwards, Taylor, and Cootes, 1998) and facial landmark detectors.

In step 2 (ROI definition) the parts of the ROI to be used for the signal extraction are defined. Approaches commonly employed include the use of whole ROI (usually the bounding box returned by VJ, which is likely to contain background pixels along with the facial pixels), a part of the ROI (e.g. 60% of the bounding box width) or specially defined areas (e.g. forehead, cheeks or shapes defined by facial landmark points).

Step 3 (ROI tracking) deals with the process of tracking the defined ROI area over the duration of the video. Ideally the PPG signal should be extracted from the pixels belonging to the same skin region over time. This is unlikely to happen due to subject motion, be it voluntary or not, i.e subjects will present movement that influences the ROI even when still (Poh, McDuff, and Picard, 2010). A common approach for this step is the re-detection of the ROI for each frame, which is computationally suboptimal and still prone to noise. Detection algorithms, e.g. VJ, are not likely to be exact, therefore results between consecutive frames might be slightly different, which causes fluctuation in the defined ROI. More elaborated approaches avoid the costly frame-by-frame re-detection procedure (and its fluctuations) by updating an already detected ROI in subsequent frames. The update is based on movement information obtained from tracking algorithms applied to features/points within the ROI.

Finally, in step 4 (raw signal extraction) a raw (untreated) signal is extracted. It is a time-varying signal whose points/samples are obtained from the content of the defined ROI in each frame of the video. One common approach used to extract the signal is based on colors, e.g. the average of the values of the pixels of each channel (R, G and B) over time produces the raw signal. Another approach is based on the movement of the head, which is a mechanical reaction to the blood flow in the aorta. Feature points within the defined ROI are tracked over time by a tracking algorithm and the horizontal and/or vertical variations in the trajectory of the points produce the raw signal. The signal extraction phase can result in multiples raw signals, which are dependent on the approach applied. A color-based approach, for instance, might result in three raw signals, one for each color channel (RGB).

6.1.2 SIGNAL ESTIMATION

The raw signals are filtered and combined to produce a single, estimated plethysmographic signal. The filtering step aims to remove unwanted noise from the raw signal, which could be caused by the capturing device (e.g. camera noise), subject movement, changes in illumination, etc. The raw signal is usually normalized (subtracted by its mean and divided by its standard deviation) first. The commonly used filters are based on high/low pass, such as bandpass, moving average window and detrending based on Smoothness Priors Approach (SPA) (Eleuteri et al., 2012). Since the feasible frequencies for the HR band are known, filters are usually configured to eliminate frequencies outside that band, e.g. cut-off frequencies of [0.7 Hz, 4.0 Hz], which eliminates, from the raw signal, frequencies outside the [42 bpm, 240 bpm] interval.

In the dimensionality reduction step, the raw signal is used to estimate the plethysmographic signal, i.e. the one containing the HR. As previously mentioned, depending on the approach used for the raw signal extraction, one or more raw signals might be available for use in the estimation. One approach for the estimation is to simply choose one of the raw PPG signals as the estimated plethysmographic signal, e.g. the raw signal extracted from the G channel. This choice is acceptable since the plethysmographic signal
is known to be stronger in the green channel of an RGB video (Verkruysse, Svaasand, and Nelson, 2008); however, it is more likely to contain noise since a raw (untreated) component is being used. Another approach relies on Blind Source Separation (BSS) methods, which try to mitigate noise by assuming the estimated plethysmographic signal is a linear combination of the raw signals. Independent Component Analysis (ICA) (Hyvärinen and Oja, 2000) and Principal Component Analysis (PCA) (Jolliffe, 2011) are common BSS algorithms used to find the weight that each raw signal has in the linear combination to produce the estimated plethysmographic signal. Finally, another approach assumes fixed weights for the linear combination, which are derived from models of skin illumination, for instance.

6.1.3 HEART RATE ESTIMATION

Finally, the estimated plethysmographic signal is analyzed and the HR is extracted. There are two steps in this phase, frequency analysis and peak detection; however, only a single one of them is usually performed. In the frequency analysis step, approaches include Continuous Wavelet Transform (CWT), used to construct a time-frequency representation of the plethysmographic signal, and Fourier Transform to analyze the signal in the frequency domain. Regarding the latter, the estimated plethysmographic signal is assumed to contain a distinct periodicity (the HR). As a result, when converted into the frequency domain via Fast Fourier Transform (FFT), for instance, the signal should present a high spectral power associated with the frequency of such a distinct periodicity. As a consequence, the index of the frequency with the highest peak in the power spectra of the frequency domain corresponds to the HR.

In the peak detection step, however, the signal is not converted into the frequency domain, instead, it is interpolated and peaks are detected (usually with a local maxima algorithm). The distances between the peaks correspond to the inter-beat interval (IBI), which is then used to calculate the HR (as \(60/IBI\)) and/or the heart rate variability (HRV).

6.2 CONSOLIDATED TECHNIQUES

The previously described general algorithm framework for rPPG techniques illustrates the wide range of variations possible by different combinations of approaches within each step of the process. Early initiatives regarding rPPG, for instance, used manually detected/defined ROI, little or no use of signal filtering, and signal extraction based on the average of image brightness (Takano and Ohta, 2007) or colors (Verkruysse, Svaasand, and Nelson, 2008).

The work of Poh, McDuff, and Picard (2010) was the first in the field to rely on BSS. Figure 6.3 illustrates the proposed approach. The authors extracted the signal by using automatic definition/tracking of the ROI (based on 60% width of a VJ bounding box) and the average of the color channels. The three raw signals (derived from R, G and B) were filtered, detrended and decomposed using ICA. The second component generated by ICA was interpolated and a custom algorithm detected peaks to identify the HR. The authors further improved the technique by selecting the ICA component whose power spectrum contained the highest peak (Poh, McDuff, and Picard, 2011) and by incorporating alternate frequency bands (i.e. orange and cyan) into the extraction phase (McDuff, Gontarek, and Picard, 2014a).
Figure 6.3: Overall structure of rPPG approach based on ICA. Adapted from Poh, McDuff, and Picard (2010).

Datcu et al. (2013) present a similar approach, however, they use AAM to segment the face of the subject into ROIs. Li et al. (2014) also use a different ROI, based on facial landmarks, and a combination of additional steps to mitigate noise caused by illumination, motion and facial expressions by removing signal outliers. Bousefsaf, Maouci, and Pruski (2013a) propose a variation of the previously described approaches, by using a skin detection procedure to select pixels in the signal extraction phase. Additionally, CWT is used in the signal estimation phase instead of the commonly used FFT, which authors claim is more suitable for rapid changes in frequencies in time. As a consequence, the authors were able to detect the instantaneous HR (iHR) and significantly reduce the waiting time for the detection of HR measurements.

Different approaches that are not based on the average of colors of the face can also be found in the literature. Such techniques use knowledge of the color vector of the different components to perform the signal extraction. For example, W. Wang, Stuijk, and Haan (2016) focus on the definition of a plane orthogonal to the skin-tone, ignoring pixels outside the subspace of skin pixels in the signal estimation phase. Similarly, Haan and Jeanne (2013) also use a skin model to generate a chrominance-based PPG signal, which is calculated as a combination of the intensities of the color channels in the video instead of their average. These techniques aim to be more resilient than BSS-based techniques regarding motion noise. Balakrishnan, Durand, and Guttag (2013) are the first to completely move away from color-based initiatives and perform the signal extraction on the basis of head movements. Irani, Nasrollahi, and Moeslund (2014) further improve the technique by using a moving average filter applied to the trajectory of the feature points being tracked, in order to remove the noise produced by other sources of motion, e.g. respiratory activity.

The different settings in each phase of an rPPG technique result in a trade-off between advantages and disadvantages. Estimation based on head movement, e.g. Balakrishnan, Durand, and Guttag (2013), does not rely on previous knowledge about colors or require visible skin area to work. However, it is outperformed by other methods when subjects are not completely still (Li et al., 2014) since it is affected significantly by subject motion. Techniques based on pre-defined skin-tone models, e.g. W. Wang, Stuijk, and Haan...
and Haan and Jeanne (2013), better adapt to changes in illumination (including non-white light sources); however, they suffer performance degradation when the skin mask is not properly defined (or is noisy) or the pre-defined skin model is inaccurate (W. Wang, Brinker, et al., 2016). Finally, BSS-based methods, e.g. Poh, McDuff, and Picard (2011), rely on BSS techniques (e.g. ICA) which are ideal to de-mix the estimated PPG signal from noise. However, such techniques are unable to deal with periodic motion, i.e. exercise situation, and the technique’s statistical nature requires a long signal to enable an accurate measurement (W. Wang, Brinker, et al., 2016). Despite such limitations, the ICA-based rPPG technique by Poh, McDuff, and Picard (2011) presents the best signal-noise ratio (SNR) for HR estimation under stationary situations, i.e. non-exercising, with different illumination conditions, compared to other techniques (W. Wang, Stuijk, and Haan, 2016). The work is also extensively cited in the literature and often used as a benchmark for new techniques, which makes it a consolidated and robust solution.

6.3 ACCURACY AND LIMITATIONS

The structure of rPPG techniques, as previously described, requires the analysis of each frame of a video to estimate the HR. As a consequence, any rPPG technique is influenced by the frame rate of the video, which accounts for the number of frames displayed per second (expressed in frames per second, or FPS), the resolution of each of those frames and the number of subsequent frames required to allow an estimation.

The frame rate of the video used in the estimation is directly connected to the sampling frequency, named $F_s$. Assuming that all frames of a video are used by an rPPG technique, a video running at 30 FPS allows a $F_s$ of 30 Hz, since each frame generates a sample. The minimum FPS required for the estimation must adhere to the Nyquist limit, which states that the sampling frequency must be at least twice as high as the highest frequency being measured:

$$F_s \geq 2 \cdot f_{Max}$$

where $f_{Max}$ denotes the highest frequency being measured. Since rPPG techniques are used to estimate HR, the value for $f_{Max}$ can be derived from a valid HR frequency interval. As previously described, a normal person presents a HR between the interval $[45 \text{ bpm}, 240 \text{ bpm}]$, which is equivalent to the interval of $[0.75 \text{ Hz}, 4 \text{ Hz}]$. As a consequence, $f_{Max}$ is 4 Hz (240 bpm) and $F_s$ is calculated as $F_s \geq 2 \cdot 4$, resulting in a minimum $F_s$ of 8 Hz, which translates to a video with a frame rate of 8 FPS.

A Fourier transform, e.g. FFT, is usually employed in rPPG techniques. The FFT divides the frequency spectrum of the input signal into $N$ number of discrete frequency blocks equal to the number of samples in the input signal. The number of samples $N$ used for the estimation is commonly referred to as window size. The smallest frequency that can be differentiated within that spectrum is named frequency resolution, $\Delta f$, which is calculated as:

$$\Delta f = \frac{F_s}{N}$$

(6.1)

The frequency resolution, $\Delta f$, influences the error rate in the estimation of the HR. For instance, if $\Delta f$ is 0.1 Hz, the HR will be estimated with an error of ±0.3 bpm. A low HR estimation error is desired, so a low value for $\Delta f$ is desired as well. By analysis of equation 6.1, it is noted that $\Delta f$ decreases when $F_s$ decreases or when $N$ increases.
The value for $F_s$ is usually dependent on the camera device being used, which follows certain standards, e.g. 30 FPS. As a consequence, $\Delta f$ can be altered by changes on the window size $N$. Such changes, however, directly impact the estimations, since there is a trade-off between the window size and the estimation error. The higher the window size, the longer the video segment required for the estimation. Longer video segments are likely to contain movement of the subject being analyzed, which adds noise to the rPPG estimation.

Roald (2013) presents this trade-off by detailing the frequency resolution as a function of window size and video frame rate, demonstrated in Table 6.1. For example, a video of 50 FPS and a frequency resolution of 0.07 Hz (which corresponds to an estimation error of $\pm 2.1$ bpm), would require a window size of 700 samples (equivalent of a video segment of 14 seconds). Additionally, it has been proven that limitations in the sampling frequency, e.g. low $F_s$, can be compensated in the PPG signal detection by the use of interpolation (Sun et al., 2012).

<table>
<thead>
<tr>
<th>Window size</th>
<th>10 FPS</th>
<th>15 FPS</th>
<th>25 FPS</th>
<th>30 FPS</th>
<th>50 FPS</th>
<th>60 FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.1</td>
<td>0.15</td>
<td>0.25</td>
<td>0.3</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>200</td>
<td>0.05</td>
<td>0.075</td>
<td>0.125</td>
<td>0.15</td>
<td>0.25</td>
<td>0.3</td>
</tr>
<tr>
<td>300</td>
<td>0.033</td>
<td>0.05</td>
<td>0.083</td>
<td>0.1</td>
<td>0.17</td>
<td>0.2</td>
</tr>
<tr>
<td>400</td>
<td>0.025</td>
<td>0.037</td>
<td>0.062</td>
<td>0.075</td>
<td>0.12</td>
<td>0.15</td>
</tr>
<tr>
<td>500</td>
<td>0.02</td>
<td>0.03</td>
<td>0.05</td>
<td>0.06</td>
<td>0.1</td>
<td>0.12</td>
</tr>
<tr>
<td>600</td>
<td>0.017</td>
<td>0.025</td>
<td>0.042</td>
<td>0.05</td>
<td>0.08</td>
<td>0.1</td>
</tr>
<tr>
<td>700</td>
<td>0.014</td>
<td>0.02</td>
<td>0.03</td>
<td>0.04</td>
<td>0.07</td>
<td>0.08</td>
</tr>
<tr>
<td>1000</td>
<td>0.01</td>
<td>0.015</td>
<td>0.025</td>
<td>0.03</td>
<td>0.05</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Roald (2013) presents this trade-off by detailing the frequency resolution as a function of window size and video frame rate, demonstrated in Table 6.1. For example, a video of 50 FPS and a frequency resolution of 0.07 Hz (which corresponds to an estimation error of $\pm 2.1$ bpm), would require a window size of 700 samples (equivalent of a video segment of 14 seconds). Additionally, it has been proven that limitations in the sampling frequency, e.g. low $F_s$, can be compensated in the PPG signal detection by the use of interpolation (Sun et al., 2012).
CHAPTER 7
MULTIFACTORIAL EMOTION ESTIMATION

Previous chapters demonstrate the separate use of physiological signals and facial information to emotion estimation. While the use of a single signal for the assessment is feasible, a mapping process based on a multifactorial analysis, when more than one signal is used, is more likely to produce accurate results (Kukolja et al., 2014). A combination of signals can reduce the interference/noise caused by signal manipulation.

Physiological signals, e.g. HR, are considered reliable sources since they are hard to fake (because of their link to the ANS), in contrast to facial expressions (Landowska, 2014), for instance. When combined in the same analysis, however, these signals can complement each other and provide more information about emotional states.

The following sections present works related to multifactorial emotion estimation categorized by contact (use of physical sensors) and non-contact (remote analysis) approaches.

7.1 APPROACHES BASED ON PHYSICAL CONTACT AND SENSORS

The first multifactorial analysis approaches were based on obtrusively obtaining measurements of the input signals, using physical sensors. Chanel et al. (2011), for instance, demonstrate the use of multifactorial input analysis to measure emotions and involvement in a game context. In their study, participants play Tetris in different conditions of difficulty while a variety of sensors, including a respiration belt and an electroencephalogram (EEG) system, monitor them. The data from the EEG and the other peripheral signals are computed into two groups of features, which are used to train classifiers to recognize the states: boredom (low pleasure and low pressure, associated with low difficulty level), engagement (higher pleasure and motivation, associated with medium difficulty level) and stress (anxiety and low pleasure, associated with high difficulty level). The classification accuracy of the model varies between 48% and 55% for different input signals, classifiers and feature selection.

Grundlehner et al. (2009) also use physical sensors to perform real-time continuous arousal monitoring. The authors record and use four signals from subjects to estimate arousal: electrocardiogram (ECG), respiration, skin conductance and skin temperature. The ECG is used to calculate HRV, which is then applied in the estimation. The data used for emotion-triggering is based on videos, sounds and a cognitive demanding game. A regression analysis is performed to identify the importance of the features in the estimation of arousal. HRV_LF and HRV_HF are not significant, compared to the other signals, e.g. skin conductance, while the standard deviation of HRV presents a significant weight. The arousal prediction matches the hypothesized arousal events marked by the authors in each of the emotion-triggering events. The results, however, were derived with controlled, pre-defined events that are expected to cause reactions, which might be related to arousal. More subtle or dynamic interactions, such as the ones obtained when a subject plays a digital game, might not be identified or detected by the approach proposed.
Figure 7.1: Non-obtrusive and remote approach based on multifactorial analysis to identify user emotion. Reproduced from D. Zhou et al. (2015).

Bailenson et al. (2008) use a combination of physiological signals (obtained from physical sensors) and facial expressions. The authors use a machine learning model to map the input signals to emotions (sadness or amusement). The training data used for the machine learning model is based on recordings of participants while they watched a video containing different emotion-triggering segments (neutral, amusement and sadness). Additionally, 15 physiological signals are also used, among them HR, skin conductance level and finger temperature. The video recordings are annotated by professional coders. The annotated video frames are used in conjunction with the physiological signals to produce the predicting model. The authors compare the performance of models built from different data sources, such as the data from all subjects (general model), from the female/male population (gender-specific model), or from a single individual (person-specific model). The model performs better when categorizing emotions instead of predicting their intensities and when detecting amusement instead of sadness. Additionally, the person-specific model outperforms the other two variations, suggesting that a person-tailored model might be more effective in identifying features (even the more subtle ones) than a general-purpose model. The results also demonstrate that a model built with a combination of facial and physiological information is more efficient than a model built with either one alone.

7.2 APPROACHES BASED ON REMOTE, NON-CONTACT ANALYSIS

D. Zhou et al. (2015) propose a completely non-obtrusive and remote approach based on multifactorial analysis to identify user emotion, illustrated in Figure 7.1. The main goal of such an approach is aimed at monitoring patients’ mental health states. The model was created by weighting different features created from input signals, all ob-
tained unobtrusively from a video of the subject’s face and the textual content he/she created during the session (e.g. a reply to a tweet). The features are physiological (HR and pupil radius), physical (head movement rate, facial expression, eye blinking rate), based on human-computer interactions (mouse and keyboard usage rate) and sentiment analysis of social content (images and textual tweets). All the features are used to train a multi-class classifier (based on logistic regression) which learns about the three possible emotion states: negative, neutral and positive. The emotion inference is performed in real-time and based on a probability rule: the state suggested by the classifier with the highest intensity is assumed to be the current subject’s emotional state. The results show an accuracy rate of 89% for negative, 56% for neutral and 78% for positive state identification. However, the authors do not specify how much each input signals contributes to the emotional output. Additionally, it is not possible to infer whether such approach could be used outside the controlled environment created by the authors, since this would require a simulation of a social network filled with previously defined and known content in order to work.

McDuff, Gontarek, and Picard (2014b) also use a camera to remotely measure cognitive stress via HRV. Participants are recorded while resting and while silently performing a mental arithmetic task. Facial landmarks are automatically detected and a ROI containing part of the subject’s face is selected for analysis. Using a spacial average of the pixel intensities in each frame, a PPG signal is calculated through independent component analysis (ICA) and the blood volume pulse (BVP) is extracted. Based on the discovered BVP, several other physiological parameters are calculated, such as HR, respiratory rate (RR), \( HRV_{LF} \) and \( HRV_{HF} \). The remote measurement of all physiological signals is in agreement with a physical device used as ground truth. Using these parameters, the authors construct a classifier, modeled with Naive Bayes and support vector machine (SVM), to predict whether the subject is under cognitive stress. The input features used for the models are mean heart rate, mean respiratory rate, normalized \( HRV_{LF} \) power, normalized \( HRV_{HF} \) power and \( HRV_{LF}/HF \) power ratio for each session. The results show that the prediction accuracy is 85% using SVM, demonstrating that the input signals are sensitive enough to measure the cognitive stress state. The HRV components and the RR are the strongest predictors, while the HR was not significantly different between the two detectable states. McDuff, Hernandez, et al. (2016) perform further investigations, but use different cognitive tasks (two cognitive demanding games). A person-independent machine learning model based on HR, HRV, \( HRV_{LF} \), \( HRV_{HF} \) (along with normalized and combined versions of these signals) and breathing rate is used to classify the stress level of the subjects. According to the authors, the average heart rate and breathing rate are not significantly different in any case, which differs from previous findings of other authors. The variations of \( HRV_{LF} \) and \( HRV_{HF} \), however, are significantly different during the cognitive tasks compared to the rest period; higher \( HRV_{LF} \) and lower \( HRV_{HF} \) power are found in both cognitive tasks compared to the rest period, which aligns with findings of previous work. The authors also point out that the stress predictions made by the model are consistent with the self-reported answers. The two participants with the highest self-reported stress level show the highest predicted stress level, while the two participants with the lowest self-reported stress level also present the lowest predictions.

Finally, Giannakakis et al. (2017) present an approach for the detection of stress/anxiety on the basis of eye blink, mouth/head movements and HR estimations using rPPG. The participants are recorded while performing a set of tasks designed to trigger emotional responses, such as talking in a foreign language, remembering a sad incident, visualizing images/videos and playing a gamified cognitive test, i.e. Stroop test. Facial cues are obtained from an automatically detected ROI. These cues are used to extract facial features,
which are related to eyes, mouth, and head movement. Figure 7.2 highlights the facial analysis used in the calculation of eye aperture, which is one of the features extracted and used in the emotion detection. The extracted features, including HR information estimated using rPPG, are selected and ranked accordingly to maximize a machine learning classification step. Different classifiers are employed, which yield different classification accuracy rates. For each task performed by the subjects, classification accuracy ranges between 80% and 90% taking into account the most efficient classifier. It is noted by the authors that the observation of stressful images and the interaction with the Stroop test appear to be the most consistent across the classifiers employed.
CHAPTER 8
EXPERIMENT 1: PROVOKED STRESS AND BOREDOM

The experiment described in this chapter, the first one conducted, aimed at gathering data and exploring the relations regarding facial actions (FA), HR and emotional states, particularly stress and boredom. The experiment is based on the previously mentioned findings that HR varies according to stress/frustration and that facial expressions can convey contextual information about an emotional state (Giannakakis et al., 2017). As opposed to previously mentioned works, in this experiment each subject spends an average of 25 minutes in the session, playing three different games that were custom-made to provoke similar emotional reactions to off-the-shelf games. Subjects were also not instructed regarding how they should move, therefore, body and facial reactions were likely to be the ones that the subject would normally exhibit in a gaming context. The approach consists of using induced boring to stressful mechanics in the games to produce variations in the emotional state and HR of participants.

In total, five studies were conducted on the data obtained from the experiment. The first study focuses on empirical exploration of how FA, defined as being any facial movement different from a neutral face, e.g. lips contraction, related to emotional states. The second study focuses on the variations of HR that occurred during the interaction with the games, especially during situations that were designed to provoke boredom and stress. This should confirm the hypothesis that the HR during boring and stressful parts of a game is in fact different. The third study focuses on the accuracy evaluation of an rPPG technique, applied to gaming sessions where subjects behave naturally. Finally, the fourth and fifth studies focus on automated facial analysis and remote detection of emotions, respectively.

The following sections present information regarding the participants, the experiment structure and each one of the mentioned studies.

8.1 PARTICIPANTS

Twenty adult participants of both genders (10 female) and different ages (22 to 59, mean 35.4, SD 10.79) as well as different gaming experience gave their informed and written consent to participate in the experiment. The study population consisted of staff members and students of the University of Skövde, as well as inhabitants of the community. When questioned about their levels of skill at playing video games, 1 subject (5%) reported no skill, 10 (50%) reported not very skilled, 7 (35%) reported moderately skilled and 2 (10%) reported very skilled. When asked the number of hours per week they had played any type of video game over the last year, 2 subjects (10%) reported more than 10 hours, 6 (30%) reported 5 to 10, 2 (10%) reported 3 to 4, 2 (10%) reported 1 to 3, 4 (20%) reported 0 to 1, and 4 (20%) reported no activity. These numbers indicate that the population has a diversity of gaming experience and playing frequency, which provides the experiment with information that is less skewed towards a specific profile of players,
8.2 MATERIALS AND PROCEDURES

The subjects were seated alone in a room in front a computer, while being recorded by a camera and measured by a heart rate sensor. The camera was attached to a tripod placed in front of the subjects at a distance of approximately 0.6m; the camera was tilted slightly up. A spotlight, tilted 45° up, placed at a distance of 1.6m from the subject and 45cm higher than the camera level, was used for illumination; no other light source was active during the experiment. Figure 8.1 illustrates the setup.

Figure 8.1: Experiment setup. On the left, an image illustrating the position of the equipment, including the angle of the external light source. On the right, an image highlighting the position and angle of the video camera.

Each participant was recorded for approximately 25 minutes, during which they played three games. Each game was followed by a questionnaire related to the game and stress/boredom. The first two games were followed by a 138 seconds’ rest period, during which the subjects listened to calm classical music. The third game was followed by an additional questionnaire about age and gaming experience/profile. The order in which the games were played was randomized among the subjects. Participants received instructions from a researcher informing them that they should play three games, answer a questionnaire after each game and rest. In addition, they were told that their gaming performance was not being analyzed, that they should not give up in the middle of the games and that they should remain seated during the whole process.

After each game, the subjects answered a questionnaire in order to provide self-reported stress and boredom measurements. The questionnaire consisted of six questions: of which the first four were a 5-point Likert scale regarding how the player felt about the stress/boredom level at the beginning/end of each game (1: not stressed/bored at all, 5: extremely stressed/bored); a question to identify the most enjoyable part of the game (very beginning, after beginning and before middle, middle, after middle and before end, very end); finally, a question regarding whether the subject understood the game. Before the conclusion of the experiment, the subjects responded to a final questionnaire with nine questions related to: age; gender; number of gaming hours per week over the last year, i.e. question from the video game experience questionnaire (Unsworth et al., 2015); self-assessed level of game proficiency or skill, i.e. question from the Survey of Spatial...
Representation and Activities - SSRA (Terlecki and Newcombe, 2005); familiarity with puzzle, platform and Tetris games; current state of mind compared to other days (e.g. normal, unusually stressed, etc.); and gaming profile (like, dislike challenging games).

Regarding the self-reported levels of stress and boredom provided by the subjects after each game, the answers were given according to the participant’s own interpretation of such levels. As a consequence it is not possible to treat the answers as a uniform scale with a defined degree of difference among the values. Therefore, it was decided to use a Wilcoxon Signed-rank test to statistically check whether the reported boredom levels at the end significantly differ from the ones at the beginning of the games, as well as whether the reported stress levels at the end differed from the ones at the beginning of the games. For the Mushroom game, the reported boredom levels at the beginning (median 3.5) were significantly higher than the reported levels at the end (median 1), \( Z = -2.69, p < 0.01 \). Regarding the reported stress levels, the values at the beginning (median 1) were significantly lower than the reported levels at the end (median 3), \( Z = 3.63, p < 0.01 \). For the Platformer game, boredom levels at the beginning (median 3) were higher than the ones at the end (median 1), \( Z = -2.47, p < 0.05 \). Regarding stress levels, the values at the beginning (median 1) were lower than the ones at the end (median 4), \( Z = 3.79, p < 0.01 \). Finally for the Tetris game, boredom levels at the beginning (median 4) were higher than the ones at the end (median 2), \( Z = -2.97, p < 0.01 \). Regarding stress levels, the values at the beginning (median 1) were lower than the ones at the end (median 4), \( Z = 3.95, p < 0.01 \).

The self-reported answers support the idea that the subjects perceived the three games as boring at the beginning and stressful at the end, which was the intended result of our design process.

8.3 DATA COLLECTION

During the whole experiment, the subjects were recorded using a Canon Legria HF R606 video camera. All the videos were recorded in color (24-bit RGB with three channels \( \times 8 \) bits/channel) at 50p frames per second (fps) with a pixel resolution of 1920 \( \times 1080 \) and saved in AVCHD-HD format, MPEG-4 AVC as the codec. At the same time, the subject’s HR was measured by a TomTom Runner Cardio watch (TomTom International BV, Amsterdam, Netherlands), which was used as ground truth. The watch was placed on the left arm, approximately 7cm from the wrist, like a regular wrist watch. The use of the watch was unobtrusive, and therefore did not affect the movements of the subjects who could still use both hands to play the games. The watch recorded the HR at 1 Hz.

8.4 GAMES AND STIMULI ELICITATION

The three games\(^1\) used in the experiment were 2D and casual-themed, and played with mouse or keyboard in a web browser. The games were carefully designed to provoke boredom at the beginning and stress at the end, with a linear progression between the two states (adjustments of such progression are performed every 1 minute). The game mechanics were chosen on the basis of the capacity to fulfill such linear progression, along with the quality of not allowing the player to instantly kill the main character (by mistake or not), e.g. by falling into a hole. The mechanics were also designed/selected

\(^1\)Source code available at: https://fernandobevilacqua.com/link/phd-experiment1
UNIVERSITY OF SKÖVDE

to ensure that all the subjects would have the same game pace, e.g. a player could not deliberately control the game speed on the basis of his/her will or skill level.

The Mushroom game, illustrated in Figure 8.2 (left), is a puzzle where the player must feed a character by dragging and dropping mushrooms in rounds. In a given round, \( M \) mushrooms are displayed in a grid and the player has \( K \) seconds (a decreasing time bar at the top informs the remaining time) to collect good and discard bad (poisonous) mushrooms. At the upper-right corner of the screen, a sign informs the player about the bad/poisonous mushroom of the round. The player must drag and drop all the good mushrooms (the ones different from the poisonous indication) into the character, while dragging and dropping the bad ones into the trash can. The mushrooms are differentiated by the colors of their features (circles). The player is rewarded with score points, a health bar increase (\( HB_I \)) and a pleasant sound when a correct move is performed. If a mistake is made, a health bar decrease (\( HB_D \)) and an annoying/aggressive alarm sound are applied. If the time \( K \) is over and the player has not finished moving all the mushrooms of the round, each remaining mushroom in the grid is counted as a mistake. If the grid is clean and there is still time available, the player must wait until the time is over. The values of \( M \), \( K \), \( HB_I \) and \( HB_D \) are used to induce boredom/stress. At the beginning, \( M \) is low (starts with 2) and \( K \) is high (starts with 45 seconds), so the player spends a significant amount of time waiting for the game to continue; every 1 minute the value of \( M \) is increased and \( K \) is decreased. The changes continue until the player is unable to deal with the amount of mushrooms within the available time. This leads to mistakes that will eventually decrease the health bar to zero, terminating the game. After the mark of 6 minutes, the game becomes virtually impossible to beat.

The Platformer, illustrated in Figure 8.2 (center), is a side-scrolling, endless runner game where the player must control the main character while collecting hearts and avoiding obstacles (skulls with spikes). The character can jump (by pressing the up arrow key in the keyboard) or slash (S key), however, the player cannot move the main character left or right; it remains in the same position on the screen (towards the left side of the screen). The character moves on top of platforms, which are always perfectly connected, so there are no gaps (holes) among them; however, the height of the platform can vary, so there might be a slope up/down connecting two platforms, for instance. If the character hits an obstacle, the health bar is decreased (\( HB_D \)) and a sound effect related to pain is heard. If any heart is collected, the health bar increases (\( HB_I \)) and a pleasant sound effect is heard. The position where the hearts appear on each platform is adjustable (defined by \( HH \)), so they can appear close to the platform (no action is required to collect the heart) or somewhat higher from the ground (jump action is required to collect the heart). The speed of the character (\( S \), which is the velocity of the elements moving on the screen), the height variation of each new platform that appears on the screen (\( HV \)), the amount of hearts (\( G \)) and obstacles (\( E \)) per platform are all controlled by the game and used to adjust boredom/stress. At the beginning, boredom is induced by keeping all the previously mentioned parameters with low values, which means the game is slow, the character moves from platform to platform at the same height and almost no hearts or obstacles appear on the screen. The few hearts that are available are placed close to the ground to destimulate jumping actions. As time progresses, the values of \( S \), \( E \), \( HV \), \( HB_D \) and \( HH \) increase, while \( G \) and \( HB_I \) decrease to induce a stressful state in the player. At the mark of 5 minutes, for instance, the game is significantly fast, with several obstacles on the screen and almost no hearts to collect; the damage caused to the character when hit by an obstacle is also greater than at the beginning of the game. The linear increase in difficulty will eventually result in consecutive hits (mistakes), which will decrease the health points to zero, when the game ends.
Finally, the game Tetris, shown in Figure 8.2 (right), is a modification of the original Tetris game. In this version of the game, the next block to be added to the screen is not displayed, so the player is unable to predict future moves. Additionally, the down key, usually used to speed up the descendant trajectory of the current piece, is disabled. The keyboard controls comprise the arrow keys to move the piece left/right and the R key to rotate the piece. The game was also modified to ensure that all the subjects received the same sequence of pieces (the same seed for the generation of random numbers was used). The speed of the pieces’ fall ($S$) is used to control boredom and stress; at the beginning of the game, boredom is induced by using a low value for $S$, which makes the game slow since the pieces fall slowly and the player is unable to speed them up. As time progresses, $S$ increases linearly, making the game faster and harder to play, which should induce stress. At the mark of 5 minutes, for instance, a single piece takes almost 1 second to traverse the whole screen.

### 8.5 STUDY 1: VARIATIONS OF FACIAL ACTIONS

This study presents information regarding FA that the subjects presented during the experiment. The 6 hours of recordings of all the subjects were analyzed manually and FA were annotated empirically. The annotations were categorized according when they occurred (the boring first part or the stressful second part of the games). An analysis of these annotated FA was conducted at group and individual level, which aimed to find patterns between the featured FA and the boring/stressful periods of the games.

The following sections presents the analysis, discussion and results of the gathered information.

#### 8.5.1 ANALYSIS AND METHODS

The recordings of all the subjects were analyzed by a single researcher who took notes of any facial actions (FA) that were different from a neutral (resting) face, e.g. lips contraction, brow movement, etc. The annotations were not conducted periodically, e.g. every 5 seconds, instead they were made only when the subject’s face changed from its neutral/resting state. As a consequence, if the subject’s face remained in a neutral state for
a long period of time, no annotations were made during that period.

This empirical and non-standard approach for facial annotation was used because the focus is not on facial expressions per se, but in the exploration of any facial action (standardized or not) that might be used to infer patterns in boredom/stressful states. This approach is not without its limitations, however it provides a reasonable empirical perception of facial activity that is different from a neutral face, which is satisfactory for the investigation. Since FA are subtle and not necessarily part of a complete facial expression, e.g. surprise face, they might be better identified in a context where annotations are made only when facial changes happen, as opposed to a frame-by-frame analysis/annotation of a video, for instance.

![Annotated facial actions (FA).](image)

According to the design of the games, the subjects were supposed to experience the beginning of the games as more boring than the end, while the experience at the end was to be perceived as more stressful than the beginning of the games. As a result, if the game sessions of each subject are divided in half, in theory, one of the two resulting parts is more likely to be perceived as more boring by the subjects, while the other is more likely to be perceived as more stressful. Using that assumption, the FA annotations were divided into two groups, the ones made during the period that corresponds to the first half ($H_0$) of the games and the ones made in the second half ($H_1$). This division of the annotations aimed to identify any pattern regarding FA that occurred during periods theoretically perceived as boring or stressful. After all the annotations were made, an identification of uniqueness was performed and, based on that information, the repetitions of such unique actions across the games for all the subjects were counted. As a result, the frequency of each FA during all the game sessions was obtained, as well as when they occurred (in $H_0$ or $H_1$). Any FA that appeared just once during the entire 6 hours recorded was excluded from the list, under the assumption that such action was noise or probably part of another action. As a result, 17 unique FA that appeared in the recordings at least twice were identified. Excluding the talking and laughing FA, Figure 8.3 illustrates all the annotated FA. Finally, after all the annotations were counted
and categorized according to the period in the game, a per-subject evaluation of the frequency of FA was conducted. For each subject, an inspection was performed regarding FA that appeared with a higher frequency in $H_1$ of all three games than in $H_0$, and vice versa (appeared with a higher frequency in $H_0$ of all three games than in $H_1$).

8.5.2 RESULTS

Table 8.1: Number of FA annotations made for all subjects during periods $H_0$ and $H_1$ of the games

<table>
<thead>
<tr>
<th>Game</th>
<th>Period $H_0$</th>
<th>Period $H_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mushroom</td>
<td>90</td>
<td>98</td>
</tr>
<tr>
<td>Platformer</td>
<td>88</td>
<td>181</td>
</tr>
<tr>
<td>Tetris</td>
<td>110</td>
<td>159</td>
</tr>
</tbody>
</table>

The number of subjects that featured a particular FA was analyzed, together with the number of repetitions of such FA, for all three games. The analysis also took into account the period of the game. In addition, only FA featured by two or more subjects were considered, since this analyzed more frequent FA among the whole group of subjects instead of the peculiarities of a single person. Table 8.1 presents the number of FA annotations made for all the subjects during the games. According to the results, the number of FA annotations made during $H_1$ (second half) of all three games was greater than the number of annotations made during $H_0$ (first half). The increase in annotations during $H_1$ compared to $H_0$ was 8.8%, 105.6% and 44.5% higher for the Mushroom, Platformer and Tetris games, respectively.

The FA annotated during each game are the following: for the Mushroom game, the three most frequent FA in $H_0$ are frown (repeated 16 times among 5 subjects), talking (12 times, 3 subjects) and tongue touching lips (9 times, 3 subjects). The three most frequent FA in $H_1$ are frown (repeated 16 times among 3 subjects), talking (13 times, 5 subjects) and lips parted (13 times, 5 subjects). A comparison of the most frequent FA in the two periods reveals that while both frown and talking are present, they are not featured by a significant number of participants. In fact, no more than 5 subjects (25% of the participants) featured one of these FA. This suggests that individuals present distinct facial behaviors that are not easily generalizable, even in the same context. Curiously, two particular FA presented a significant change in the number of repetitions and subjects between the two periods: lip pressors (from 7 to 11 repetitions, 2 to 4 subjects) and lips parted (from 5 to 13 repetitions, 2 to 5 subjects). When compared to the whole group of participants, such an increase is not significant (again they represent less than 25% of the participants), but it might be the indication of a pattern for two or three subjects. As suggested by previous work, the combination of such particular changes with another physiological signal, e.g. HR, might produce an acceptable detector for boredom/stress emotional states.

For the Platformer game, the three most frequent FA for $H_0$ are frown (19 repetitions among 3 subjects), tongue touching lips (12 repetitions, 3 subjects) and smile not showing teeth (11 repetitions, 3 subjects). For $H_1$, the FA are frown (49 repetitions, 5 subjects), smile not showing teeth (21 repetitions, 7 subjects) and lips parted (17 repetitions, 5 subjects). A comparison of the FA in both periods reveals that frown is featured by more subjects (5, representing 25%) during the stressful part of the game, but more par-
participants (7, representing 35%) also feature smiles not showing teeth as well. In addition to these FA, 25% of the participants feature talking behavior during $H_1$, externalizing game decisions.

For the Tetris game, the three most frequent FA for $H_0$ are frown (36 repetitions among 4 subjects), smile not showing teeth (14 repetitions, 4 subjects) and lip pressor (11 repetitions, 4 subjects). For $H_1$, the FA are frown (42 repetitions among 4 subjects), lip pressor (28 repetitions, 6 subjects) and smile not showing teeth (16 repetitions, 5 subjects). Comparing these results to the most frequent FA in the Mushroom game reveals that only frown is present in both. It is important to stress that frown is featured by less than 25% of the participants in both games, which highlights the difficulties in finding a pattern that can be applied to all subjects, in order to identify a boring or stressful situation, even when the most frequent FA are used. On the other hand, two FA present a significant change from one period to another in the Tetris game: lip pressor (from 11 to 28 repetitions, 4 to 6 subjects) and talking (from 0 to 15 repetitions, 0 to 6 subjects). Both facial actions are featured by 30% of the participants, which could be further investigated in the pursuit of FA that can help in the identification of emotional states. Regarding the talking FA, it has been observed in the recordings that some subjects tended to externalize in words any wrong decisions they made in the game, such as how pieces were positioned. This is similar to observations made during the Platformer game; in that sense, talking could be used as an indicator of activity in the game, since it is a clear facial manifestation that happened, in this case, when players were frustrated.

For further FA analysis based on a group level, see Bevilacqua, Backlund, and Engström (2016).

Finally, a per-subject inspection of all annotated FA was conducted according to the procedure described in Section 8.5.1. The aim was to identify, for each subject, which FA appeared in $H_0$ (or $H_1$) of all three games with a higher frequency than they did in $H_1$ (or $H_0$), if any. Table 8.2 shows the results of this inspection. The marked numbers represent the frequency of a FA that was present in all three games for the specified subject and period. In total, 10 participants (50%) featured at least one FA that appeared in all three games, in the same period (boring or stressful part), with a frequency equal to or greater than its appearance in the counter-period. Subject 2, for instance, featured one lip pressor during $H_0$, while the same FA appeared a total of 18 times in $H_1$ for all three games combined. It is important to highlight that subject 16 was the only one who featured a FA more frequently in $H_0$ of all three games than he/she did during $H_1$; all the other subjects featured FA more frequently in $H_1$ than in $H_0$.

8.5.3 DISCUSSION

With regard to the FA, even though further investigation is required, calculations indicate that the subjects featured a neutral face for a longer period of time during the first half ($H_0$) of all the games compared to the second half ($H_1$). Since FA annotations were made only when the subject’s face featured anything different than the neutral face, more annotations indicate more facial activity. Additionally the results might indicate that the subjects featured more FA (different from the neutral face) during stressful situations than they did under boring situations, where a neutral face/expression is probably dominant.

The games used in the experiment were designed to gradually increase in the level of difficulty until the subject could no longer proceed. As a consequence, it is possible to postulate that the smiles and laughs during the second half could be connected to the subject’s perception that the games were too difficult to continue playing properly. On
Table 8.2: Subject-based frequency of FA that appeared in the same period of all three games

<table>
<thead>
<tr>
<th>Subject</th>
<th>FA</th>
<th>Period $H_0$</th>
<th>Period $H_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Lip pressor</td>
<td>1</td>
<td>18&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>15</td>
<td>Lip pressor</td>
<td>2</td>
<td>9&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>10</td>
<td>Laughing</td>
<td>2</td>
<td>19&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>14</td>
<td>Laughing</td>
<td>3</td>
<td>9&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>12</td>
<td>Smile not showing teeth</td>
<td>2</td>
<td>8&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>13</td>
<td>Smile not showing teeth</td>
<td>0</td>
<td>6&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>18</td>
<td>Smile not showing teeth</td>
<td>4</td>
<td>10&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>11</td>
<td>Lips parted</td>
<td>1</td>
<td>10&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>17</td>
<td>Lip stretcher</td>
<td>0</td>
<td>8&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>16</td>
<td>Talking</td>
<td></td>
<td>1&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>b</sup> FA was present in all three games for the specified subject and period.

The other hand, they could indicate genuine manifestations of enjoyment during the moments the subjects felt the game was properly balanced and engaging. Regarding other FA, such as lip pressor and lips parted, further investigation is required to accurately relate or use these actions to predict/detect emotional states. However, the results reveal a clue about how FA variations can be different at the individual level. As previously discussed, the analysis and generalization of FA at a group level is less clear than an individual approach, since FA behavior might be specific to each person. The per-subject analysis indicates that, for some of the participants, at least one FA was featured in the three games, in the same period, by the same person. Such information could be used as the starting point for further research on FA and, for example, an individual-tailored detection model for boredom/stress.

8.5.4 CONCLUSION

The results show that more FA annotations were made during the stressful part of the games, which indicates that the participants maintained a neutral face for longer periods of time during the boring part. The analysis at group level reveals that any FA pattern is related to 5 subjects (25% of the group) at most. In the analysis conducted at the individual level, particular patterns were found for 10 subjects (50% of the group).

8.6 STUDY 2: VARIATIONS OF HEART RATE

This study presents information regarding the variations of HR of subjects during the experiment. The HR data related to the game session of each subject were divided into periods of 1 minute each; the HR mean was calculated and compared to a baseline value obtained from the HR mean of the subject during rest. Based on the self-reported answers regarding stress and boredom, the HR mean was analyzed at specific periods, such as the second minute of gameplay (perceived as boring) and the last minute of gameplay (perceived as stressful).
The result of this study contributes information regarding HR in the context of games. It is a key element to creating user-tailored models for emotion detection based on different data sources, e.g. HR and facial actions. The following sections present the analysis, discussion and results of the study.

8.6.1 ANALYSIS AND METHODS

Firstly, the set of HR was filtered by removing all the readings obtained during the experiment whose values were equal to zero, assuming they were miss-readings. Thereafter, the baseline HR value for each subject ($B_s$) was calculated as:

$$B_s = \frac{1}{2}(\overline{HR}_{r1,s} + \overline{HR}_{r2,s})$$ (8.1)

where $s$ indicates the subject and $\overline{HR}_{r1,s}$, $\overline{HR}_{r2,s}$ are the mean HR during the first and second rest period (for subject $s$), respectively. $B_s$ is assumed to be the "expected" HR of a given subject while at rest. The average difference between $\overline{HR}_{r1,s}$ and $\overline{HR}_{r2,s}$ for each subject was 2.34 bpm.

The HR mean coefficient $C_{g;t}^s$ was then calculated; this is the HR mean of a subject while playing a game during a given period of 60 seconds:

$$C_{g;t}^s = \frac{1}{60} \sum_{n=1}^{60} HR_{s,g}(t \cdot 60 + n)$$ (8.2)

where $s$ is the subject, $g$ is the game being played ($M$ for Mushroom, $P$ for Platformer or $T$ for Tetris), $t$ is the period and $HR_{s,g}(k)$ is the HR of subject $s$ measured in game $g$ at the mark of $k$ seconds. Since each subject played each game for more than 60 seconds, there is more than one period for each subject and a given game. The $t$ component of $C_{g;t}^s$ specifies which of these periods the HR mean refers to. For instance, $t = 0$ comprises the period from time 0:00 until time 1:00 of a given game, $t = 1$ is the period from time 1:01 until 2:00, and so on. As an example, the HR mean coefficient $C_{P;1}^2$ is the HR mean of subject 2 playing the Platformer game from time 1:01 to 2:00.

HR values are specific to each individual, so the relativized HR mean coefficient, $V_{g;t}^s$, was calculated by subtracting $B_s$ from $C_{g;t}^s$ as:

$$V_{g;t}^s = C_{g;t}^s - B_s$$ (8.3)

$V_{g;t}^s$ accounts for values that are related to changes instead of absolute HR measurements, which are significantly more suitable for comparison among different subjects, or within the same subject.

Based on previous work regarding variations of HR and emotions, the following hypothesis was proposed: the HR mean during the last minute of gameplay is greater than the HR mean during the second minute of gameplay. More specifically, the true difference in means between $V_{g;n}^0$ (i.e. HR means when $t = n$, where $n$ is the last minute of gameplay) and $V_{g;1}^1$ (i.e. HR means when $t = 1$, the second minute of gameplay) is greater than zero. The dependent variable is $V$ and the null hypothesis is that the true difference in means between $V_{g;n}^0$ and $V_{g;1}^1$ is less than or equal to zero. The reason $t = 1$ (second minute of gameplay) was chosen instead of $t = 0$ (first minute of gameplay) for the hypothesis is...
due to the belief that the first minute of the game might not be ideal for a fair comparison. Firstly, during the first minute of gameplay, subjects are less likely to be in their usual neutral emotional state. They are more likely to be stimulated by the excitement of the initial contact with a game soon to be played, which interferes with any feelings of boredom. Secondly, subjects need a basic understanding of and experimentation with the game, in order to assess whether it is boring or not. As per the understanding of the author, it is less likely for such conjecture to be fulfilled during the first minute of gameplay than it is during the second minute of gameplay.

8.6.2 RESULTS

Table 8.3 presents the values of $V$, the relativized HR mean coefficient, for all the subjects in all the games, grouped by intervals of 1 minute, calculated according to the description in Section 8.6.1. Column $g$ is the game played, $t$ is the period in the game and $s$ is the subject. Since all the games were constantly changing in level of difficulty and the subjects have different gaming skills, there are subjects with no data entry for some $t$ intervals, which means the subject was defeated by the game sooner than other subjects were. Subject 9 had problems playing the Platformer game, therefore, data for that subject in that game were not used in the calculations.

A positive value in Table 8.3 represents a $V$ (HR mean) that is above the subject’s baseline $B_s$ (mean HR while at rest) for a specific period $t$. A negative value indicates that $V$ in that period is below the subject’s baseline $B_s$. Assuming $n$ is the last minute of gameplay of a given subject in a game, by comparing the values at $t = 0$ (first minute of gameplay, perceived as boring) and $t = n$ (last minute of gameplay, perceived as stressful) in the Mushroom game, 19 subjects (95%) presented $V^{M;n}$ greater than $V^{M;0}$. The same comparison regarding the Platformer game indicates that 16 subjects (84.2%) had higher $V^{P;n}$ than $V^{P;0}$. In the Tetris game 13 subjects (65%) presented higher $V^{T;n}$ than $V^{T;0}$.
Table 8.3: Values of $V_{g,s}$, the relativized HR mean coefficient, for all subjects ($s$) in a given game $g$ (M is for Mushroom, P for Platformer and T for Tetris), grouped by intervals ($t$) of 1 minute

<table>
<thead>
<tr>
<th>$g$</th>
<th>$t$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>0</td>
<td>-3.8</td>
<td>2.2</td>
<td>-2.5</td>
<td>-3.1</td>
<td>-3.4</td>
<td>-2.5</td>
<td>0.4</td>
<td>3.5</td>
<td>-4.9</td>
<td>-2.8</td>
<td>-3.4</td>
<td>-0.2</td>
<td>-1.8</td>
<td>5.9</td>
<td>-4.9</td>
<td>6.5</td>
<td>-0.4</td>
<td>-3.3</td>
<td>4.4</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-9.1</td>
<td>-1.4</td>
<td>0.3</td>
<td>-2.6</td>
<td>0.2</td>
<td>-3.7</td>
<td>-1.5</td>
<td>2.7</td>
<td>-4.9</td>
<td>-4.1</td>
<td>-10.3</td>
<td>0.0</td>
<td>-3.1</td>
<td>-3.1</td>
<td>0.2</td>
<td>5.6</td>
<td>2.5</td>
<td>-0.8</td>
<td>0.5</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-4.8</td>
<td>-1.3</td>
<td>-0.1</td>
<td>-0.6</td>
<td>7.0</td>
<td>0.9</td>
<td>2.0</td>
<td>4.5</td>
<td>-0.8</td>
<td>-3.0</td>
<td>-9.2</td>
<td>4.1</td>
<td>0.1</td>
<td>-0.5</td>
<td>-0.1</td>
<td>5.4</td>
<td>2.8</td>
<td>2.4</td>
<td>2.6</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-4.9</td>
<td>-0.7</td>
<td>-2.8</td>
<td>-1.5</td>
<td>1.5</td>
<td>0.3</td>
<td>2.4</td>
<td>5.1</td>
<td>-2.4</td>
<td>1.7</td>
<td>-4.6</td>
<td>2.4</td>
<td>0.4</td>
<td>-0.2</td>
<td>0.7</td>
<td>4.5</td>
<td>3.4</td>
<td>3.8</td>
<td>2.4</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>-3.9</td>
<td>-1.1</td>
<td>0.9</td>
<td>1.5</td>
<td>5.3</td>
<td>0.8</td>
<td>4.5</td>
<td>6.3</td>
<td>2.0</td>
<td>1.3</td>
<td>-3.6</td>
<td>3.3</td>
<td>1.6</td>
<td>6.9</td>
<td>0.0</td>
<td>4.5</td>
<td>2.9</td>
<td>3.4</td>
<td>9.9</td>
<td>9.1</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.3</td>
<td>2.4</td>
<td>1.4</td>
<td>1.7</td>
<td>11.9</td>
<td>-1.2</td>
<td>-</td>
<td>-</td>
<td>4.4</td>
<td>10.2</td>
<td>1.6</td>
<td>6.2</td>
<td>-</td>
<td>3.2</td>
<td>5.9</td>
<td>3.2</td>
<td>17.9</td>
<td>5.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>-</td>
<td>-</td>
<td>-1.2</td>
<td>-0.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3.3</td>
<td>-</td>
<td>6.7</td>
</tr>
<tr>
<td>P</td>
<td>0</td>
<td>-1.7</td>
<td>1.3</td>
<td>0.4</td>
<td>-0.2</td>
<td>0.1</td>
<td>9.9</td>
<td>2.4</td>
<td>-1.7</td>
<td>-</td>
<td>19.7</td>
<td>6.0</td>
<td>0.2</td>
<td>5.9</td>
<td>1.2</td>
<td>4.2</td>
<td>5.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-1.6</td>
<td>-0.4</td>
<td>3.4</td>
<td>-0.9</td>
<td>-0.3</td>
<td>2.2</td>
<td>2.4</td>
<td>5.5</td>
<td>-</td>
<td>4.4</td>
<td>0.7</td>
<td>-1.5</td>
<td>4.1</td>
<td>4.9</td>
<td>-1.0</td>
<td>0.8</td>
<td>-0.2</td>
<td>3.7</td>
<td>7.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.9</td>
<td>9.7</td>
<td>0.8</td>
<td>-0.6</td>
<td>3.0</td>
<td>-1.1</td>
<td>1.3</td>
<td>3.9</td>
<td>-</td>
<td>15.1</td>
<td>0.2</td>
<td>3.5</td>
<td>3.9</td>
<td>3.8</td>
<td>11.7</td>
<td>-0.7</td>
<td>2.8</td>
<td>0.4</td>
<td>4.0</td>
<td>10.6</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3.0</td>
<td>9.3</td>
<td>2.5</td>
<td>-2.6</td>
<td>2.8</td>
<td>10.3</td>
<td>4.9</td>
<td>5.2</td>
<td>-</td>
<td>21.6</td>
<td>3.2</td>
<td>5.4</td>
<td>9.2</td>
<td>4.6</td>
<td>9.9</td>
<td>-</td>
<td>2.1</td>
<td>2.6</td>
<td>7.9</td>
<td>10.4</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5.9</td>
<td>6.8</td>
<td>-</td>
<td>8.0</td>
<td>5.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4.9</td>
<td>-13.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>2.1</td>
<td>6.5</td>
<td>-2.1</td>
<td>-1.3</td>
<td>-4.0</td>
<td>5.7</td>
<td>3.4</td>
<td>4.2</td>
<td>8.3</td>
<td>3.2</td>
<td>2.1</td>
<td>-0.1</td>
<td>3.5</td>
<td>3.4</td>
<td>4.4</td>
<td>-1.2</td>
<td>7.8</td>
<td>-3.9</td>
<td>5.8</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-2.7</td>
<td>0.0</td>
<td>-3.3</td>
<td>-1.2</td>
<td>-4.9</td>
<td>0.1</td>
<td>4.3</td>
<td>2.7</td>
<td>2.9</td>
<td>0.0</td>
<td>2.6</td>
<td>2.2</td>
<td>-2.5</td>
<td>5.9</td>
<td>-1.3</td>
<td>4.2</td>
<td>-0.4</td>
<td>5.7</td>
<td>0.0</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-1.7</td>
<td>2.6</td>
<td>2.4</td>
<td>-0.1</td>
<td>-2.3</td>
<td>4.3</td>
<td>3.5</td>
<td>-0.4</td>
<td>2.7</td>
<td>5.1</td>
<td>2.6</td>
<td>5.9</td>
<td>1.1</td>
<td>-1.1</td>
<td>5.3</td>
<td>-1.8</td>
<td>7.4</td>
<td>0.1</td>
<td>8.1</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-1.9</td>
<td>-0.2</td>
<td>0.3</td>
<td>-2.2</td>
<td>0.8</td>
<td>5.4</td>
<td>2.1</td>
<td>3.8</td>
<td>-</td>
<td>5.2</td>
<td>2.2</td>
<td>5.4</td>
<td>-0.5</td>
<td>-2.5</td>
<td>4.7</td>
<td>-1.2</td>
<td>10.6</td>
<td>1.5</td>
<td>3.8</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>-0.8</td>
<td>3.0</td>
<td>-</td>
<td>0.9</td>
<td>-</td>
<td>-</td>
<td>7.8</td>
<td>-</td>
<td>9.2</td>
<td>0.2</td>
<td>6.6</td>
<td>-</td>
<td>3.4</td>
<td>5.6</td>
<td>-1.2</td>
<td>-</td>
<td>2.0</td>
<td>6.8</td>
<td>4.3</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1.5</td>
<td>7.4</td>
<td>-</td>
<td>-0.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>12.9</td>
<td>7.4</td>
<td>4.5</td>
<td>-</td>
<td>3.5</td>
<td>6.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>6.9</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

* Subject 9 had problems playing the Platformer game, thus data from this subject during this game were excluded.
Table 8.4: Mean of the differences of \( V_{g,t} \) at the periods \( t = 1 \) (second minute of gameplay) and \( t = n \) (last minute of gameplay), for all subjects in each game \( (g) \). Values in bpm (beats per minute). Significance was tested with a one-tailed paired t-test

<table>
<thead>
<tr>
<th>Game ((g))</th>
<th>( V_{g,n}, V_{g,1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mushroom (M)</td>
<td>6.11 ***</td>
</tr>
<tr>
<td>Platformer (P)</td>
<td>5.10 ***</td>
</tr>
<tr>
<td>Tetris (T)</td>
<td>3.33 ***</td>
</tr>
</tbody>
</table>

*** \( p < 0.001 \)

Table 8.5: Mean of the differences of \( V_{g,t} \) at key periods, for all subjects in a given game \( g \). Values in bpm (beats per minute)

<table>
<thead>
<tr>
<th>Game ((g))</th>
<th>( V_{g,1}, V_{g,0} )</th>
<th>( V_{g,n}, V_{g,n-1} )</th>
<th>( V_{g,n}, V_{g,0} )</th>
<th>( V_{g,n-1}, V_{g,1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mushroom (M)</td>
<td>-0.87</td>
<td>2.39</td>
<td>5.23</td>
<td>3.71</td>
</tr>
<tr>
<td>Platformer (P)</td>
<td>-1.31</td>
<td>2.57</td>
<td>3.78</td>
<td>2.52</td>
</tr>
<tr>
<td>Tetris (T)</td>
<td>-1.71</td>
<td>1.22</td>
<td>1.62</td>
<td>2.10</td>
</tr>
</tbody>
</table>

As previously mentioned, the null hypothesis is that the true difference in means between \( V \) at the last minute of gameplay \( (t = n) \) and at the second minute of gameplay \( (t = 1) \) is less than or equal to zero. Table 8.4 shows the mean of the differences of a one-tailed paired t-test on the values of \( V_{g,n} \), i.e. last minute of gameplay for a given game \( g \), and \( V_{g,1} \), i.e. second minute of gameplay for a given game \( g \), for all games and subjects. The results indicate that the difference is greater than zero with statistical significance for all games. For the Mushroom game, the mean of the differences between the last \( (V_{M,n}) \) and the second \( (V_{M,1}) \) minutes of gameplay is 6.11 bpm \((p < 0.001)\). For the Platformer game, the mean of the differences of \( V_{P,n} \) and \( V_{P,1} \) is 5.1 bpm \((p < 0.001)\). Finally, for the Tetris game, the mean of the differences of \( V_{T,n} \) and \( V_{T,1} \) is 3.33 bpm \((p < 0.001)\). These numbers reject the null hypothesis, thus supporting the experimental hypothesis that the HR mean during the last minute of gameplay is greater than the HR mean during the second minute of gameplay, for all games.

In order to further explore the mean variation of HR at key periods other than the ones in our hypothesis, the mean of the differences involving \( V_{g,0}, V_{g,1}, V_{g,n} \) and \( V_{g,n-1} \) was calculated for all games and subjects. The results are presented in Table 8.5. \( V_{g,0} \) and \( V_{g,1} \) are the values of \( V \) for a given game \( g \) during the first and the second minute of gameplay, respectively. \( V_{g,n} \) and \( V_{g,n-1} \) represent the values of \( V \) for a given game \( g \) during the last and immediately before the last minute of gameplay, respectively. As previously mentioned, the value of \( n \), the last minute of gameplay, is different for each subject since they might have been defeated by the game at different moments due to personal skill levels.

In the first two minutes of gameplay \((t = 0 \text{ and } t = 1)\), the mean of the differences between \( V_{g,1} \) and \( V_{g,0} \) is negative for all games. These numbers suggest a higher HR mean during the first minute of the games \((t = 0)\) than during the second minute \((t = 1)\). At the last two minutes of gameplay \((t = n \text{ and } t = n - 1)\), the mean of the differences between \( V_{g,n} \) and \( V_{g,n-1} \) is positive for all games. These numbers suggest a higher HR mean during the
last minute of the game \((t = n)\) compared to the penultimate minute \((t = n − 1)\). Finally, regarding the last \((t = n)\) and first \((t = 0)\) minutes of gameplay, as well as the penultimate \((t = n − 1)\) and second \((t = 1)\) minutes of gameplay, the numbers suggest a higher HR mean during the last minute of the game \((t = n)\), compared to the first minute \((t = 0)\). There is also a higher HR mean during the penultimate minute of gameplay \((t = n − 1)\), compared to the second minute \((t = 1)\).

### 8.6.3 DISCUSSION

A number of subjects presented a higher value for \(V\), the relativized HR mean coefficient, towards the end of the Mushroom and the Platformer games, compared to the same period of the Tetris game, as shown in Table 8.3. Both the Mushroom and the Platformer game were completely new to the subjects, since they were developed exclusively for the experiment. For the self-reported 5-point Likert scale regarding familiarity with the games/genres, the mean value was 2.75 for the Mushroom, 2.8 for the Platformer and 3.35 for the Tetris game (5 denoting extremely familiar). Such numbers could indicate that it would be less likely for the subjects to predict what was going to happen in the Mushroom and the Platformer games, compared to the Tetris game. It could explain the greater number of subjects with higher \(V\) during the end (stressful) part of those two games, compared to the smaller number of subjects with higher \(V\) at the end of Tetris. The latter is a popular game and the subjects were more familiar with it; therefore, it is more likely they would guess what is about to happen in the game, reducing anxiety levels. This is especially true if the subject is trained to deal with the inherent stress of the mechanic, for instance.

A significant number of subjects presented a negative value for \(V\) in some periods. In total 16 subjects (80%) in the Mushroom game, 11 (57.8%) in the Platformer and 12 (60%) in the Tetris game presented negative values. A negative value indicates that the subjects had a lower HR mean while playing the game at specific periods than while resting. After the experiment, some subjects reported discomfort during the rest period, mentioning that it was too long and boring. The rest period might have been stressful for some subjects, as they were required to rest while seated without any entertainment, e.g. mobile phones. Another explanation for the negative values is that the calculation of the subject’s baseline \(B_s\) might be a weak approximation of the real HR mean of each subject during rest, since only two 140-seconds long rest periods for each subject were measured. However, the baseline calculation is probably still a good parameter, since the average difference between the mean HR of the two rest periods was significantly low, as explained in Section 8.6.1.

Regarding the confirmation of the hypothesis, the mean of the differences between \(V\) in the last \((t = n)\) and the second \((t = 1)\) minutes of gameplay, presented in Table 8.4, shows statistical significance in the difference for all games. It reinforces findings of previously mentioned works (Vandeput et al., 2009; Garde et al., 2002; Boussofsaf, Maaoui, and Pruski, 2013b; Rodriguez et al., 2015; Yamakoshi et al., 2007) which indicate that HR tends to be higher (above the subject’s baseline) during stressful moments and lower (closer to subject’s baseline) during boring moments in a gaming context. As previously described, the reason the \(t = 1\) (second minute of gameplay) was used instead of \(t = 0\) (first minute of gameplay) for the main comparison is because the first minute of all games might not be ideal for a fair comparison. During the first minute, subject are less likely to be in their usual neutral emotional state. This line of reasoning is supported by the exploratory analysis of the mean of the differences of \(V\) at periods other than the ones used in the hypothesis, as presented in Table 8.5. At the beginning of the games, the HR
mean during $t = 0$ (0:00 to 1:00) was higher than during $t = 1$ (1:01 to 2:00) for all three games. It could indicate that the subjects were more stimulated at the very beginning, probably due to the excitement of the initial contact with a game to be played. Such a difference between $t = 1$ and $t = 0$ could also be explained by the fact that the subjects probably understood the game mechanic. During the first minute of gameplay, subjects are probably still working to understand the game, so an opinion regarding boredom is still being formed. After the one minute mark, subjects are more likely to fully understand the game, so they could assess that it was boring. Additionally, a better understanding of the mechanic combined with the fact that the subjects were not allowed to change the game pace, e.g. to make it more interesting, probably increased the feeling of boredom.

8.6.4 CONCLUSION

The results indicate that the average HR mean for all the subjects during the last minute of gameplay was greater than the average HR mean during the second minute of gameplay, for all the games with statistical significance, i.e. $p < 0.001$. The findings are aligned with and reinforce previous research that indicate higher HR mean during stressful situations in a gaming context. The design of the games permitted a more elaborated analysis of boring and stressful periods, which contributes information regarding variations of HR mean during such conditions in gaming sessions. Additionally an exploratory investigation regarding HR mean during other key periods in the games was conducted, e.g. first and penultimate minutes of gameplay. Although further analysis is still required, the numbers suggest that the average HR mean during the first minute of gameplay was greater than during the second minute of gameplay, probably as a consequence of unusual excitement during the first minute, e.g. the prospect of playing a new game. The findings suggest that changes in the HR during gaming sessions are a promising indicator of stress, which could be incorporated into a model aimed at emotion detection. As indicated by previous work, a user-tailored model based on several signals, e.g. HR and FA, is more likely to detect the emotional states of users.

8.7 STUDY 3: HEART RATE AND ACCURACY OF RPPG MEASUREMENTS

This study presents information regarding the accuracy evaluation of a remote photoplethysmography (rPPG) technique in a gaming context. The technique was applied to estimate the HR of subjects behaving naturally in gaming sessions with induced boredom and stress. Previous research with experiments involving emotions and rPPG was conducted under extremely controlled situations with few game-related stimuli. Subjects did not interact with a complete digital game in any of the experiments, which restricted the accuracy evaluation of rPPG techniques within the context of games research, for instance. Authors commonly used images, videos or text as content to produce the emotional stimuli, in experimental sessions lasting from 20 seconds to 10 minutes.

The aforementioned, non-game stimuli content is less likely to produce the reactions of a real gaming session, e.g. spontaneous body movement and facial actions. In contrast to such research, in this study each subject spent an average of 25 minutes in the gaming session, playing three different games that were custom-made to provoke the emotional reactions similar to a natural play session. Furthermore, the subjects were not instructed
regarding how they should move, thus body and facial reactions would likely be the ones
the subject would normally exhibit in a gaming context.

The recordings of the game sessions of each subject were divided into video segments of
1 minute each. The rPPG technique by Poh, McDuff, and Picard (2011) was applied to
each of the video segments to estimate the HR of the subjects. An accuracy evaluation
of the estimated HR obtained from the video segments was made in relation to the HR
calculated from ground truth.

8.7.1 ANALYSIS AND METHODS

Firstly, the videos of each game session were divided into several video segments of 60
seconds each, denoted as $V_i$, where $i \in [1, 2, ..., n]$ represents the interval (1 represents
the time from 0:00 to 1:00, 2 represents the time from 1:00 to 2:00, and so on). The
use of 60 seconds as the duration of each video segment is based on the work by Poh,

Since the difficulty level of the games constantly increases, different subjects might have
played the same game for longer or shorter periods of time. As a consequence, the in-
terval $n$ represents the last available interval for each subject, which is likely to be dif-
ferent among the subjects. Any remaining video segment of less than 60 seconds was
discarded, i.e. if the duration of a game session was not a multiple of 60. Thereafter, the
$HR_{gt}(V_i)$ was calculated, which is the mean HR obtained from the ground truth data
of a subject while playing during the video segment $V_i$. Any HR value equal to zero was
excluded from the calculation, under the assumption it was a miss-reading.

As previously mentioned in Section 6.2, the rPPG technique proposed by Poh, McDuff,
and Picard (2011) is consolidated, extensively mentioned in the literature and presents
the best SNR for HR estimation under non-exercising situations. For those reasons,
this technique (referred to as the rPPG technique from now on) was selected to perform
the extraction of HR from the video segments. For the comparison, $HR_{video}(V_i)$ was
calculated, which is the estimated HR from video segment $V_i$ obtained with the rPPG
technique. The evaluation of the accuracy of $HR_{video}$ compared to $HR_{gt}$ was based on
statistical methods used by previous works (Poh, McDuff, and Picard, 2011; Rouast et
al., 2016; Li et al., 2014). The measurement error is calculated as:

$$HR_{error} = HR_{video} - HR_{gt}$$ (8.4)

where $HR_{video}$ is the set of HR estimations by the rPPG technique from the video seg-
ments and $HR_{gt}$ is the set of HR means calculated from the ground truth obtained from
the watch, as previously described in Section 8.7. $HR_{error}$ was calculated with video
segments of a given game (for the analysis of that game) or with all available segments
(for the analysis of all games combined).

The following measurements were also calculated: mean of $HR_{error}$ denoted as $M_e$;
standard deviation of $HR_{error}$ denoted as $SD_e$; root mean square error (RMSE) of $HR_{error}$;
mean of error-rate percentage, calculated as:

$$M_eRate = \frac{1}{N} \sum_{i=1}^{N} \frac{|HR_{error}(V_i)|}{HR_{gt}(V_i)}$$ (8.5)

where $V_i$ is a video segment and $N$ is the total of video segments for a given game (or
for all games); finally, the linear correlation between $HR_{video}$ and $HR_{gt}$ accessed using
Table 8.6: Performance of the rPPG technique applied to the testing set

<table>
<thead>
<tr>
<th>$M_e$ (bpm)</th>
<th>$SD_e$ (bpm)</th>
<th>RMSE (bpm)</th>
<th>$M_{eRate}$ (%)</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.25</td>
<td>1.41</td>
<td>1.40</td>
<td>1.52</td>
<td>0.99*</td>
</tr>
</tbody>
</table>

* $p < 0.001$

Pearson’s correlation coefficient $r$.

The selected rPPG technique was implemented in Matlab R2016a according to the original paper by Poh, McDuff, and Picard (2011). Since the custom algorithm to detect peaks mentioned in the article is unknown, this step was replaced by the identification of the highest peak in the frequency domain after an FFT operation. This operation is commonly used in the HR estimation phase of rPPG techniques, as explained in Section 6.1.

The implementation of the rPPG technique was validated with a procedure similar to the one described by Li et al. (2014). Firstly, a manual inspection of all video recordings from the experiment was conducted and a segment $V'_i$ of 30 seconds (1500 frames) was exerted from each video where the subject presented the least amount of body motion and facial activity. This resulted in a testing set of 20 video segments of 30 seconds each, totaling 30000 frames of data. The mean HR calculated from the ground truth for the testing set was 76.8 bpm (SD 13.4 bpm, min. 55 bpm, max. 110 bpm).

The rPPG technique was then applied to each of the $V'_i$ segments to estimate the HR, evaluating the estimated values using ground truth and the statistical methods previously described. The results are presented in Table 8.6. The numbers indicate that the implemented rPPG technique produces accurate and statistically significant results for the estimations, which are aligned with those reported in the original paper. Therefore, it is assumed that the rPPG technique is correctly implemented and further accuracy measurements obtained during the analysis are due to subject activity, not implementation errors.

8.7.2 RESULTS

The performance of the rPPG technique regarding the extraction of the HR is presented in Table 8.7. The first three rows of the table present the performance evaluation calculated with data from each game, while the last row presents the same performance evaluation calculated with data from all the games combined. Regarding the analysis of all the games combined, the mean estimation error $M_e$ was 2.99 bpm ($SD_e$ 18.83 bpm), RMSE was 19.03 bpm and $M_{eRate}$ was 10.31%. The low value for $M_e$ suggests the feasible overall accuracy of the technique; however, the high values for $SD_e$ and $M_{eRate}$ suggest significant variation among the estimations.

As demonstrated by $M_{eRate}$, which is the mean of error-rate percentage, the estimation error of the rPPG technique was equivalent to 10.31% of the expected HR value calculated from ground truth, on average. At a game level, the mean estimation error $M_e$ was 2.96 bpm ($SD_e$ 19.45 bpm) in the Mushroom game, 0.31 bpm (13.51 bpm) in the Platformer game and 5.18 bpm (21.45 bpm) in the Tetris game. RMSE and $M_{eRate}$ were 19.59 bpm and 10.88% in the Mushroom game, 13.43 bpm and 7.82% in the Platformer game, and 21.97 bpm and 11.64% in the Tetris game, respectively.
All three games presented acceptable values for \( M_e \) and significantly higher values for \( SD_e \) and \( M_eRate \), which also suggests feasible results with considerable variations in the estimation error during the analysis of the subjects for each game. In particular, the estimations calculated during the Platformer game presented the lowest values for \( M_e \), RMSE and \( M_eRate \), which indicates that the rPPG technique performed with fewer errors and variations among subjects in the Platformer game than it did in the other two games.

The Pearson’s correlation coefficient \( r \) regarding \( HR_{gt} \) (mean HR calculated from the ground truth) and \( HR_{video} \) (mean HR estimated via rPPG) was 0.45, 0.55 and 0.37 for the Mushroom, Platformer and Tetris game, respectively. All correlations have statistical significance, \( p < 0.001 \). The correlation is illustrated in Figure 8.4. For all three games, there is a positive and medium strength correlation between \( HR_{gt} \) and \( HR_{video} \), which also indicates that the estimations performed by the rPPG technique are feasible. The correlation is stronger in the Platformer game, followed by the Mushroom game and finally by the Tetris game.

![Pearson correlation](image)

Figure 8.4: Statistical correlation of \( HR_{gt} \) and \( HR_{video} \) applied to the video segments of each game, as well as to the video segments of all games.

To better analyze the variations regarding estimation errors among subjects, Figures 8.5 and 8.6 show a distribution of values of \( M_e \), RMSE and \( M_eRate \) for all games combined and individually. The x-axis represents intervals of values of \( M_e \), RMSE or \( M_eRate \) while the y-axis represents the percentage of subjects that presented an estimation error within the interval informed in the x-axis.

Regarding the distribution of values of \( M_e \), shown in Figure 8.5, an overall 66.1% of the subjects presented estimations with \( M_e \) within the interval [-5 bpm, 5 bpm]. For the remaining 33.9% of the subjects, \( M_e \) was spread within the interval [-20 bpm, 35 bpm].

Table 8.7: Accuracy measurements of the rPPG technique when applied to the video segments of a given game and of all games

<table>
<thead>
<tr>
<th>Game</th>
<th>( M_e ) (bpm)</th>
<th>( SD_e ) (bpm)</th>
<th>RMSE (bpm)</th>
<th>( M_eRate ) (%)</th>
<th>( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mushroom</td>
<td>2.96</td>
<td>19.45</td>
<td>19.59</td>
<td>10.88</td>
<td>0.45*</td>
</tr>
<tr>
<td>Platformer</td>
<td>0.31</td>
<td>13.51</td>
<td>13.43</td>
<td>7.82</td>
<td>0.55*</td>
</tr>
<tr>
<td>Tetris</td>
<td>5.18</td>
<td>21.45</td>
<td>21.97</td>
<td>11.64</td>
<td>0.37*</td>
</tr>
<tr>
<td>All</td>
<td>2.99</td>
<td>18.83</td>
<td>19.03</td>
<td>10.31</td>
<td>0.43*</td>
</tr>
</tbody>
</table>

* \( p < 0.001 \)
At a game level, $M_e$ was within the interval $[-5 \text{ bpm}, 5 \text{ bpm}]$ for 65%, 68.4% and 65% of the subjects of the Mushroom, Platformer and Tetris game, respectively. The values for $M_e$ are more equally distributed for the Platformer game, which explains the lower values of $SD_e$ for that game compared to the Mushroom and the Tetris game, which present less equally distributed values of $M_e$.

The distribution of values of RMSE, shown in the first row of Figure 8.6, indicates that overall values were lower than 10 bpm for 59.4% of the subjects, while the remaining of the subjects had RMSE varying from 10 bpm to 50 bpm. At a game level, RMSE was lower than 10 bpm for 50%, 68.5% and 60% of the subjects of the Mushroom, Platformer and Tetris game, respectively.

Regarding $M_{eRate}$, shown in the second row of Figure 8.6, overall 69.5% of subjects had HR estimations that were up to 10% different than the expected HR from ground truth. At a game level, in total 73.7% and 70% of the estimations performed by the rPPG technique during the Platformer and the Tetris game, respectively, presented $M_{eRate}$ less or equal to 10%. These values are slightly better than the 65% of subjects with $M_{eRate}$ up to 10% in the Mushroom game. Despite the fact that $M_{eRate}$ was similar for both the Platformer and Tetris games, the former presented no subjects whose $M_{eRate}$ was greater than 30%, while the latter presented 10% of the subjects with $M_{eRate}$ greater than 30%.

![Figure 8.5: Distribution of values of $M_e$ for all games. The x-axis represents intervals of values of $M_e$ while the y-axis represents the percentage of subjects that presented an estimation error within the interval informed in the x-axis.](image)

8.7.3 DISCUSSION

The results obtained indicate that the use of the selected rPPG technique to estimate HR from videos of gaming sessions is feasible. When the technique was applied to a testing set of 20 manually selected 30 seconds long video segments, whose subject’s facial activity and body movement were minimal, the estimations were significantly accurate. As demonstrated in Table 8.6, the mean of error-rate $M_{eRate}$ was 1.52% and the Pearson’s correlation coefficient was $r = 0.99$ for that testing set. Those results were expected since the videos featured an unrealistic condition where the subjects remained mostly still with a neutral face.

When the rPPG technique was applied to all gaming sessions, however, body movement and facial activity significantly impacted the estimation performance. It is aligned with
previously described works in the literature, which indicate that the estimation error increases when subject activities increase (W. Wang, Stuijk, and Haan, 2016).

The elevated values for $SD_e$, the standard deviation of $Me_e$, suggest significant variations in the estimations among subjects in each video segment. The estimation discrepancies do not seem to be caused by errors equally spread among all the gaming sessions, but due to a subset of problematic ones instead. The discrepancies and skewness of the estimations are visible in the scatter plot of the estimated and expected HR values in Figure 8.4. It shows a cluster of points for each game, however, it is surrounded by significantly wrong estimation points. In the Mushroom game, for instance, 5 estimations (bottom right of the chart) were in the interval [120 bpm, 181 bpm] bpm, which is significantly outside the expected ground truth interval of [80 bpm, 110 bpm]. Similar, significantly wrong estimations can also be seen in the Platformer and the Tetris game.

The skewed distribution of values of $Me_e$, $MeRate_e$ and RMSE illustrated in Figures 8.5 and 8.6 also support that indication. Considering the estimations for all games, in total 69.5% of them presented $MeRate_e$ related to an estimation value that was less than 10% different than the expected HR from ground truth. Additionally 59.4% of all estimations presented RMSE lower than 10 bpm. That result is slightly worse than similar works that used rPPG techniques and subjects featuring natural movements, with reported RMSE between 0.11 and 7.28 bpm.

A direct comparison of the results of this study to the ones of such similar works is unfair, however. Despite the fact that the aforementioned works present experiments where subjects are told to behave naturally, their accuracy evaluation is based on artificial human-computer interactions, as previously described in Section 8.7. The accuracy results of the present study account for body and facial movement caused by games whose focus is entertainment, not artificial interactions. As a consequence, the results are more connected to a scenario involving real and spontaneous reactions to games,

Figure 8.6: Distribution of values of RMSE and $MeRate_e$ for all games. The x-axis represents intervals of values of RMSE or $MeRate_e$ while the y-axis represents the percentage of subjects that presented an estimation error within the interval informed in the x-axis.
showing that the estimations of the rPPG technique are feasible, but skewed by other factors such as natural facial activity and subject movement.

The differences in estimation also seem to be connected to the particularities of each game and subject. Considering the distribution of values of RMSE and $M_{rate}$, both the Platformer and the Tetris games presented more estimations with fewer errors than the Mushroom game. The Mushroom game presented 15% of its estimations with RMSE greater than 30 bpm and $M_{rate}$ greater than 30%, which are significantly wrong estimations.

In order to further explore such differences in estimations, the variations of movement and size of the ROI used to track the subject’s face along the videos were analyzed. A stable ROI (both in shape and movement) is required for a precise extraction of the photoplethysmographic signal, thus significant variations in the ROI lead to estimation errors. The mean position of the center point of the ROI for each subject in each gaming session was calculated. For each subject in each game session, the Euclidean distance between the center point of the ROI of each frame and the mean center point of the ROI previously calculated (for that subject in that session) was measured.

Similarly, the mean length of the ROI diagonal for each subject in each game session was calculated, subtracting it from the length of the ROI diagonal of each frame in that gaming session. Since the time duration of game sessions differs, the subject’s progress in the game was normalized using the interval $[0, 1]$, where 0 is the start point of the gaming session and 1 its end. The measurements were also subtracted from the sessions mean to facilitate analysis and comparison among different games/subjects.

Figure 8.7 illustrates some of the patterns observed in the investigation of the distance of the ROI central position. Each row in the figure contains three charts showing the variations of the ROI central position along the gaming sessions of a given subject. The first row contains the investigation of subject 17, who presented, for all his/her gaming sessions combined, -0.33 for $M_{e}$ ($SD_{e}$ 1.4) and 1.39 for RMSE (low estimation errors). The second row shows subject 3, who presented 11.47 for $M_{e}$ ($SD_{e}$ 16.47) and 19.62 for RMSE (moderate to high estimation errors). Finally the third row shows subject 1, who presented 15.94 for $M_{e}$ ($SD_{e}$ 28.5) and 31.96 for RMSE (high estimation errors).

The estimations performed on subject 17 were significantly accurate and the charts regarding the variation of the ROI central position show a stable progression along all gaming sessions. The distance variation (y-axis) remains mostly concentrated within the interval of $[0, 100]$ pixels for all the games, which suggests the subject presented few or short movements during gaming sessions. Subject 3 also presented low variation in the Platformer and the Tetris game, however, there is a significant variation in the ROI central position in the Mushroom gaming session. The chart indicates significant distance variations of the ROI that are above 200 pixels in a certain period of the game. Finally subject 1 presents high variations in the ROI distance in all gaming sessions, as demonstrated by points above the 200 pixels mark regarding the difference to mean. The Tetris game, in particular, presents distance variations above 200 pixels during almost the whole session.

Figure 8.8 illustrates the same subjects regarding the investigation of the variations of the ROI diagonal length. Similar to the variations of the ROI central position, the variation of the ROI diagonal length (y-axis) is lower for subject 17 (first row of charts in the figure), since the majority of the values are close to zero. Subject 3 also presents low variations in the ROI diagonal length during the Platformer and the Tetris game, however there are significant changes in the ROI size during a period in the Mushroom game. In this period, the length of the ROI diagonal is negative, i.e. -400 pixels, which indicates
that the size of the detected ROIs for those frames was smaller than the mean ROI diagonal length, perhaps caused by a wrongly detected face (false-negative), for instance. Finally, subject 1, to some extent, presents variations of the ROI diagonal length during the majority of his/her gaming sessions. These constant variations could be caused by the inability of the face tracking algorithm to steadily and continuously detect the subject’s face along the frames of the video. The chart shows a distribution of values along the zero mark regarding difference to mean, however they are more spread than those of subject 17, for instance, which indicates higher instability of the ROI size/detection for subject 1. In the Tetris session of subject 1, for instance, there are extreme variations in the ROI diagonal length, with values close to -400 pixels, similar to the ones of subject 3 in the Mushroom game. Such an extreme variation could also be explained by an incorrectly detected face area in those frames.

An inspection of the videos of the subjects with patterns similar to those of subjects 3 and 1 revealed a sensitive amount of movement and facial activity, including occlusion of the face by the subject’s hand, as illustrated by Figure 8.9. Any facial occlusion influences the face tracking algorithm used (Viola&Jones), since it could detect the face position incorrectly or not detect it at all. A flawed face detection step affects the extraction of the plethysmographic signal, because noise is extracted along with the raw signal, preventing the rPPG technique from being able to separate it properly.

Despite the efforts to create games that prevent face occlusion by a subject’s hand, such behavior seems to be natural in boring situations. In both the Mushroom and Tetris
It is possible to speculate that the variations regarding the movement and size of the ROI, which directly influence estimation accuracy of the rPPG technique, seem to be related to the unique behavior of each user as well. As illustrated by Figures 8.7 and 8.8, the subjects present different movement patterns. A previous analysis of the videos indicates significant differences regarding facial activities among subjects (Bevilacqua, Backlund, and Engström, 2016). It strengthens the idea of a user-tailored model able to deal with such peculiarities, and thus more likely to produce better estimations. A method that operates its estimations on the basis of average user behavior is prone to

Figure 8.8: Variations of the ROI diagonal length for subjects 17 (low estimation errors), 3 (moderate to high estimation errors) and 1 (high estimation errors) during their gaming sessions. Values were subtracted from the session mean to facilitate analysis and comparison among different games/subjects.

Figure 8.7: Progress in the game (normalized) and Difference to mean (pixels) for Mushroom, Platformer, and Tetris games.
be significantly affected by specific user behavior outside the expected mean pattern, causing a skewed distribution of estimation errors, such as those presented in Figure 8.6 regarding $McRate$ and RMSE.

8.7.4 CONCLUSIONS

Overall, the estimation of the rPPG was feasible, showing a mean estimation error of 2.99 bpm (SD 18.83 bpm), RMSE of 19.03 bpm, as well as a positive and medium strength Pearson correlation of $r = 0.43$, $p < 0.001$. On average, the estimation error of the rPPG technique was up to 10.31% of the expected value calculated from ground truth. Additionally the exploratory investigation regarding factors that impacted the accuracy of the rPPG technique, such as variations in the region of interest (ROI) used to remotely extract the HR signal, suggests that factors connected to the type of game being played and the unique behavior of each subject influenced the estimations. Among the identified causes of such influence were body movement, e.g. head tilt and rotation, and facial occlusion by subjects hand.

8.8 STUDY 4: AUTOMATED FACIAL ANALYSIS

This study presents a method for the automated analysis of facial cues from videos, as well as an empirical evaluation of its application as a potential tool for detecting stress and boredom in players. The proposed automated facial analysis is based on the measurement of facial features ($F_1$ to $F_7$) calculated from facial landmarks obtained unobtrusively via computer vision.
8.8.1 FACIAL FEATURES

The proposed automated facial analysis is based on the measurement of 7 facial features calculated from 68 detected facial landmarks. Table 8.8 presents the facial features, which are illustrated in Figure 8.10(b). The facial features are mainly based on the Euclidean distances between landmarks, similar to some works previously mentioned. However, the approach does not rely on pre-defined expressions, i.e. the 6 universal facial expressions, training of a model, or the use of the MPEG-4 standard, which specifies representations for 3D facial animations, not emotional interactions in games. Additionally, the method does not use an arbitrarily selected frame, e.g. the 100th frame (Giannakakis et al., 2017), as a reference for calculations, since features are derived from each frame (or a small set of past frames). The features are obtained unobtrusively via computer vision analysis that focuses on detecting activity of facial muscles reported by previous work involving EMG and emotion detection in games.

The process of extracting the facial features has two main steps: face detection and feature calculation. In the first step, computer vision techniques are applied to a frame of the video and facial landmarks are detected. In the second step, the detected landmarks are used to calculate several facial features related to eyes, mouth, and head movement. The following sections present a detailed description of how each step is performed, including the calculation of features.
Table 8.8: Information regarding calculated facial features

<table>
<thead>
<tr>
<th>Name</th>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouth outer</td>
<td>$F_1$</td>
<td>Sum of the Euclidean distance between the mouth contour landmarks and the anchor landmarks. It monitors the zygomatic muscle.</td>
</tr>
<tr>
<td>Mouth corner</td>
<td>$F_2$</td>
<td>Sum of the Euclidean distance between the mouth corner landmarks and the anchor landmarks. It monitors the zygomatic muscle.</td>
</tr>
<tr>
<td>Eye area</td>
<td>$F_3$</td>
<td>Area of the regions bounded by the closed curves formed by the landmarks in the contour of the eyes. It monitors the orbicularis oculi muscle.</td>
</tr>
<tr>
<td>Eyebrow activity</td>
<td>$F_4$</td>
<td>Sum of the Euclidean distance between eyebrow landmarks and the anchor landmarks. It monitors the corrugator muscle.</td>
</tr>
<tr>
<td>Face area</td>
<td>$F_5$</td>
<td>Area of the region bounded by the closed polygon formed by the most external detected landmarks.</td>
</tr>
<tr>
<td>Face motion</td>
<td>$F_6$</td>
<td>Average value of the Euclidean norm of a set of landmarks in the last $N$ frames. It describes the total distance the head has moved in any direction in a short period of time.</td>
</tr>
<tr>
<td>Facial COM</td>
<td>$F_7$</td>
<td>Average value of all detected landmarks. It describes the overall movement of all facial landmarks.</td>
</tr>
</tbody>
</table>
Face detection

The face detection procedure is performed for every frame of the input video. Initially, the face is detected using a Constrained Local Neural Field (CLNF) model (Baltrusaitis, Robinson, and Morency, 2013; Baltrusaitis, Robinson, and Morency, 2016). CLNF uses a local neural field patch expert that learns the nonlinearities and spatial relationships between pixel values and the probability of landmark alignment. The technique also uses a non-uniform regularized landmark Mean Shift fitting technique that takes into consideration patch reliabilities. It improves the detection process under challenging conditions, e.g. extreme face pose or occlusion, which is likely to happen in game sessions (see Section 8.5). The application of the CLNF model to a given video frame produces a vector $L$ of 68 facial landmarks:

$$L = [p_0, p_1, p_2, \ldots, p_{67}]^T$$

where $p_i$ is a detected facial landmark that represents a 2D coordinate $(x_i, y_i)$ in the frame. Facial landmarks are related to different facial regions, such as eyebrows, eyes and lips. Figure 8.10(a) illustrates the landmarks of $L$ in a given frame.

Anchor landmarks

The calculation of the proposed facial features involves the Euclidean distance among facial landmarks. Subsequently, the Euclidean distance between two landmarks $a_1 = (x_1, y_1)$ and $a_2 = (x_2, y_2)$ is given as:

$$d(a_1, a_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Landmarks in the nose area are more likely to be stable, presenting fewer position variations in consecutive frames (Giannakakis et al., 2017). Consequently, they are good
reference points for use in the calculation of the Euclidean distance among landmarks. In order to provide stable reference points for the calculation of our facial features, three highly stable landmarks located in the nose line were selected, denoted as the anchor vector $\mathbf{A} = [p_{28}, p_{29}, p_{30}]^T$. The landmarks of the anchor vector $\mathbf{A}$ are highlighted in yellow in Figure 8.10(a).

Feature normalization

The subjects moved towards and away from the camera during the gaming sessions. This movement affects the Euclidean distance between landmarks, as it tends to increase when the subject is closer to the camera, for instance. Additionally, the subjects have unique facial shapes and characteristics, which also affect the calculation and comparison of the facial features between subjects. To mitigate that problem, a normalization coefficient $K$ was calculated as the Euclidean distance between the upper and lowermost anchor landmarks in $\mathbf{A}$. In other words, $K$ represents the size of a subject’s nose line. Since all features are divided by $K$, their final value is expressed as normalized pixels (relative to $K$) rather than pixels per se.

Mouth related features

Mouth related features aim to detect activity in the zygomatic muscles, illustrated in Figure 4.1 (c and d, page 33), which are related to changes in the mouth, such as lips activity (stretch, suck, press, parted, tongue touching, bite) and movement (including talking). Two facial features related to the mouth area were calculated: mouth outer and mouth corner.

Mouth outer ($F_1$): given vector $\mathbf{M} = [p_{48}, p_{49}, \ldots, p_{60}]^T$ containing the landmarks in the outer part of the mouth (highlighted in orange in Figure 8.10(a)). The mouth outer feature is calculated as the sum of the Euclidean distance among the landmarks in $\mathbf{M}$ and the anchor landmarks in $\mathbf{A}$:

$$F_1 = \frac{1}{K} \sum_{i=1}^{12} \sum_{j=1}^{3} d(A_j, M_i)$$

where $A_j$ and $M_i$ are the $j$-th and $i$-th element of $\mathbf{A}$ and $\mathbf{M}$, respectively.

Mouth corner ($F_2$): given vector $\mathbf{C} = [p_{48}, p_{54}]^T$, containing the two landmarks representing the mouth corners (highlighted in pink in Figure 8.10(a)). The mouth corner feature is the sum of the Euclidean distance among the landmarks in $\mathbf{C}$ and $\mathbf{A}$:

$$F_2 = \frac{1}{K} \sum_{i=1}^{2} \sum_{j=1}^{3} d(A_j, C_i)$$

where $A_j$ and $C_i$ are the $j$-th and $i$-th element of $\mathbf{A}$ and $\mathbf{C}$, respectively.

Eye related features

Eye related features aim to detect activity related to the orbicularis oculi and the corrugator muscles, illustrated in Figure 4.1(b) and Figure 4.1(a) respectively, which comprehend changes in the eyes region, including eye and eyebrow activity. Two facial features related to the eyes were calculated: eye area and eyebrow activity.
Eye area ($F_3$): given vector $Y_l = [p_{36}, p_{37}, \ldots, p_{41}]^T$ containing the landmarks describing the left eye, highlighted in green in Figure 8.10(a), and vector $Y_r = [p_{42}, p_{43}, \ldots, p_{47}]^T$ containing the landmarks describing the right eye, highlighted in green in Figure 8.10(a). The eye area feature is the area of the regions bounded by the closed curves formed by the landmarks in $Y_l$ and $Y_r$, divided by $K$. The area of the curves is calculated using OpenCV’s contourArea() function, which uses Green’s theorem (Stewart, 2011).

Eyebrow activity ($F_4$): calculated as the sum of the Euclidean distances among the eyebrow landmarks and the anchor landmarks in $A$. Given the vector $W_l = [p_{17}, p_{18}, \ldots, p_{21}]^T$ containing the landmarks describing the left eyebrow, highlighted in blue in Figure 8.10(a), and the set $W_r = [p_{22}, p_{23}, \ldots, p_{26}]^T$ containing the landmarks describing the right eyebrow, highlighted in blue in Figure 8.10(a). The eyebrow activity feature is calculated as:

$$F_4 = \frac{1}{K} \sum_{i=1}^{5} \sum_{j=1}^{3} \left[ d(A_j, W_{l,i}) + d(A_j, W_{r,i}) \right]$$

where $A_j$, $W_{l,i}$ and $W_{r,i}$ are the $j$-th, $i$-th and $i$-th element of $A$, $W_l$ and $W_r$, respectively.

Head related features

Head related features aim to detect body movements, in particular, variations of head pose and amount of motion the head/face performs over time. Three features related to the head were calculated: face area, face motion and facial center of mass (COM).

Face area ($F_5$): during the interaction with a game, subjects tend to move towards (or away from) the screen, which causes the facial area in the video recordings to increase or decrease. Given vector $F = [p_0, p_1, \ldots, p_{16}]^T$ containing the landmarks describing the contour of the face, highlighted in red in Figure 8.10(a). The face area feature is the area of the region bounded by the closed curves formed by the landmarks in $F \cup W_r \cup W_l$, divided by $K$. Similar to the eye area, the area under the curves is calculated using OpenCV’s contourArea() function.

Face motion ($F_6$): accounts for the total distance the head has moved in any direction in a short period of time. For each frame of the video, the currently detected anchor vector $A$ is saved, which produces vector $D = [A_1, A_2, \ldots, A_n]^T$, where $A_i$ is the vector $A$ detected in the $i$-th frame of the video and $n$ is the number of frames in the video. Subsequently the face motion feature is calculated as:

$$F_6 = \frac{1}{K} \sum_{j=1}^{Z} \sum_{t=1}^{3} ||D(f - t, j) - D(f - Z, j)||$$

where $Z$ is the number of frames to include in the motion analysis, $D(i, j)$ is the $j$-th element of $A_i \in D$, $f$ is the number of the current frame, and $||.||$ is the Euclidean norm. In the analysis presented here, a value of $Z = 50$ (50 frames, equivalent to 1 second) was used.

Facial COM ($F_7$): describes the overall movement of all facial landmarks. A single 2D point, calculated as the average of all landmarks in $L$, is used to monitor the movement. The COM feature is calculated as:
Figure 8.11: Extraction of video segments $H_0$ and $H_1$ containing boring and stressful game interactions, respectively. Initial $D$ seconds of any video $V_{s, i}$ are ignored and the remaining is divided into three pieces, from which the first and the last ones are selected. Stripes highlight discarded video segments.

\[
F_T = \frac{1}{K} \frac{1}{N} \sum_{i=1}^{N} ||p_i||
\]

where $N$ is the total number of detected landmarks (elements in $L$) and $||.||$ is the Euclidean norm.

8.8.2 ANALYSIS AND METHODS

Data pre-processing

The pre-processing of video recordings involved the extraction of parts containing the interaction with the games and the discarding of noisy frames. Firstly, the periods showing subjects playing each available game were extracted from the video recordings. This resulted in three videos per subject, denoted as $V_{s, i}$ where $s$ is the $s$-th subject and $i \in \{1, 2, 3\}$ represents the game.

As previously mentioned, the games used as emotional elicitation material in the experiment induced variations of physiological signals in the subjects, who perceived the games boring at the beginning and stressful at the end. Since the aim of the present study was to test the potential of facial features to differentiate emotional states of boredom and stress, two video segments were extracted from each video $V_{s, i}$, named $H_0$ and $H_1$, whose subject’s emotional state was assumed to be known and related to boredom and stress. In order to achieve that, the following extraction procedure was performed, illustrated in Figure 8.11. Firstly a duration of $D$ seconds of any given video $V_{s, i}$ was ignored. For this pre-processing, $D = 60$ was used. The remainder of the video was then divided into three segments, from which the first and the last were selected as $H_0$ and $H_1$, respectively.

The reason $D = 60$ seconds was discarded from all the video’s segments is because the initial minute might not be ideal for a fair analysis. During the first minute of gameplay, subjects are less likely to be in their usual neutral emotional state. They are more likely to be stimulated by the excitement of the initial contact with a game soon to be played, which interferes with any feelings of boredom. Additionally, subjects need basic experimentation with a game to learn how to play it and assess whether it is boring or not. This claim is supported by the empirical analysis of the first minute of the video recordings, which show repeated head and eye movements to and from the keyboard/display. Consequently, the second minute and onward in the videos is more likely to portray facial activity related to emotional reactions to the game than facial activity connected to...
gameplay learning. Regarding the division of the remaining part of the video into three segments, from which two were selected as $H_0$ and $H_1$, it followed the reasoning that the emotional state of the subjects was unknown in the middle part of $V_{s,i}$. Based on the self-reported emotional states, subjects reported the beginning part of the games as boring and the final part as stressful. Additionally, there are significant differences in the HR mean between the second and the last minute of gameplay in the games (Bevilacqua, Engström, and Backlund, 2018c). As a consequence, it is assumed that the video segments $H_0$ and $H_1$ accurately portray the interaction of the subjects during boring and stressful periods of the games, respectively.

The pre-processing of the recordings resulted in 6 video segments per subjects: 3 segments $H_0$ (one per game) and 3 segments $H_1$ (one per game). A given game $i$ contains $N = 20$ pairs of $H_0$ and $H_1$ video segments (20 segments $H_0$, one per subject, and 20 segments $H_1$, one per subject). With regard to all the subjects and games, there are $N = 60$ pairs of $H_0$ and $H_1$ video segments (3 games × 20 subjects, resulting in 60 segments $H_0$ and 60 segments $H_1$). Subject 9 had problems playing the Platformer game, consequently, segments $H_0$ and $H_1$ from subject 9 in the Platformer game were discarded. Therefore, the Platformer game contains $N = 19$ pairs of $H_0$ and $H_1$ video segments; regarding all the games and subjects, there are $N = 59$ pairs of $H_0$ and $H_1$ video segments.

Feature analysis

The previously mentioned features can be calculated for each frame of any given video, however facial cues might be better contextualized if analyzed in multiple frames. For that reason, the facial analysis was applied to all the frames of all the video segments $H_0$ and $H_1$. The mean value of each facial feature in each video segment was then calculated. As a result, any facial feature $F_i$ has $N = 59$ pairs of mean values (59 from $H_0$ and 59 from $H_1$). Henceforth, the set of mean values in $H_0$ or $H_1$ of a given feature $F_i$ will be referred to simply as feature value in $H_0$ or $H_1$, respectively.

Based on a previous manual analysis of facial actions of the video recordings (Bevilacqua, Backlund, and Engström, 2016) and findings of related work, the values of facial features during boring periods of the games are expected to be different than those during stressful periods. Since the subjects perceived the games as boring at the beginning and stressful at the end, it is assumed that values in $H_0$ and $H_1$, for all features, are likely to correlate with an emotional state of boredom and stress, respectively. Therefore, the overarching hypothesis is stated as follows: the mean value of features in $H_0$ is different than the mean value in $H_1$, for all the subjects and games. More specifically, the overarching hypothesis can be described as 7 sub-hypotheses, denoted $u_i$, where $i \in \{1, 2, ..., 7\}$. Hypothesis $u_i$ states that the true difference in means between the value of a given feature $F_i$ in $H_0$ and $H_1$, for all the subjects, is greater than zero. The dependent variable of $u_i$ is $F_i$ and the null hypothesis is that the true difference in means between $H_0$ and $H_1$ for feature $F_i$, for all the subjects and games, is equal to zero.

Hypothesis $u_i$ was tested by performing a paired two-tail t-test on the values $H_0$ and $H_1$ of feature $F_i$. In total, 7 tests were performed: $u_1$ (mouth outer), $u_2$ (mouth corner), $u_3$ (eye area), $u_4$ (eyebrow activity), $u_5$ (face area), $u_6$ (face motion), and $u_7$ (facial COM).

### 8.8.3 RESULTS

Table 8.9 presents the mean of differences of all the features between periods $H_0$ and $H_1$, calculated for all the subjects in all the games, according to the description in Section 8.8.1, and analyzed according to the procedures described in Section 8.8.2. The
mean of differences of all the features shows a decrease from $H_0$ to $H_1$. Comparing the mean difference of a feature to its mean value in $H_0$, the decrease from $H_0$ to $H_1$ was 10.7% for mouth outer ($F_1$), 11.8% for mouth corner ($F_2$), 10.4% for eye area ($F_3$), 8.1% for eyebrow activity ($F_4$), 9.4% for face area ($F_5$), 8.2% for face motion ($F_6$), and 11% for facial COM ($F_7$). Changes related to $F_6$ and $F_7$ were not statistically significant. All the remaining features presented statistically significant changes from $H_0$ to $H_1$. The greatest decrease with statistical significance was associated with mouth corner, followed by mouth outer, eye area, face area, and eyebrow activity. These numbers support the experimental expectations that values for facial features are different when a comparison is made between two distinct parts of the games, i.e. boring and stressful ones.

Table 8.9: Mean of differences ($\pm$SD) of features between periods $H_0$ and $H_1$ ($N = 59$). Units expressed in normalized pixels.

<table>
<thead>
<tr>
<th>Feature (notation)</th>
<th>Mean difference ($\pm$SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouth outer ($F_1$)</td>
<td>$-20.59 \pm 57.36^{**}$</td>
</tr>
<tr>
<td>Mouth corner ($F_2$)</td>
<td>$-3.90 \pm 10.16^{**}$</td>
</tr>
<tr>
<td>Eye area ($F_3$)</td>
<td>$-0.019 \pm 0.064^{*}$</td>
</tr>
<tr>
<td>Eyebrow activity ($F_4$)</td>
<td>$-15.59 \pm 49.71$</td>
</tr>
<tr>
<td>Face area ($F_5$)</td>
<td>$-2.60 \pm 7.90^{*}$</td>
</tr>
<tr>
<td>Face motion ($F_6$)</td>
<td>$-44.97 \pm 326.74$</td>
</tr>
<tr>
<td>Facial COM ($F_7$)</td>
<td>$-0.029 \pm 0.113$</td>
</tr>
</tbody>
</table>

$p < 0.05$  ** $p < 0.01$

The two facial features related to mouth, i.e. mouth corner and mouth outer, presented a combined average decrease of 11.24% from $H_0$ to $H_1$. The change was the most significant compared to all other features. The mean of differences of $F_1$ and $F_2$ between periods $H_0$ and $H_1$ was $T(59) = -20.59$ (SD 57.36, $p < 0.01$) and $T(59) = -3.90$ (SD 10.16, $p < 0.01$), respectively. Both features had a statistically significant change from $H_0$ to $H_1$, which supports the claim that they are different in those periods. Additionally, both features presented SD as considerably greater than the mean, which indicates that the differences of such features for each subject between periods $H_0$ and $H_1$ are likely to be spread out rather than clustered around the mean value. Features related to eyes, i.e. eye area and eyebrow activity, presented a combined average decrease of 9.28% from $H_0$ to $H_1$. The mean of differences of $F_3$ and $F_4$ between periods $H_0$ and $H_1$ was $T(59) = -0.019$ (SD 0.064, $p < 0.05$) and $T(59) = -15.59$ (SD 49.71, $p < 0.05$), respectively. Similar to mouth-related features, eye-related features had a statistically significant change from $H_0$ to $H_1$, indicating that they are different in those periods. Following the same pattern of change for $F_1$ and $F_2$, both features $F_3$ and $F_4$ also presented a SD considerably greater than the mean, also suggesting that the differences of such features for each subject between periods $H_0$ and $H_1$ are likely to be spread out rather than clustered around the mean value.

Finally, features related to the whole face, i.e. face area, face motion, and facial COM, presented a combined average decrease of 9.52% from $H_0$ to $H_1$. The mean of differences of $F_5$, $F_6$ and $F_7$ were $T(59) = -2.60$ (SD 7.90, $p < 0.05$), $T(59) = -44.97$ (SD 326.74, $p = 0.29$), and $T(59) = -0.029$ (SD 0.113, $p = 0.052$), respectively. Face area was the only feature in this category to present a change that was statistically significant between periods $H_0$ and $H_1$, supporting the idea that $F_5$ is different in those periods. In constrast, $F_6$ and $F_7$ lack statistically significant differences between periods $H_0$ and $H_1$. Similar to
8.8.4 DISCUSSION

The overarching hypothesis states that the mean value of features in $H_0$ is different than the mean value in $H_1$. Furthermore, the overarching hypothesis is composed of 7 sub-hypotheses, i.e. $u_i$, one for each feature $F_i$, where $u_i$ states that the true difference in means between the value of a given feature $F_i$ in $H_0$ and $H_1$ is greater than zero. The majority of the calculated facial features, i.e. mouth outer ($F_1$), mouth corner ($F_2$), eye area ($F_3$), eyebrow activity ($F_4$), and face area ($F_5$), presented statistically significant differences in their mean values when a comparison is made between two distinct parts of the games, i.e. $H_0$ and $H_1$. As previously mentioned, the subjects perceived the first part of the games, i.e. $H_0$, as boring and the second part, i.e. $H_1$, as stressful. The results support the claim of sub-hypotheses $u_1$ to $u_5$, which indicate that facial features $F_1$ to $F_5$ can be differentiated between periods $H_0$ and $H_1$ and can consequentially have the potential to unobtrusively differentiate the emotional states of boredom and stress of players in gaming sessions. The results refute sub-hypotheses $u_6$ and $u_7$, since features $F_6$ and $F_7$ lack statistically significant differences between periods $H_0$ and $H_1$.

Mouth related facial features, i.e. mouth outer ($F_1$) and mouth corner ($F_2$), presented statistically significant differences between boring and stressful parts of the games. Both features are calculated on the basis of the distance between mouth and nose related facial landmarks, which presented a decrease in stressful parts of the games. This decrease could be attributed to landmarks in the upper and lower lips being closer to each other, which could be associated with lips pressing, lips sucking or talking, for instance. Particular to the mouth corner feature, a decrease in distance is the result of the two mouth corners being placed closer to the nose area, which could be associated with smiles or mouth deformation, e.g. mouth corner pull to left/right. Consequentially, a decrease in the mean value of both features suggests greater mouth activity that involves the proximity of mouth landmarks to the nose area in stressful parts of the games compared to boring parts. Such results are aligned with previous studies that show lip pull corner as a frequent facial behavior during gaming sessions (Kaiser, Wehrle, and P. Edwards, 1994) and talking as an emotional indicator (Blom et al., 2014). Additionally, stating that mouth related features were constructed after the zygomatic muscle activity, the results are connected with previous studies that show increased activity of the zygomatic muscle is related to self-reported emotions (Tijss, Brokken, and IJsselsteijn, 2008) and their connection to changes in a game (Ravaja et al., 2006).

Eye related features, i.e. eye area ($F_3$) and eyebrow activity ($F_4$), also presented statistically significant differences between boring and stressful parts of the games. They presented a decrease in the mean value from $H_0$ to $H_1$, which points to landmarks detected in the eyes contour becoming closer to each other in $H_1$. This suggests that more pixels in the eyes area were detected during $H_0$ (boring part) than $H_1$ (stressful part). Such numbers might indicate less blinking activity or more wide-open eyes during boring parts of the games. Additionally, they could indicate more blinking and eye tightening activity (possibly related to frowning) during stressful parts. Both indications are aligned with previous findings, which show increased blinking activity (calculated from eye area) in stressful situations (Giannakakis et al., 2017). Regarding the eyebrow feature, its calculation is based on the distance between facial landmarks in the eyebrow lines and the
A decrease in value indicates a smaller distance between eyebrows and nose, which could be explained by frowning, suggesting that the subjects presented more frowning action during stressful moments of the game. The mean value of eyebrow activity during \( H_0 \) is greater than during \( H_1 \), which indicates that the distance between eyebrows and nose was greater during boring parts of the games compared to stressful parts. It could also be the result of more eyebrow risings, e.g. facial expressions of surprise, in boring periods compared to stressful periods. Eye related features were constructed to monitor the activity of the orbicularis oculi and the corrugator supercilii muscles. The results are related to previous work that reports game events affecting the activity of the orbicularis oculi (Ravaja et al., 2006) and the corrugator (Hazlett, 2006) muscles.

Finally, features related to the whole face, i.e. face area (\( F_5 \)), face motion (\( F_6 \)) and facial COM (\( F_7 \)), are partially conclusive. These features are affected by body motion, e.g. head movement and corporeal posture, therefore, a decrease in value might indicate less corporeal activity during \( H_1 \) compared to \( H_0 \). Face area was the only feature in this category to present a change that was statistically significant. The value of the face area feature is directly connected to the subjects’ movement towards and away from the camera. A decrease in face area from \( H_0 \) to \( H_1 \) suggests that the subjects were closer to the computer screen more often during boring parts of the games than during stressful parts. The facial COM feature also presented a decrease from \( H_0 \) to \( H_1 \). This feature is connected to vertical and horizontal movements, performed by the subject’s face, that are anchored to a fixed reference point and less influenced by head rotations. Despite presenting a change that is not statistically significant (\( p = 0.519 \)), the decrease of facial COM might be an indication that the subjects were motionless more during stressful periods than during boring periods. The face motion feature also presented a decrease from \( H_0 \) to \( H_1 \), that is not statistically significant (\( p = 0.294 \)). This feature accounts for the amount of movement a subject’s face performs in a period of 50 frames (dynamic reference point), which is directly affected by vertical, horizontal and rotational movements of the head. A decrease could be associated with moving/rotating the head less often during the analyzed 50 frames periods in \( H_1 \) than in \( H_0 \). However, the lack of statistical significance suggests the change is not related to a subject’s emotional state, but other factors, such as the inherent behavior associated with game mechanics, i.e. head movement caused by the observation of cards in the Mushroom game. The results lack the statistical significance to replicate the findings of previous work, which relate head movements to changes in games, i.e. failure (Shaker, Asteriadis, et al., 2011) and frustration (Blom et al., 2014), or to stressful situations (Giannakakis et al., 2017).

### Table 8.10: Percentage of change of features from period \( H_0 \) to \( H_1 \) in the Mushroom game (\( N = 20 \)).

<table>
<thead>
<tr>
<th>Feature (notation)</th>
<th>Mean</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouth outer (( F_1 ))</td>
<td>-12.9</td>
<td>-69.1</td>
<td>22.1</td>
</tr>
<tr>
<td>Mouth corner (( F_2 ))</td>
<td>-15.0</td>
<td>-71.6</td>
<td>15.5</td>
</tr>
<tr>
<td>Eye area (( F_3 ))</td>
<td>-8.9</td>
<td>-76.9</td>
<td>8.2</td>
</tr>
<tr>
<td>Eyebrow activity (( F_4 ))</td>
<td>-8.0</td>
<td>-72.3</td>
<td>9.6</td>
</tr>
<tr>
<td>Face area (( F_5 ))</td>
<td>-11.3</td>
<td>-74.5</td>
<td>18.2</td>
</tr>
<tr>
<td>Face motion (( F_6 ))</td>
<td>47.2</td>
<td>-61.3</td>
<td>253.8</td>
</tr>
<tr>
<td>Facial COM (( F_7 ))</td>
<td>-12.9</td>
<td>-81.0</td>
<td>9.8</td>
</tr>
</tbody>
</table>
It could be argued that the characteristics of each game mechanic influence the mean change of features between the two periods. Such argument is particularly true for features that are calculated on the basis of the subject’s body movement, i.e., face area, face motion and facial COM. In that case, subjects could move the face as a result of in-game action, i.e., inspecting mushrooms, rather than an emotional manifestation. Additionally, the mean change of features between the two periods presented SD as considerably greater than the mean value, indicating that differences between periods are likely to be spread out. It suggests significant between-subject variations for each feature or game. In order to further explore such topics, the changes of all the features were analyzed at a game level. Tables 8.10, 8.11, and 8.12 present the mean, minimum and maximum change presented by the features, in percentages, from period \(H_0\) to \(H_1\), calculated from all the subjects in the Mushroom, Platformer and Tetris game, respectively.

Table 8.11: Percentage of change of features from period \(H_0\) to \(H_1\) in the Platformer game \((N = 19)\).

<table>
<thead>
<tr>
<th>Feature (notation)</th>
<th>Mean</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouth outer ((F_1))</td>
<td>-7.4</td>
<td>-54.0</td>
<td>16.9</td>
</tr>
<tr>
<td>Mouth corner ((F_2))</td>
<td>-8.2</td>
<td>-55.9</td>
<td>15.5</td>
</tr>
<tr>
<td>Eye area ((F_3))</td>
<td>-6.8</td>
<td>-30.4</td>
<td>20.0</td>
</tr>
<tr>
<td>Eyebrow activity ((F_4))</td>
<td>-4.9</td>
<td>-31.1</td>
<td>7.8</td>
</tr>
<tr>
<td>Face area ((F_5))</td>
<td>-5.9</td>
<td>-43.8</td>
<td>14.2</td>
</tr>
<tr>
<td>Face motion ((F_6))</td>
<td>0.9</td>
<td>-60.2</td>
<td>112.7</td>
</tr>
<tr>
<td>Facial COM ((F_7))</td>
<td>-3.6</td>
<td>-42.1</td>
<td>23.1</td>
</tr>
</tbody>
</table>

Mouth and eye related features, i.e., \(F_1\) to \(F_4\), presented, on average, a decrease from \(H_0\) to \(H_1\) in all three games. However, the decrease does not apply to all the subjects, since at least one presented an increase from \(H_0\) to \(H_1\), as demonstrated by the positive values in the Max column of Tables 8.10, 8.11, and 8.12. Comparatively, the mean, minimum and maximum change of mouth \((F_1, F_2)\) and eye \((F_3, F_4)\) related features is similar in the three games. Consequentially, it is possible that features \(F_1\) to \(F_4\) are not affected by the game mechanics, however they do differ on a subject basis. On the other hand, features related to the whole face, i.e., \(F_5\) to \(F_7\), seem to be affected by game mechanics. Both \(F_5\) and \(F_7\) presented, on average, a decrease in the three games. In contrast, \(F_6\) presented, on average, an increase in the Mushroom and the Platformer game. A disproportional mean increase of 47.2% from \(H_0\) to \(H_1\) for feature \(F_6\) in the Mushroom game compared to the Platformer (0.9% increase) and Tetris (11.3% decrease) game, suggests that the feature is highly influenced by the mechanic of the Mushroom game. In this game, subjects are likely to move the head to facilitate saccadic eye movements used to inspect the cards. As the difficulty of the game increases, the number of cards to be inspected on the screen also increases, which could potentially lead to more (periodic) head movements as the game progresses to its stressful part.

Finally, all the features presented changes from periods \(H_0\) to \(H_1\) whose SD is considerably greater than the mean value, as shown in Table 8.9. The considerable heterogeneous variation of features, as demonstrated in the Min and Max columns of Tables 8.10, 8.11, and 8.12, supports the claim that the differences of features between the periods are spread out rather than clustered around the mean. Even though further analysis is required, the high SD and the broad interval of percentage change of all the features in the
Table 8.12: Percentage of change of features from period $H_0$ to $H_1$ in the Tetris game ($N = 20$).

<table>
<thead>
<tr>
<th>Feature (notation)</th>
<th>Mean</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouth outer ($F_1$)</td>
<td>-1.5</td>
<td>-27.8</td>
<td>39.0</td>
</tr>
<tr>
<td>Mouth corner ($F_2$)</td>
<td>-2.1</td>
<td>-26.5</td>
<td>26.9</td>
</tr>
<tr>
<td>Eye area ($F_3$)</td>
<td>-2.6</td>
<td>-19.0</td>
<td>26.1</td>
</tr>
<tr>
<td>Eyebrow activity ($F_4$)</td>
<td>-3.3</td>
<td>-16.2</td>
<td>21.1</td>
</tr>
<tr>
<td>Face area ($F_5$)</td>
<td>-1.4</td>
<td>-24.3</td>
<td>26.7</td>
</tr>
<tr>
<td>Face motion ($F_6$)</td>
<td>-11.3</td>
<td>-85.8</td>
<td>114.3</td>
</tr>
<tr>
<td>Facial COM ($F_7$)</td>
<td>-2.7</td>
<td>-24.7</td>
<td>21.8</td>
</tr>
</tbody>
</table>

three games, showing a decrease of 76.9% and increase of 8.2% for the same feature in the same game, for instance, highlight the between-subjects’ behavioral differences. A possible interpretation is that a more user-tailored, as opposed to a group-oriented, use of our facial features is more likely to portray such subject-based differences in a context involving emotional detection and games.

8.8.5 CONCLUSIONS

The method has been applied to the video recordings of an experiment involving games as emotion elicitation sources, which were deliberately designed to cause emotional states of boredom and stress. The results show statistically significant differences in the values of facial features detected during boring and stressful periods of gameplay for the following features: mouth outer ($F_1$), mouth corner ($F_2$), eye area ($F_3$), eyebrow activity ($F_4$), and face area ($F_5$). The face motion ($F_6$) and facial COM ($F_7$) features presented variations that were not statistically significant.

The results support the claim that the proposed method for the automated analysis of facial cues can potentially be used to differentiate the emotional states of boredom and stress in players. The utilization of such a method is unobtrusive and video-based, which eliminates the need to attach physical sensors to subjects.

8.9 STUDY 5: REMOTE DETECTION OF EMOTIONS

This study presents information regarding the use of machine learning to remotely detect the emotional state of subjects playing a game. The literature review presented in chapters 4, 5, and 7 indicates that a model based on several user signals, which is a multifactorial analysis, is more efficient for emotion detection. The mentioned chapters also highlight which of those signals can be remotely acquired within the context of this research via computer vision techniques.

The majority of the previous work found in the literature mention the use of machine learning techniques to model user signals into emotional states (Moghimi, R. Stone, and Rotshtein, 2017). Different models and accuracy results are mentioned, which depend on several particularities of the approach used by the authors. Based on the literature, a neural network has been selected for this study as a promising machine learning model.
This study is a systematic evaluation of the feasibility of a user-tailored neural network trained on data samples from two calibration games of a given subject which is then used to classify samples from a third calibration game of that same subject. The following sections describe how the study was conducted and analyzed, and present the results obtained. Finally, a discussion of the results and a conclusion are provided.

8.9.1 ANALYSIS AND METHODS

Data pre-processing
The pre-processing of video recordings involved the extraction of the parts containing the interaction with the games and the discarding of noisy frames. The process is very similar to the one detailed in Section 8.8.2 (page 88). Firstly, the periods in which the subjects played each of the available games were extracted from the video recordings. This resulted in three videos per subject, denoted as $V_{s,i}$, where $s$ is the $s$-th subject and $i \in \{1, 2, 3\}$ represents the game. Then the initial $D = 45$ seconds of any given video $V_{s,i}$ were ignored. The remainder of the video was then divided into three segments, from which the first and the last were selected as $H_0$ and $H_1$, respectively.

The pre-processing of the recordings resulted in 6 video segments per subjects: 3 segments $H_0$ (one per game) and 3 segments $H_1$ (one per game). A given game $i$ contains $N = 20$ pairs of $H_0$ and $H_1$ video segments (20 segments $H_0$, one per subject, and 20 segments $H_1$, one per subject). Regarding all the subjects and games, there are $N = 60$ pairs of $H_0$ and $H_1$ video segments (3 games $\times$ 20 subjects, resulting in 60 segments $H_0$ and 60 segments $H_1$). Subject 9 had problems playing the Platformer game, therefore, all segments $H_0$ and $H_1$ from that subject were discarded. Consequently, there are $N = 57$ pairs of $H_0$ and $H_1$ video segments in total, after the pre-processing.

The emotion elicitation design of the calibration games, where $H_0$ and $H_1$ represent boring and stressful interactions, respectively, is used to label samples to train and test the model. Samples from the $H_0$ part were labeled as boredom and samples from the $H_1$ part were labeled as stress. This process accounts for the informed levels of boredom and stress of the subject, which aimed to ensure a correct labeling of the samples, based on video segments that more likely, accurately reflect the emotional state self-reported by the subjects.

Classification features
The classification efficiency of a machine learning model is related to the number of features able to accurately discriminate the elements being classified. The use of more features does not necessarily produce a better model (James et al., 2013, Chapter 6). Some features might not accurately contribute to the classification, which leads to degradation of the results, if they are included.

The features and their classification potential are highly dependent on the type of data being used. In the present study, the set of features used for classification was extracted and selected on the basis of previous reports regarding the potential of said features to differentiate emotional states in games. In total, 9 features, denoted $F_1$ to $F_9$, were available for use: $F_1$ to $F_7$ are related to facial activity, while $F_8$ and $F_9$ are related to HR activity, including remote estimations (rPPG). Table 8.13 presents a description of all the features.

Features $F_1$ to $F_7$ are based on automatically detected facial landmarks related to facial...
Table 8.13: Description of features used for classification

<table>
<thead>
<tr>
<th>Notation</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>Mouth outer</td>
<td>Monitor the zygomatic muscle.</td>
</tr>
<tr>
<td>$F_2$</td>
<td>Mouth corner</td>
<td>Monitor the zygomatic muscle.</td>
</tr>
<tr>
<td>$F_3$</td>
<td>Eye area</td>
<td>Monitor the orbicularis oculi muscle, e.g. blinking.</td>
</tr>
<tr>
<td>$F_4$</td>
<td>Eyebrow activity</td>
<td>Monitor the corrugator muscle.</td>
</tr>
<tr>
<td>$F_5$</td>
<td>Face area</td>
<td>Monitor facial movement to and away from the camera</td>
</tr>
<tr>
<td>$F_6$</td>
<td>Face motion</td>
<td>Describe the total distance the head has moved in any direction in a short period of time.</td>
</tr>
<tr>
<td>$F_7$</td>
<td>Facial COM</td>
<td>Describe the overall movement of all facial landmarks.</td>
</tr>
<tr>
<td>$F_8$</td>
<td>Remote HR</td>
<td>HR estimated using the rPPG technique proposed by Poh, McDuff, and Picard (2011).</td>
</tr>
<tr>
<td>$F_9$</td>
<td>Ground HR</td>
<td>HR calculated from a physical sensor, i.e. watch.</td>
</tr>
</tbody>
</table>

elements that express a connection with emotional states. As previously mentioned in Chapter 4, there is evidence of more frequent corrugator activity when positive game events occur (Hazlett, 2006), as well as increased activity of the zygomatic muscle associated with self-reported positive emotions (Tijs, Brokken, and IJsselsteijn, 2008). Positive and rewarding game events are also connected to an increase in zygomatic and orbicularis oculi activity (Ravaja et al., 2006). The detection of stress is also related to blinking rate (Giannakakis et al., 2017; Dinges et al., 2005), lip movement (Dinges et al., 2005) and lips deformation (Metaxas, Venkataraman, and Vogler, 2004; Giannakakis et al., 2017), mouth activity (Liao et al., 2005), and head movement/velocity (Giannakakis et al., 2017).

Feature $F_8$ is based on remote estimations of HR performed using the rPPG technique proposed by Poh, McDuff, and Picard (2011). Similarly, feature $F_9$ is based on the HR readings provided by a physical sensor, i.e. watch, used by the subjects. As mentioned in Chapter 5, HR and its derivatives, such as HRV, have been used as reliable sources of information in different emotion estimation methods (Kukolja et al., 2014). Reports in the literature show the use of HR and its derivatives for continuous arousal monitoring (Grundlehner et al., 2009), the measurement of confusion (Xiao and J. Wang, 2015), the triangulation of psychophysiological emotional reactions to digital media stimuli (Nogueira et al., 2015), the detection of mental and physical stress (Vandeput et al., 2009; Garde et al., 2002), and the measurement of frustration (Rodriguez et al., 2015).

Features extraction and calculation

The process of extracting and calculating features is achieved using a moving window applied to the videos of all the subjects. The moving window has a size of 15 seconds and a step of 1 second (93.33% overlap). For each window in the video, computer vision techniques are applied to all the frames within that window, to detect facial landmarks and
collect information regarding pixel values, e.g. the mean value of pixels in the blue channel. The detected landmarks are used to calculate the features related to facial activity, while the pixel values are used to estimate the HR.

Features $F_1$ to $F_7$, which represent facial activity, are mostly calculated using the Euclidian distance of automatically detected facial landmarks for each frame. A detailed description of the process is presented in Section 8.8 (page 82). Features $F_8$ and $F_9$, which represent HR activity, are calculated on the basis of rPPG estimations of HR and HR measurements of a physical sensor, respectively. A detailed description of the rPPG estimation process is presented in Section 8.7 (page 73).

Even though all the frames within the window are analyzed, only a single, final value is assigned to each feature per window. For features $F_1$ to $F_7$, the final value of a given feature is calculated by aggregating the values of all the frames within the window of that given feature by using mean or standard deviation. Henceforth, $F_i^\text{m}$ and $F_i^\text{s}$ are used to denote feature $F_i$ whose values in a window were aggregated using the mean and standard deviation, respectively. Empirical tests conducted for this study have shown that features connected to facial regions with fast changes within the window, e.g. eye area and face motion, are better represented by an aggregation using the standard deviation. However, facial features with slower changes, e.g. face area and mouth activity, are better represented by an aggregation using the mean. Feature $F_8$ does not require any aggregation of values, since all the frames within the window are used to produce a single value, i.e. the estimation of the mean HR in that window. Finally, feature $F_9$ is aggregated using the mean of all the HR values provided by the physical sensor within the window, i.e. the mean HR within the window.

Training and evaluation of an emotion classifier

The classification procedure uses the previously mentioned feature set and a neural network trained to identify two emotional states: boredom and stress. Both the training and evaluation of the neural network are performed in a user tailored fashion: data from a given subject $S_i$ is used to train and evaluate the emotion classification of that given subject $S_i$. Figure 8.12 illustrates the process.

![Figure 8.12: Iteration of a 3-fold Leave-One-Session-Out Cross-Validation performed on the gaming session of a given subject with 3 games, i.e. A, B and C. Data of two calibration games, e.g. A and B, are used to train the machine learning model, while data of the third calibration game, e.g. C, are used to evaluate the model.](image)

Leave-One-Session-Out Cross-Validation (LOSOCV) is used to evaluate each trained user-tailored model, as illustrated in Figure 8.12. In LOSOCV, the data from one session instance are left out and a model is constructed on the data from all other session instances. In the present study, a given subject $S_i$ played 3 calibration games, e.g. A, B and C, thus, the data from one calibration game are left out and a model is trained on the data of the other two calibration games for that subject $S_i$. This is repeated for
all three calibration games of that subject $S_i$. Consequently, the use of LOSOCV will produce 3 models per subject, resulting in 3 measurements of classification accuracy per subject, denoted $L_j$, where $j \in \{1, 2, 3\}$ represent each evaluated model. The final classification accuracy for given subject $S_i$, named $A_i$, is calculated as the mean of $L_j$ values obtained from the iterations in the LOSOCV. In other words, each subject contributes a single classification accuracy value $A_i$, which is calculated on the basis of the mean classification accuracy of the subject’s three models in the LOSOCV iterations.

In the training process of each model, which is performed 3 times per user, the hyper-parameters of each neural network, e.g. number of neurons, are optimized using random search (Bergstra and Bengio, 2012). A 10-fold cross validation method repeated 3 times is applied, which divides the dataset into 10-subsets, each of which is left out while the model is trained on all the others. The process is repeated 3 times and the final metric for the model is the mean from the number of repeats. The area under the ROC curve (AUC) is used as a metric to select the best model.

According to previous analysis, the subjects perceived the games as boring at the beginning and stressful at the end. As a consequence, it is assumed that the subject’s emotional state in $H_0$ and $H_1$ is boredom and stress, respectively. Based on this assumption, training and evaluation data obtained from the video segments in $H_0$ and $H_1$ were labeled as boredom and stress, respectively.

Analysis

In order to test the effectiveness of the neural network in classifying samples as either boredom or stress, all the trained neural networks were evaluated in conjunction. As described in the previous section, each subject’s model was evaluated using LOSOCV, which produced a classification accuracy $A_i$ for any given subject $S_i$. The minimum, maximum and mean value of $A_i$ was calculated as a metric for accuracy. In order to better contextualize the classification results, the same process was also applied to the other three metrics obtained during the LOSOCV evaluation: Precision, Recall and F1 score. Precision accounts for the correctly predicted positive observations of the total predicted positive observations, e.g. of all the samples classified as stress, how many were indeed labeled as stress. Recall accounts for the correctly predicted positive observations of all the available observations in a class, e.g. of all the available samples labeled as boredom (or stress), how many were actually classified as such. Finally, the F1 score is the weighted average of Precision and Recall.

In order to better understand the contribution of each feature for the classification process, the training/evaluation process mentioned previously was also performed using different feature sets. Each of these different feature sets was evaluated in an independent test, denoted $T_i$. Table 8.14 shows tests $T_i$ and the corresponding feature sets used in the process.

Tests MULTI_R and MULTI_G use a multifactorial set of features for their neural network, where facial and HR information is used in combination. The difference between MULTI_R and MULTI_G is that the former uses rPPG estimated HR, while the latter uses the HR obtained from the physical sensor. Test FACE uses a set of features based solely on facial information. Finally, tests HR_R and HR_G use only HR information as a feature. Similar to MULTI_R and MULTI_G, tests HR_R and HR_G use HR readings from rPPG estimations and a physical sensor, respectively.

Since the subjects perceived the games as boring and stressful, then this difference should enable a trained neural network to properly classify evaluation samples as either bore-

\[^2\text{F1 score should not be confused with } F_1, \text{ the mouth outer facial feature used in the model.}\]
Table 8.14: Tests and their respective feature sets

<table>
<thead>
<tr>
<th>$T_i$</th>
<th>Name</th>
<th>Feature set</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MULTI_R</td>
<td>$F_1^\mu, F_2^\mu, F_3^<em>, F_4^</em>, F_5^\mu, F_6^*, F_8$</td>
<td>Facial analysis, rPPG-estimated HR.</td>
</tr>
<tr>
<td>2</td>
<td>MULTI_G</td>
<td>$F_1^\mu, F_2^\mu, F_3^<em>, F_4^</em>, F_5^\mu, F_6^*, F_9^\mu$</td>
<td>Facial analysis, HR from physical sensor.</td>
</tr>
<tr>
<td>3</td>
<td>FACE</td>
<td>$F_1^\mu, F_2^\mu, F_3^<em>, F_4^</em>, F_5^\mu, F_6^*$</td>
<td>Facial analysis only.</td>
</tr>
<tr>
<td>4</td>
<td>HR_R</td>
<td>$F_8$, $F_9^\mu$</td>
<td>rPPG-estimated HR only.</td>
</tr>
<tr>
<td>5</td>
<td>HR_G</td>
<td>$F_9^\mu$</td>
<td>HR from physical sensor only.</td>
</tr>
</tbody>
</table>

dom or stress. Additionally, the use of a multifactorial approach, where facial analysis and HR information are used in combination instead of either one alone, is expected to produce better classification results (Zacharatos, Gatzoulis, and Chrysanthou, 2014). Based on those expectations, the following hypotheses state:

- $u_1$: a user-tailored neural network trained on the data samples from two calibration games of a given subject $S_i$ is able to classify samples from a third calibration game of that same subject $S_i$ with an accuracy greater than chance-level rate (random guessing);
- $u_2$: a user-tailored neural network using a multifactorial feature set, i.e. facial and HR features, performs with greater accuracy than a user-tailored neural network using facial features only;
- $u_3$: a user-tailored neural network using a multifactorial feature set, i.e. facial and HR features, performs with greater accuracy than a user-tailored neural network using HR features only.

Hypothesis $u_1$ was tested by checking whether the mean value of the classification accuracy, i.e. calculated from all $A_i$ values, is greater than $0.5$. In such a case, it is assumed that an accuracy rate of $0.5$ ($50\%$) is the theoretical probabilistic chance level achieved by a totally random classification performed by a classifier evaluated on an infinite number of data samples. Hypotheses $u_2$ and $u_3$ were tested by performing a Wilcoxon Signed Ranks test on all $J_i$ values of the two competing tests $T_i$. As previously mentioned, the use of LOSOCV produces 57 accuracy measurements $J_i$ per test $T_i$.

8.9.2 RESULTS

Table 8.15 presents the mean values of the resulting classification metrics for accuracy, precision, recall and F1 score, calculated and analyzed according to the procedures described in Section 8.9.1. Regarding the accuracy metric, the highest mean value achieved was $62.3\%$ in test MULTI_G, whose model used a combination of facial and HR features. The HR feature in that case was calculated from the physical sensor, not remotely estimated. The second and third highest accuracy rates were $60.8\%$ in test HR_G (HR from physical sensor only) and $60.4\%$ in test MULTI_R (facial and remotely-estimated HR features), respectively. The highest values achieved for precision, recall and F1 score were $65.6\%$, $62.4\%$, and $58.1\%$, respectively, all in test HR_G.
Table 8.15: Mean values of resulting classification metrics

<table>
<thead>
<tr>
<th>Test</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MULTI_R</td>
<td>0.604</td>
<td>0.612</td>
<td>0.599</td>
<td>0.521</td>
</tr>
<tr>
<td>MULTI_G</td>
<td>0.623</td>
<td>0.583</td>
<td>0.607</td>
<td>0.514</td>
</tr>
<tr>
<td>FACE</td>
<td>0.594</td>
<td>0.601</td>
<td>0.585</td>
<td>0.507</td>
</tr>
<tr>
<td>HR_R</td>
<td>0.547</td>
<td>0.541</td>
<td>0.545</td>
<td>0.497</td>
</tr>
<tr>
<td>HR_G</td>
<td>0.608</td>
<td>0.656</td>
<td>0.624</td>
<td>0.581</td>
</tr>
</tbody>
</table>

Table 8.16 presents the minimum and maximum mean values for the resulting classification metrics. At least one subject in test MULTI_G has been classified with a mean accuracy of 98%, the highest value for that metric in all the tests. The worst mean accuracy value was 19% for at least one subject in test MULTI_R. Regarding precision, the highest mean value was 97% in test MULTI_G. In all tests but HR_G, at least one subject has been classified with zero precision (all the samples were classified wrongly). Regarding precision, the highest and lowest mean values were 98% and 12% in tests MULTI_G and FACE, respectively. Finally, regarding the F1 score, the highest mean value was 98% for at least one subject in test MULTI_G. All the tests presented zero as the lowest F1 score.

Table 8.16: Minimum and maximum mean values of resulting classification metrics

<table>
<thead>
<tr>
<th>Test</th>
<th>Accuracy min</th>
<th>Accuracy max</th>
<th>Precision min</th>
<th>Precision max</th>
<th>Recall min</th>
<th>Recall max</th>
<th>F1 min</th>
<th>F1 max</th>
</tr>
</thead>
<tbody>
<tr>
<td>MULTI_R</td>
<td>0.19</td>
<td>0.91</td>
<td>0.00</td>
<td>0.95</td>
<td>0.19</td>
<td>0.87</td>
<td>0.00</td>
<td>0.91</td>
</tr>
<tr>
<td>MULTI_G</td>
<td>0.25</td>
<td>0.98</td>
<td>0.00</td>
<td>0.97</td>
<td>0.13</td>
<td>0.98</td>
<td>0.00</td>
<td>0.98</td>
</tr>
<tr>
<td>FACE</td>
<td>0.26</td>
<td>0.90</td>
<td>0.00</td>
<td>0.95</td>
<td>0.12</td>
<td>0.90</td>
<td>0.00</td>
<td>0.89</td>
</tr>
<tr>
<td>HR_R</td>
<td>0.36</td>
<td>0.72</td>
<td>0.00</td>
<td>0.79</td>
<td>0.18</td>
<td>0.77</td>
<td>0.00</td>
<td>0.67</td>
</tr>
<tr>
<td>HR_G</td>
<td>0.38</td>
<td>0.82</td>
<td>0.26</td>
<td>0.85</td>
<td>0.23</td>
<td>0.87</td>
<td>0.00</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Finally, there are indications that a multifactorial model, which uses a combination of facial and HR features, performs with greater accuracy than a model that uses either facial or HR features. A Wilcoxon Signed Ranks test indicates that the mean accuracy was greater for a multifactorial model that uses facial and remotely estimated HR features, i.e. MULTI_R, than for a model that uses only remotely estimated HR, i.e. HR_R, \( Z = -2.00, p = 0.044, r = 0.26 \). However, there are no indications that the mean accuracy of such a multifactorial model, i.e. MULTI_R, is statistically significantly greater than the mean accuracy of a model that uses only facial features, i.e. FACE, \( Z = -1.10, p = 0.267, r = 0.14 \).

8.9.3 DISCUSSION

The results indicate that the application of a user-tailored model to remotely estimate the emotional state of players from videos of gaming sessions is feasible. Previously mentioned hypothesis \( u_1 \) states that a user-tailored neural network trained on data samples from two calibration games of a given subject \( S_i \) is able to classify samples from a third calibration game of that same subject \( S_i \) with an accuracy greater than chance-level rate (random guessing). Such a model was tested in two configurations: MULTI_R and
MULTI_G. Model MULTI_R, which uses a multifactorial feature set composed of facial and remotely estimated HR information, presented a mean classification accuracy of 60.4%. This reported accuracy is greater than 50%, the theoretical probabilistic chance level, which confirms hypothesis $u_1$. Model MULTI_G, which also uses a multifactorial feature set but differs in the acquisition of HR data, i.e. physical sensor instead of remote estimation, presented a mean classification accuracy of 62.3%. This reported accuracy is also greater than the theoretical probabilistic chance level. The slightly greater classification accuracy of model MULTI_G compared to MULTI_R suggests that more precise rPPG estimations of the HR could improve the overall classification accuracy of model MULTI_R. A Wilcoxon Signed Ranks test, however, has no statistically significant indication that the accuracy was greater for MULTI_G than for MULTI_R, $Z = -1.76$, $p = 0.078$, $r = 0.23$. Despite not being statistically significant, values $p = 0.078$ and $r = 0.23$ (small effect according to Cohen’s classification of effect size) suggest a trend towards that reasoning.

Hypothesis $u_2$ states that a user-tailored neural network using a multifactorial feature set performs with greater accuracy than a user-tailored neural network using facial features only. The Wilcoxon Signed Ranks test, mentioned in the previous section, presented no statistically significant indications that the classification accuracy of MULTI_R is greater than FACE. It is possible to speculate that a set of facial features, e.g. eyebrow and mouth analysis, has a greater potential to differentiate emotional states in a classifier. In particular, it might perform better than a classifier based on rPPG-estimated HR alone.

Regarding hypothesis $u_3$, which states that a user-tailored neural network using a multifactorial feature set performs with greater accuracy than a user-tailored neural network using HR features only. The Wilcoxon Signed Ranks test, mentioned in the previous section, presents statistically significant indications that the classification accuracy of MULTI_R is greater than HR_R. It supports the claim of hypothesis $u_3$, confirming that a multifactorial model performs better than one based solely on remotely estimated HR data. The lower classification potential of remotely estimated HR, however, could be attributed to errors in the rPPG estimation process caused by noise, e.g. natural movement of subjects (see Section 8.7, page 73, for more information). As a consequence, a more precise HR estimation used in a multifactorial model allegedly contributes to producing a better classifier. A Wilcoxon Signed Ranks test confirms with statistical significance that the classification accuracy of MULTI_G, i.e. facial and HR from sensor, is greater than the accuracy of FACE, i.e. facial information only, $Z = -2.12$, $p = 0.033$, $r = 0.28$. It supports the previously mentioned idea that a precise HR estimation (from a physical sensor in the case of MULTI_G) combined with facial information is a better classifier than one using facial information alone, i.e. model FACE. Finally a Wilcoxon Signed Ranks test does not indicate that the accuracy of MULTI_G is greater than HR_G, $Z = -1.08$, $p = 0.278$, $r = 0.14$. Consequently, it seems that precise estimations of HR are important, but HR or facial information used separately is likely to be less important for classification than their joint, multifactorial use in the emotion classification process.

Finally, it is important to highlight the possible limitations of using the theoretical probabilistic chance level rate of 50% for the evaluation of the model’s accuracy. As previously mentioned, an accuracy baseline of 50% assumes a totally random classification evaluated on an infinite number of data samples. One could argue that such a scenario is unrealistic. In that case, the true evaluation of a model must consider statistical significance levels that take into account the sample size used in the process. Combisson and Jerbi (2015) present a work in the field of brain signal classification that uses analytical and empirical solutions, i.e. binomial formula and permutation tests, to show the influence of small numbers of data samples in the theoretical probabilistic chance level. According to the authors, a minimal correct classification rate of 62.5% is required to
assert statistical significance, i.e. $p < 0.05$, during the classification of two classes with a sample size of 40. In the present study, each subject was evaluated with 38 samples on average. As presented in the results of this study, the mean accuracy rate of models $\text{MULTI}_R$ and $\text{MULTI}_G$ are 60.4% and 62.3%, respectively. Following the analysis of Combrisson and Jerbi (2015), the mean accuracy rate of these multifactorial models would not be enough to assert statistical significance of the results. Nevertheless, both models presented a mean accuracy that trends towards the minimal 62.5%. Additionally, the previously mentioned Wilcoxon Signed Ranks test does not indicate that the accuracy was greater for $\text{MULTI}_G$ than for $\text{MULTI}_R$, so their performance could be the same. It is also important to stress the use of Leave One Session Out Cross Validation in the evaluation of each subject. It uses a completely independent and different game as the sampling source for the evaluation of each model, which strengthens the evaluation process. The reduced number of subjects in the study, i.e. 19, as well as samples used in the evaluation of each model, i.e. 38 on average, are in fact limiting factors. The reported results, however, provide insights into the feasibility of a multifactorial remote approach for emotion classification. The accuracy of the models presented in this study is not statistically confirmed without the assumption of a baseline produced by a random classifier evaluated on an infinite number of data samples. However, there is in fact a trend indicating that the use of a user-tailored model to remotely estimate the emotional states of players is worth further investigation.

8.9.4 CONCLUSIONS

This study presents a systematic evaluation of the feasibility of using a user-tailored neural network trained on data samples from two calibration games of a given subject to classify emotional states from a third calibration game of that same subject. Further investigation is required, however, the results suggest that a user-tailored neural network, based on remotely acquired data from video recordings, is able to classify emotional states with an accuracy greater than chance-level rate (random guessing).

Regarding the efficiency of a multifactorial model, which uses facial and HR information in combination instead of separately, there are no statistically significant indications that the classification accuracy of such a model is greater than a model using facial information alone. However, a multifactorial model based on remotely acquired data performs better than one based solely on remotely estimated HR data. Finally, it seems that precise estimations of HR are important, but HR or facial information used separately is likely to be less important for classification than their combined use in a multifactorial emotion classification model based on remotely acquired data. The analysis performed in this study supports further research on a user-tailored model to remotely estimate the emotional states of players.

\footnote{This number refers to the amount of samples used for the evaluation of a model, not its training. During the training phase of each model, more than 38 samples were used per subject.}
Chapter 9

Experiment 2: Validation of Remote Detection of Emotions

The experiment described in this chapter, the second conducted, aimed at validating the proposed approach for the remote detection of emotions, described in the objectives of this thesis (Section 1.3, page 7). The approach uses remotely acquired signals, namely, heart rate (HR) and facial actions (FA), to create a user-tailored model, i.e. trained neural network, that is able to detect the emotional states of boredom and stress of a given subject. The approach is composed of two phases: training (or calibration) and testing. In the training phase, the model is trained by applying a user-tailored approach, i.e. data from subject $S_a$ playing 3 calibration games (Mushroom, Platformer and Tetris) are used to create model $N_a$. The result of the training phase is a user-tailored model, i.e. model $N_a$, a trained neural network for use on subject $S_a$. The testing phase occurs in a game session involving subject $S_a$ playing any ordinary, non-calibration game, e.g. COTS game. During the testing phase, the signals of subject $S_a$ are remotely acquired and fed into the previously trained model $N_a$, which outputs the estimated emotional state of subject $S_a$ for that particular testing game.

In summary, the aim of this experiment was to answer the following research question:

"How accurate is an emotion detection approach that uses remotely acquired signals, i.e. heart rate and facial actions, as input of a machine learning model, i.e. neural network, that is trained on a user-tailored basis (one subject produces one model) using calibration games as emotion elicitation?"

The overall goal of this experiment is to validate the proposed approach and prove its feasibility by analyzing the emotion classification accuracy during the testing phase. The following sections present a detailed explanation of the experiment, including its participants, structure and results, as well as a discussion and a conclusion.

9.1 Participants

Sixty two ($N = 62$) adult participants of both genders (38.7% female, 61.3% male) with different ages (19 to 66, mean 27.2, SD 7.2) and different gaming experience gave their informed and written consent to participate in the experiment. The study population consisted of staff members and students of the University of Skövde, as well as inhabitants of the community/city.

1Participants of this experiment (Experiment 2) and those of Experiment 1 (detailed in Chapter 8, page 59) are different. There is no overlap of subjects in both experiments.
In response to the question regarding the participants’ level of skill playing video games, 6 (9.7%) reported no skill, 19 (30.6%) reported not very skilled, 25 (40.3%) reported moderately skilled and 12 (19.9%) reported very skilled. In response to the question regarding the number of hours per week playing any type of video game over the last year, 25 (40.3%) reported more than 10, 7 (11.3%) reported 5 to 10, 6 (9.7%) reported 3 to 4, 5 (8.1%) reported 1 to 3, 10 (16.1%) reported 0 to 1, and 9 (14.5%) reported no activity.

These numbers indicate that the sample population has a diversity of ages, gaming experience and playing frequency. Such diversity provides heterogeneous data that allow a more realistic and broad analysis of the proposed method.

9.2 METHOD

The following sections present the experiment structure and the methods employed to collect and analyze data.

9.2.1 EXPERIMENTAL DESIGN AND SETUP

Subjects were seated alone in the room, in front a computer, while recorded by a camera and measured by a heart rate sensor, as illustrated in Figure 9.1. The camera was attached to a tripod and placed in front of the subjects at a distance of approximately 0.6m; the camera was tilted slightly up. A spotlight, tilted 45° up, placed at a distance of 1.6m from the subject and 45cm higher than the camera level, was used for illumination; no other light source was active during the experiment.

![Figure 9.1: Experiment setup. (a) Position of equipment, showing computer, camera, and external light source. (b) Highlight of the video camera and its angle.](image)

Each participant was recorded for an average of 45 minutes during two (uninterrupted) parts of the experiment, i.e. the calibration and testing phase, as illustrated by Figure 9.2. In the calibration part, aimed at gathering data for training a user-tailored model, the subjects played 3 calibration games (described in Section 9.2.2). Each of these games was followed by a questionnaire related to the game and the stress/boredom levels. The games were also followed by a rest period of 138 seconds during which the subjects listened to calm classic music. In the testing part, aimed at gathering data to test the ac-
curacy of the user-tailored model, the subjects played 7 levels of an evaluation game, i.e. Infinite Mario (described in Section 9.2.3). In the calibration phase of the experiment, the order in which the games were played was randomized among the subjects.

In the testing part, the subjects played 3 batches of Infinite Mario levels: batches A, B and C containing 3, 3 and 1 level each, respectively. The levels in batches A and B were designed to present an increase in difficulty within the batch; therefore, levels $A_1$ and $B_1$ were expected to be less difficult/challenging than levels $A_3$ and $B_3$, for instance. Similarly, the levels in batch B were designed to be more challenging than the levels in batch A, also following an increase in difficulty. Consequently, levels $B_i$ are expected to be slightly more difficult/challenging than levels $A_i$. This pattern intended to mimic the balance curve of a commercial game, where levels, and game parts, commonly tend to increase their difficulty as the game story progresses. In order to ensure that the subjects would experience some level of boredom during the testing phase, which is required for the evaluation of the proposed method, levels $B_1$ and $C_1$ were designed using a particular set of changes, including the use of Mario’s auto-scrolling camera mechanics. In such a configuration, the player has no control of the speed of the level. After each level was played, the subjects were required to answer a questionnaire about how boring/stressful the game level had been. The order in which the levels were played was not randomized among the subjects during the testing phase of the experiment. As a consequence, all the subjects played the evaluation game in the same order: levels $A_1$ to $A_3$, then by levels $B_1$ to $B_3$, finally level $C_1$.

After the subjects finished playing the last level in the testing part, i.e. level $C_1$, they answered a final questionnaire about their age and gaming experience/profile. Before starting the experiment, the participants received the following instructions from a researcher: they should play a few games, answer questionnaires after each game and rest when instructed; they were told that their gaming performance was not being evaluated, that they should not give up in the middle of the games, that a time limit exists for some levels to prevent them from playing too long, and that they should remain seated during the whole process.

9.2.2 CALIBRATION GAMES

The three calibration games, i.e. Mushroom, Platformer and Tetris, have already been described in detail in Section 8.4 (page 61). The present experiment used exactly the same calibration games as in the first experiment (see Chapter 8). The games are 2D, casual-themed, and played with mouse or keyboard in a web browser. They were carefully designed to provoke boredom at the beginning and stress at the end, with a linear progression between the two states.
9.2.3 EVALUATION GAME

The game used in the evaluation phase of the experiment is a modified version of Markus Persson’s Infinite Mario, a public domain clone of Nintendo’s platform game *Super Mario Bros*. In the case of this experiment, the game is played with a keyboard in a web browser. Infinite Mario has been widely mentioned in the literature, including studies involving the modeling of player experience (Pedersen, Togelius, and Yannakakis, 2009; Pedersen, Togelius, and Yannakakis, 2010; Shaker, Asteriadis, et al., 2011) and detection of affective states (Shaker, Yannakakis, and Togelius, 2011).

![Figure 9.3: Screenshots from Infinite Mario. From left to right, level types Over-ground, Underground and Castle, respectively.](image)

The gameplay in Super Mario, consequently in Infinite Mario as well, consists of controlling the main character, Mario, along the level. Mario can move left or right, jump, run, duck, and throw fireballs (if the power-up Flower has been collected). The objective of the game is to complete each level, which is accomplished by traversing it from left to right until the “end of level” checkpoint is reached. Mario can be in three different states: small, big, and power-up. If Mario is small, any interaction with enemies that is different from landing on top of them after a jump results in Mario getting killed immediately. If Mario is big, the same “wrong” interaction with enemies causes Mario to be hurt and transform into the small state. If Mario is in the power-up state, the “wrong” interaction with enemies causes Mario to be hurt and transform into the big state. Consequently, keeping Mario in the big or power-up state is a strategic advantage that prevents Mario from being killed, which is likely to calm players, i.e. relaxed emotional state. On the other hand, keeping Mario in the small state is less beneficial, since mistakes are fatal, thus likely causing players to feel anxious/stressed in such conditions.

Along the level, Mario might encounter enemies, which can be killed or ignored. Mario can kill enemies by jumping and landing on top of them, which is rewarded with score points. Some enemies, e.g. Koopa Troopa (a sort of turtle), leave a shell behind when killed by Mario. The shell can be picked up by Mario and carried around, serving as a weapon when released. The levels might also contain terrain obstacles of varying sizes, e.g. gaps, that must be jumped over. If Mario falls into a gap, he dies immediately. Mario can also find collectable items, i.e. coins and power-ups, or interactable items, e.g. blocks. Mario interacts with blocks by bumping into them from below, e.g. jumping

---

2The version of the game used in the experiment is a HTML5, web-based version built by Robert Kleffner, available at: https://github.com/robertkleffner/mariohtml5. Robert ported to HTML5 the original Java version created by Markus Persson. Source code of both versions, Robert’s and Markus’, are in the public domain. The source code of the final version used in this experiment is available at: https://fernandobevilacqua.com/link/phd-experiment2
and hitting Mario’s head on the bottom of a block destroys it. A destroyed block might give a collectable item as a reward, e.g. coin, *Mushroom* (Mario transitions to big state) or *Flower* (Mario transitions to power-up state).

During gameplay, information about Mario, the score and the current level is displayed at the top of the screen. This information includes the number of lives Mario has left to complete the level, the level score, number of coins collected (collecting 100 coins results in an extra life), name of the current level, and the amount of time available to complete the level (constantly ticking down). When the time remaining to complete the level reaches the 60 seconds mark, a hurry up sound is played, then the background music starts to play in a faster tempo. Unless informed otherwise, all the levels of Infinite Mario in the experiment start with 3 lives and 200 seconds of available time. Every time Mario dies, the time remaining to complete the level is reset to its initial value, e.g. 200 seconds.

Originally, Infinite Mario procedurally generates all its gameplay content, e.g. level design and position of items/enemies. This behavior was not desired for the experiment, since all the subjects should experience the same Mario levels. Additionally, subjects should feel stressed and bored in the game at some points, so that the proposed emotion detection method can be properly evaluated when such moments are detected. As a consequence, Infinite Mario was adapted and tweaked, thus made to fit as an ideal evaluation game in the experiment. The procedural content generation was constrained by a seed and a set of parameters was introduced to control the creation of the content, e.g. length of the level, amount and width of terrain obstacles, such as gaps and platforms, availability of coins and power-ups, among others. It ensured that all the subjects experienced exactly the same levels.

Previous works using Infinite Mario (Pedersen, Togelius, and Yannakakis, 2009; Pedersen, Togelius, and Yannakakis, 2010) have shown a correlation between anxiety and 1) difficulty of jumping, e.g. overcoming obstacles, and 2) gap width. There is also a correlation between boredom and the width of gaps, i.e. the wider the gap, the less boring the level. Based on those findings and the guidance provided by game design experts, the Mario levels used in the experiment were adjusted according to the description presented in Table 9.1. Column *Level* refers to the level name/number. Column *Type* refers to the overall visual representation of the level. Possible types are *Overground* (open sky and green landscape), *Underground* (closed ceiling, dirt-like environment), and *Castle* (closed ceiling with bricks resembling the interior of a castle). Each level type features different background music and visual elements, as illustrated in Figure 9.3. Column *Emotion* refers to the expected emotional state most subjects will experience. Finally, column *Adjustments* refer to the constraints used to generate the levels content.
Table 9.1: Levels of Infinite Mario and adjustments made to induce a particular emotional state.

<table>
<thead>
<tr>
<th>Level</th>
<th>Type</th>
<th>Emotion</th>
<th>Adjustments</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>Overground</td>
<td>Any</td>
<td>Reduced number of interactable/collectable items and terrain obstacles; no power-ups; only 2 enemies and 1 gap (with width of Mario himself); Mario starts in big state.</td>
</tr>
<tr>
<td>$A_2$</td>
<td>Underground</td>
<td>Any</td>
<td>Regular number of interactable/collectable items, terrain obstacles, power-ups and enemies. Mario starts in small state.</td>
</tr>
<tr>
<td>$A_3$</td>
<td>Castle</td>
<td>Stress</td>
<td>Several gaps (with varying widths); reduced number of interactable items; no collectables/power-ups; several enemies; reduced time to complete level. Mario remains in small state.</td>
</tr>
<tr>
<td>$B_1$</td>
<td>Overground</td>
<td>Boredom</td>
<td>Auto-scrolling camera; reduced number of interactable/collectable items; few terrain obstacles; no gaps, power-ups, or enemies. Mario remains in big state.</td>
</tr>
<tr>
<td>$B_2$</td>
<td>Underground</td>
<td>Any</td>
<td>Regular number of interactable/collectable items, terrain obstacles, power-ups and enemies. Mario starts in small state.</td>
</tr>
<tr>
<td>$B_3$</td>
<td>Castle</td>
<td>Stress</td>
<td>Several gaps (with varying widths); reduced number of interactable items; no collectables/power-ups; several enemies; reduced time to complete level. Mario remains in small state.</td>
</tr>
<tr>
<td>$C_1$</td>
<td>Overground</td>
<td>Boredom</td>
<td>Auto-scrolling camera; reduced number of interactable/collectable items; few terrain obstacles; no gaps, power-ups, or enemies. Mario remains in big state.</td>
</tr>
</tbody>
</table>
Level $A_1$ is an introduction to the game to familiarize the subjects with the mechanics, e.g. move, jump, collect items. Levels $A_2$ and $B_2$ were designed to be regular Mario levels with a compelling and enjoyable challenge scale.

Levels $A_3$ and $B_3$ were designed to be more stressful by including more enemies and several gaps which were wider than usual. The absence of power-ups, the number of challenges, i.e. enemies and wide gaps, and the fact that Mario is continuously in the small state should force the subjects to better time actions, e.g. jump, and constantly pay attention to the surroundings. These levels also use the Castle type, which is usually associated with “boss levels” in Super Mario (commonly more challenging). Finally, levels $A_3$ and $B_3$ have an available time of 80 seconds to be finished, a considerably lower value compared to 200 seconds in other levels. As a consequence, after 20 seconds of gameplay, the hurry up sound is played and the background music starts to play faster, a configuration that is likely to cause an emotional state of stress.

In contrast, levels $B_1$ and $C_1$ were designed to be more boring. These levels include an auto-scrolling camera mechanic, which enables the camera to automatically traverse the level independently of Mario’s movements. The speed of the auto-scrolling camera was adjusted to be at constant, but slow pace. Additionally, the reduced number of interactable/collectable items, the existence of only a few terrain obstacles, as well as the absence of gaps, power-ups and enemies are likely to cause an emotional state of boredom. Furthermore, levels $B_1$ and $C_1$ are very similar visually, which might cause subjects to perceive level $C_1$ as a repetition of level $B_1$. In that case, subjects might perceive level $C_1$ as even more boring, since the level topology is already known and the player is unable to move the camera at a faster pace.

As previously mentioned, the levels were adjusted and play-tested by game design experts. It ensured that the content of all levels and the constraints/modifications applied to them did not affect the subject’s perception of playing a clone of a Mario. For instance the order in which the levels were played, i.e. repeating the pattern of an overground, then an underground and finally a castle level, was kept as an important element. It should mimic the expected world progression of the original Mario game, where the final level of a particular world is usually a castle level with a boss. Finally particular attention was invested to make Infinite Mario levels difficulty as different as possible from the linear difficulty progression present in the three calibration games. The aim was to make Infinite Mario as similar to Super Mario as possible respecting the content constraints mentioned previously.

9.2.4 DATA COLLECTION

During the whole experiment, the subjects were recorded with a Canon Legria HF R606 video camera. All the videos were recorded in color (24-bit RGB with three channels × 8bits/channel) at 50p frames per second (fps) with a pixel resolution of 1920 × 1080 and saved in the AVCHD-HD format, MPEG-4 AVC as the codec. At the same time, the subject’s HR was measured by a TomTom Runner Cardio watch (TomTom International BV, Amsterdam, Netherlands). The watch was placed on the left arm, approximately 7cm away from the wrist, like a regular wrist watch. The use of the watch was unobtrusive, i.e. it did not affect the movements of the subjects, who could still use both hands to play the games. The watch recorded the HR at 1 Hz.

In the calibration phase of the experiment, the subjects answered a questionnaire after each game in order to provide a self-reported level of stress and boredom. The questionnaire had six questions: the first four were a 5-point Likert scale related to how the
player felt about their level of stress/boredom at the beginning/end of each game (1: not stressed/bored at all, 5: extremely stressed/bored); a question asking the player to identify the part of the game that best describes the moment of play the subject enjoyed the most (very beginning, after beginning and before middle, middle, after middle and before end, very end); finally, a question asking whether the subject understood the game. In the testing phase of the experiment, the subjects answered a questionnaire after each level of Infinite Mario, in order to provide a self-reported level of stress and boredom experienced during the level played. The questionnaire had two questions, one concerning stress and the other concerning boredom; both used a 5-point Likert scale related to the level of stress/boredom the player experienced during the level played (1: not stressed/bored at all, 5: extremely stressed/bored). Finally, before the end of the experiment, the subjects answered a questionnaire with ten questions related to their age; gender; number of hours per week spent playing games over the last year, i.e. question from the video game experience questionnaire (Unsworth et al., 2015); self-assessed level of proficiency or skill at playing video games, i.e. question from the Survey of Spatial Representation and Activities - SSRA (Terlecki and Newcombe, 2005); familiarity with puzzle, platform, Tetris and Mario games; current state of mind compared to other days (e.g. normal, unusually stressed, etc.); and familiarity with the research related to the experiment (unfamiliar, not very familiar, moderately familiar, very familiar).

9.2.5 DATA PREPROCESSING

Before any analysis was conducted, the video recordings of the experiment were preprocessed to allow the extraction of training and validation data. The process involved the extraction of the parts containing the interaction with games and levels, as well as the discarding of noisy frames. The pre-process procedure is notably similar to the one described in Section 8.8.2 (page 88). The periods during which the subjects played the available games and levels were extracted from the video recordings. The pre-processing of the calibration phase, illustrated in Figure 9.4, resulted in 3 videos per subject, denoted $C_{i,g}$ where $i$ is the $i$-th subject and $g \in \{1, 2, 3\}$ represents a calibration game. The initial $D = 45$ seconds of any given video $C_{x,i}$ were removed since they were deemed noisy regarding emotion information. The value of $D$ is a configuration parameter for the method that directly affects the amount of samples available for training the model. A value of $D = 45$ has been used in favor of the previously mentioned $D = 60$ to increase the number of samples collected for the training of the model. The remainder of the video was then divided into 3 segments, from which the first and the last were selected as $H_0$ and $H_1$, respectively. The middle part was discarded because its emotional state is unknown. Segments $H_0$ and $H_1$ represent the boring and stressful parts of the calibration games, respectively. The pre-processing of the testing phase resulted in 7 segments per subject, denoted $M_{i,m}$ where $i$ is the $i$-th subject and $m \in \{1, 2, ..., 7\}$ represents a level of Infinite Mario. Video segment $M_{1,4}$, for instance, represents the recording of subject 1 playing level $B_4$. No parts were discarded from video segments collected in the testing phase.

The pre-processing of all the recordings resulted in 13 video segments per subject: 3 segments $H_0$ (one per calibration game), 3 segments $H_1$ (one per calibration game), and 7 segments $M$ (one per level of Infinite Mario). Considering all the subjects, the pre-processing of the calibration phase resulted in $N = 186$ pairs of $H_0$ and $H_1$ video segments (3 calibration games $\times$ 62 subjects, resulting in 186 segments $H_0$ and 186 segments $H_1$). The testing phase resulted in $N = 434$ video segments $M$ (7 levels of Infinite Mario $\times$ 62 subjects).
Figure 9.4: Extraction of video segments $H_0$ and $H_1$ containing boring and stressful interactions, respectively, in the calibration games. Initial $D = 45$ seconds of any video $C_{i,g}$ are ignored and the remainder is divided into three segments, from which the first and the last ones are selected. Stripes highlight discarded video segments.

9.2.6 FEATURES EXTRACTION

Features used in the classification model were extracted remotely via the analysis of the video segments of each subject. In total, 8 features, denoted $F_1$ to $F_8$, were used in the process. Detailed information regarding how each feature was extracted, calculated, and aggregated, as well as the reasoning for its use is presented in Sections 8.8 and 8.9 (pages 82 and 94, respectively).

9.2.7 TRAINING OF THE EMOTION CLASSIFIER

The classification procedure uses the previously mentioned feature set and a neural network to identify two emotional states: boredom and stress. Both the training and evaluation of the neural network were performed on a user-tailored basis: data from the calibration games of a given subject $S$ were used to train a model for that subject, i.e. model $N_a$, which is then used to classify the emotional state of that given subject $S$ on levels of Infinite Mario. Figure 9.5 illustrates the process.

Figure 9.5: Training and evaluation of a user-tailored emotion classifier. (a) Training of the emotion classifier; (b) Constructions of a testing dataset; (c) Evaluation of the emotion classifier.

In the training process of each user-tailored model, illustrated in Figure 9.5(a), features are extracted from the video segments $H_0$ and $H_1$ (calibration data) of a given subject. This information is used to create a training dataset for that given subject. According to previous analysis, the subjects perceived the calibration games as boring at the beginning and stressful at the end. As a consequence, it is assumed that a subject’s emotional state...
in $H_0$ and $H_1$ is boredom and stress, respectively. Based on this assumption, training data obtained from video segments in $H_0$ and $H_1$ were labeled as boredom and stress, respectively.

The training dataset was then used to train the user-tailored model, i.e. neural network. The hyper-parameters of the subject’s model, e.g. number of neurons, were optimized using random search (Bergstra and Bengio, 2012). A 10-fold cross-validation method repeated 3 times was applied, so the training data was split into 10-subsets and each of those subset is left out while the model was trained on all the others. The process is repeated 3 times and the final metric for the model is the mean from the number of repeats. The area under the ROC curve (AUC) was used as a metric to select the best model.

The result of the training process is a trained neural network, i.e. $N_i$ where $i$ is the $i$-th subject. The model is said to be user-tailored because it was trained using only data from a given subject, e.g. subject $S_a$ produces model $N_a$.

### 9.2.8 Construction of a Testing Dataset

The majority of the works in the literature validate an emotion classifier by applying it to a share of the samples not used for training. Generally, all available data samples are split into two sets, e.g. one with 80% and one with 20% of all the samples, which are then used for training and testing/validation, respectively. In such a configuration, all data used in the process come from the same source, the only difference is how the data are distributed in the different sets. In contrast to that approach, the evaluation of the emotion classifier proposed in this experiment was validated using a completely different and independent dataset.

As mentioned in the previous section, data extracted from the calibration games, i.e. video segments $H_0$ and $H_1$ of a given subject $S_a$, are used to train a model $N_a$. On the other hand, data extracted from the Infinite Mario game, i.e. video segments $M_a$ of subject $S_a$, are sampled to produce a testing dataset. It is important to highlight how unique and challenging such a configuration is, since the user-tailored model is trained on one kind of dataset (calibration games) and tested/validated on another (evaluation game). Each dataset is derived from different and independent sources. Despite this configuration, the game used for evaluation, i.e. Infinite Mario, still shares common characteristics with the calibration games, such as the 2D and casual mechanic.

The feature extraction procedure described previously uses a moving window of 15 seconds with a step of 1 second. When it is applied to the video segments $M_i$, a new value for each feature is extracted per second. It has been reported in the literature that HR-based emotion estimation is possible every 10 seconds (Valenza et al., 2014), however reports also show changes in the inter-beat interval of HR within 4 seconds after in-game events (Ravaja et al., 2006). Regarding facial actions, they might change faster than the HR, therefore, it is reasonable to believe they could significantly change within a time span of 10 seconds. As a consequence, a sampling of 5 seconds was selected to collect feature values for the testing dataset of a given subject from video segments $M_i,m$. A sampling of 5 seconds was expected to cover changes both in HR and facial actions as often as possible, without risking the collection of samples that are not independent.

The testing dataset of a given subject $S_a$ contains samples (acquired every 5 seconds) from all the levels of that given subject $S_a$ selected for evaluation. The self-reported emotional state provided by each subject in the selected levels is used as ground truth to test the accuracy of the model. The levels of Infinite Mario used for evaluation are
selected according to the following procedure. Assuming that $r_{stress_{i,j}}$ and $r_{boredom_{i,j}}$ represent the self-reported levels of stress and boredom of subject $S_i$ in a Mario level $j$, respectively, a stress score $stress_{i,j}$ and a boredom score $boredom_{i,j}$ are calculated as:

$$stress_{i,j} = r_{stress_{i,j}} - r_{boredom_{i,j}}$$  \hfill (9.1)

$$boredom_{i,j} = r_{boredom_{i,j}} - r_{stress_{i,j}}$$  \hfill (9.2)

Two Mario levels, i.e. video segments $M_{i,m}$, of a given subject $i$ with the highest values for $stress_i$ are selected and used to sample the stress entries. Similarly, the two levels with the highest values for $boredom_i$ are selected and used to sample the boredom entries. In order to avoid sampling levels whose self-reported emotional state is inconclusive, e.g. stress and boredom levels are equal, the levels already selected for sampling, whose values of $stress_i$ or $boredom_i$ are not greater than or equal to 1 are excluded from the sampling process.

### 9.2.9 EVALUATION OF THE EMOTION CLASSIFIER

Similar to the training process, the evaluation process happens on a user basis, as illustrated by Figure 9.5(c). After the user-tailored model $N_a$ of a given subject $S_a$ has been trained, it is applied to the testing dataset of that subject. The testing dataset of a given subject $S_a$ contains all the samples (acquired every 5 seconds) from all the levels of that given subject $S_a$ selected for evaluation, as described in Section 9.2.8.

The evaluation of the proposed emotion classifier is based on classification accuracy. As mentioned previously, each user-tailored model $N_i$ is applied to a testing dataset sampled for that particular subject. Consequentially, each subject $S_i$ produces one single accuracy metric, named $A_i$. The overall classification accuracy of the proposed method is calculated on the basis of the mean of all $A_i$ values.

### 9.2.10 ANALYSIS

The aim of the experiment is to validate and prove the feasibility of the proposed emotion detection approach, i.e. the use of remotely acquired signals and a user-tailored model (trained on data from calibration games) to detect emotional states of stress and boredom. The feasibility of the approach will be tested in terms of classification accuracy. Thus, the following hypothesis states:

- **H1**: A user-tailored model, i.e. neural network, trained on data samples from three calibration games of a given subject $S_a$, i.e. Mushroom, Platformer and Tetris, is able to classify the emotional state of samples extracted from an evaluation game, i.e. Infinite Mario, played by that same subject $S_a$ with a mean accuracy greater than the chance-level rate;

The chance level is thereby the accuracy achieved assuming it is equally likely for a data sample to fall in any of the existing classes (Kassraian-Fard et al., 2016). For a balanced two-class problem, the chance-level classification accuracy equals 50%. However, a chance-level accuracy rate of 50% assumes that a classifier performs random guessing.
on datasets of infinite size. As a consequence, random guessing approximates chance-level accuracies, if the testing data set is large enough. If the testing data set is small, random classification can deliver accuracies that significantly deviate from chance level (Combrisson and Jerbi, 2015).

In order to account for that, a minimal correct classification rate for accuracy has been calculated using the binomial cumulative distribution to assert statistical significance with a confidence level of 95% \( (p < 0.05) \) as a function of sample size \( n \), i.e. size of validation dataset, and number of classes, i.e. boredom and stress (Combrisson and Jerbi, 2015). Since the subjects had different gaming skills, the time spent playing the levels of Infinite Mario was likely to differ. This variance produced validation datasets of different sizes among the subjects. An analysis of all the validation datasets shows a mean size of 64.4 samples with 32.1 and 32.3 as the mean number of stress and boredom samples in each set, respectively. Therefore, it was assumed that the proposed method was to be validated as a balanced two-class problem with a sample size of \( n = 64 \) (on average) used for the evaluation of each classification. Based on these numbers, a mean classification accuracy rate of 60% was found to be the minimal rate to assert classification better than chance-level. Consequentially, the null hypothesis associated with \( u_1 \) is: a user-tailored model trained on data samples from three calibration games of a given subject \( S_a \) is not able to classify the emotional state of samples extracted from an evaluation game played by that same subject \( S_a \) with a mean accuracy greater than chance-level rate.

Hypothesis \( u_1 \) was tested by checking whether the mean value of the classification accuracy, i.e. calculated from all \( A_i \) values, is greater than 0.6.

### 9.3 RESULTS

The following sections present the results obtained from the data analysis performed according to the previously described method.

#### 9.3.1 SELF-REPORTED EMOTIONAL STATE

The emotional state of the subjects during the interaction with the Infinite Mario game is an important element of the experiment. Table 9.2 shows the mean value and standard deviation of the answers given in the self-reported emotional state questionnaire after each level of Infinite Mario.

Table 9.2: Mean value of the answers given in the self-reported emotional state questionnaire after levels of Infinite Mario

<table>
<thead>
<tr>
<th>Level</th>
<th>Stress</th>
<th>Boredom</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A_1 )</td>
<td>1.6 ± 0.8</td>
<td>2.3 ± 1.2</td>
</tr>
<tr>
<td>( A_2 )</td>
<td>2.1 ± 0.9</td>
<td>1.8 ± 1.1</td>
</tr>
<tr>
<td>( A_3 )</td>
<td>2.9 ± 0.9</td>
<td>1.9 ± 1.2</td>
</tr>
<tr>
<td>( B_1 )</td>
<td>1.5 ± 1.0</td>
<td>3.9 ± 1.2</td>
</tr>
<tr>
<td>( B_2 )</td>
<td>2.0 ± 0.8</td>
<td>2.2 ± 1.2</td>
</tr>
<tr>
<td>( B_3 )</td>
<td>3.0 ± 1.1</td>
<td>2.1 ± 1.2</td>
</tr>
<tr>
<td>( C_1 )</td>
<td>1.3 ± 0.7</td>
<td>4.0 ± 1.2</td>
</tr>
</tbody>
</table>
Levels $A_3$ and $B_3$ presented 2.9 and 3.0 as the mean value for reported stress, respectively, the two highest mean values for stress. Using level $A_1$ as a baseline, since it presented the lowest median values for stress and boredom, i.e. 1.0 and 2.0 respectively, a Wilcoxon Signed-ranks test indicated different stress levels between $A_1$ (median 2.0) and $A_3$ (median 3.0), $Z = -5.78, p < 0.001, r = 0.73$. The same test also indicated different stress levels between $A_1$ (median 2.0) and $B_3$ (median 3.0), $Z = -5.55, p < 0.001, r = 0.70$.

Levels $B_1$ and $C_1$ presented 3.9 and 4.0 as the mean values for reported boredom, respectively, the two highest mean values for boredom. Repeating the use of level $A_1$ as a baseline, a Wilcoxon Signed-ranks test indicated different boredom levels between $A_1$ (median 2.0) and $B_1$ (median 4.0), $Z = -6.14, p < 0.001, r = 0.77$. The same test also indicated different boredom levels between $A_1$ (median 2.0) and $C_1$ (median 4.0), $Z = -6.21, p < 0.001, r = 0.78$.

As mentioned in Section 9.2.3, levels $A_3$ and $B_3$ were adjusted to be perceived as more stressful. Similarly, levels $B_1$ and $C_1$ were adjusted to be perceived as more boring. The results with statistical significance confirm that the adjustments applied to these levels of Infinite Mario indeed caused a particular emotional state. Even though the analysis presented here was performed on a group basis and the levels of Infinite Mario used in the evaluation of the method were selected on a user basis (according to individually self-reported emotional states), it is possible to conclude that there are indeed levels that were perceived with different emotional states. This is essential for the evaluation and validation of the proposed method.

### 9.3.2 EMOTION CLASSIFICATION

A subject’s emotional state during the interaction with particular levels of Infinite Mario was classified as stress or boredom using the proposed method. Table 9.3 presents the mean value of the resulting classification metrics for accuracy, precision, recall and F1 score, calculated and analyzed according to the procedures described in Section 9.2.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6158</td>
<td>0.6163</td>
<td>0.6127</td>
<td>0.5786</td>
</tr>
</tbody>
</table>

The proposed method was able to identify the emotional state of subjects with a mean accuracy of 61.6%. As previously mentioned, hypothesis $u_1$ states that a user-tailored model, i.e. neural network, trained on data samples from three calibration games of a given subject $S_a$, i.e. Mushroom, Platformer and Tetris, is able to classify the emotional state of samples extracted from an evaluation game, i.e. Infinite Mario, played by that same subject $S_a$ with a mean accuracy greater than 60% (calculated chance-level rate). A mean accuracy of 61.6% refutes the null hypothesis, supporting the claim of hypothesis $u_1$. It confirms the feasibility of the proposed method to perform better than chance-level estimations.

Since the subjects were evaluated independently, the mean classification accuracy is not enough to contextualize the estimations at a user level. Figure 9.6 provides a better context of the accuracy values distribution at a subject level, as demonstrated by a histogram and density curves. As illustrated in the histogram of Figure 9.6(a), the majority of the subjects presented an emotion classification accuracy close to 60%. Particularly inaccu-
rate estimations can be seen in a group of 13 subjects (20.9%) that were classified with a mean accuracy less than 50%. In contrast, a group of 14 subjects (22.5%) presented particularly precise emotion estimations with a classification accuracy greater than 75%.

Figure 9.6: Distribution of accuracy values at a subject level. (a) Histogram showing the number of subjects and the accuracy rate obtained in their emotion classification. (b) Density curve regarding the distribution of accuracy values.

Figure 9.6(b) shows a density curve with default bandwidth and an adjustment of 0.25 regarding the distribution of accuracy values. In this illustration, the area under any part of the curve provides the probability that the accuracy value would equal that particular group of samples. As expected, there is a greater likelihood that samples are classified with an accuracy of 60% (mean accuracy). Interestingly, there is the likelihood that samples are classified with an accuracy of 80% or 90%, which is more likely than a classification accuracy of 40%.

### 9.4 DISCUSSION

The results of the proposed method for emotion detection based on remotely acquired signals show that the mean classification accuracy of 61.6% is better than a calculated 60% chance-level rate. As described in Section 9.2, the evaluation of the method has been performed as a balanced two-class problem with a sample size of \( n \approx 64 \) (on average). In particular, the chance-level mean classification accuracy rate of 60% has been found by assuming a binomial distribution for the classification error to ensure a statistically significant classification (Combrisson and Jerbi, 2015). The achieved mean classification accuracy of 61.6% supports the proposed hypothesis \( H_1 \), proving that a user-tailored model trained on data samples from three calibration games of a given subject is able to classify the emotional state of samples extracted from an evaluation game played by that same subject. The use of calibration games as emotion elicitation for training an emotion classifier is a novel aspect of the method presented here. The results support such an idea, showing that calibration games could be used as emotion elicitation material.
Despite the fact that a mean classification accuracy of 61.6% is better than the calculated chance-level accuracy of 60%, it is still below the mean classification accuracy achieved in other affective computing studies. A survey conducted by Moghimi, R. Stone, and Rotshtein (2017) of over 33 affective computing studies undertaken since 1993 shows a mean classification accuracy of 77.91% (±12.76%, minimum of 50%, i.e. random classification, and maximum of 96.5%). The method proposed here is within such a reported range, however, a fair comparison of evaluation metrics is virtually impossible considering the different methods, setups and aims. For instance, the mentioned survey presents only 6 studies (18%) that used game-related stimuli, however all of them used physical sensors to acquire user signals. Compared in isolation, the mean accuracy is a simple metric that can be used to estimate how well an approach can classify emotional states. However, the evaluation method and the procedures used for training/testing the model can profoundly influence accuracy results. For instance, Kukolja et al. (2014) classify 5 emotions using kNN (nearest neighbors) based on physiological signals obtained with physical sensors. When Leave-One-Out Cross-Validation (LOOCV) is employed, i.e. available data is divided into parts and one part is left out while the rest is used for training, the mean evaluation accuracy is 78.76%. When LOSOVC is used, i.e. one experimental session is left out for testing and the remaining ones are used for training, the mean evaluation accuracy drops to 56.18%. Studies using LOOCV usually report mean classification accuracy in the range of 60-80%, however, LOOCV is less likely to be encountered in a real world situation. When the data available is divided into two groups, e.g. training and testing datasets, samples that are highly correlated, i.e. samples from the same game or session, could exist in both datasets. In the present experiment, for instance, data samples from the game being evaluated, i.e. Infinite Mario, are not in the training dataset. They are, in fact, used exclusively in the testing/evaluation dataset, which is completely independent from the training data. Classification performance on fresh data from the validation set is a better measure for how well the classifiers generalize (James et al., 2013, Chapter 5). Therefore, the use of LOOCV, even when k-fold cross-validation is used, presents testing samples that are considerably similar to those found in the training dataset, which could lead to higher mean classification accuracy.

It is also important to highlight how the signals used in the emotion classification are acquired. In the method proposed in the present experiment, all the information used to build the user-tailored model is collected remotely, in a non-obtrusive manner. The previously mentioned mean classification accuracy of 77.91% from other affective computing studies depends on physical sensors to acquire a subject’s signals. A completely remote, emotion estimation setup presents a significant set of challenges. The results of the completely remote data acquisition employed by previous studies show an accuracy rate of 89% for negative, 56% for neutral and 78% for positive state identification (D. Zhou et al., 2015). For only stress detection, the mean classification accuracy reached best marks of 80% and 90% in contexts involving interactions with stressful images and the Stroop test (Giannakakis et al., 2017). Finally, in a context involving the detection of cognitive stress, the mean accuracy classification of rest vs. cognitive load has been reported as 86% (McDuff, Hernandez, et al., 2016). All those studies rely on a completely remote method for data acquisition, however, the context of the classification is not focused on games research or in the use of games as the source of emotion elicitation for the training of a model. In the majority of cases, the subjects are also instructed to remain still, which is unlikely to happen in the natural interaction with games. Additionally, LOOCV or equivalent is used in some cases, which influences the model accuracy, as previously mentioned. As detailed in the survey by Moghimi, R. Stone, and Rotshtein (2017), there are many different experimental setups and approaches for affective computing. Given the peculiarities of each approach, including how the model is trained and
evaluated, it would be unfair and naive to make a direct comparison of the studies. Attention should focus on the method, aims and evaluation of any presented approach, so merit can be decided.

Finally, an important factor in the present experiment is the material used to produce the emotional stimuli. In contrast to previous studies, the subjects interacted with complete digital games, not images, videos or text as content, to produce the emotional stimuli. The evaluation game, and the calibration games used in the experiment are not gamified cognitive tests, e.g. Stroop test (Golden, 1978). It strengthens the applicability of the results in the field of games research, which is the foundation and the aim of the proposed approach. Another remark is that a mean classification accuracy of 61.6% might not necessarily be connected to flaws in the proposed approach, but due to limitations in the labeling of ground data. The most reliable technique to assess and label emotional experiences in order to perform appropriate psycho-physiological signal classification is self-assessment of the emotional state (Moghimi, R. Stone, and Rotshtein, 2017). However, even if a subject reported a particular level of Infinite Mario as stressful, it does not mean that all the samples collected from that level represent an emotional state of stress. It is plausible to believe that the subjects experienced fluctuations of emotions during a single level, e.g. stress, happiness, and even boredom. Such nuances are not captured by the labeling process used to create the evaluation dataset, which could lead to lower classification accuracy. The heterogeneous nature of the subject population, however, should ensure that such a factor is accounted for. It should be noted that the considerable number of subjects in the experiment, i.e. $N = 62$, is greater than the average number of participants in previous affective computing studies, which is $N = 25.5$ subjects per experiment (Moghimi, R. Stone, and Rotshtein, 2017). This allows a broad evaluation of the proposed approach, accounting for different player profiles and supporting the claims of the previously mentioned hypothesis.

9.5 CONCLUSION

The experiment has validated and shown the accuracy of the proposed method for emotion detection. Such a method uses remotely acquired signals, i.e. heart rate and facial actions, and machine learning to detect emotional states of stress and boredom on a user-tailored basis. In order to test the method, calibration games, i.e. Mushroom, Platformer and Tetris, have been used as emotion elicitation material. A fourth game, i.e. Infinite Mario, has been used as an evaluation game. Some levels of Infinite Mario were adjusted so that the subjects would more likely perceive them as stressful or boring, thus allowing the proposed method to be evaluated according to such differences in the emotional state. The analysis performed on the levels of Infinite Mario has shown with statistical significance that some of levels were indeed perceived as more stressful or boring than others.

Regarding the evaluation of the emotion classification, the results confirm that the proposed method was able to identify the emotional state of subjects with a mean accuracy of 61.6%. The achieved accuracy confirms with statistical significance that the proposed method performs better than a calculated chance-level estimation of 60%. Despite the fact that a mean classification accuracy of 61.6% is better than chance-level rate, it is still below the mean classification accuracy achieved in other affective computing studies, i.e. 77.91%. Finally, the results suggest that a multifactorial, user-tailored model trained on data samples extracted from calibration games is a feasible method to classify the emotional states of users during their interaction with games.
CHAPTER 10

NON-OBTRUSIVE DETECTION OF EMOTIONS

The initial chapters of this thesis present the theoretical foundations and the work needed to create a novel method for the remote detection of users’ emotions during their interaction with games. As highlighted by previous research, the understanding of human emotions, as well as the process of automatically detecting them, is the aim of a number of researchers in many different fields. As detailed in Chapter 3, different theories for modeling and studying emotions in a variety of contexts, including those related to games, have been proposed. A considerable number of those theories are based on the human physiology, connecting emotional reactions to psychophysiological signals, e.g. HR and facial activity. Several approaches to putting such models and theories into practice, to achieve the ultimate goal of detecting what a person is feeling, have been proposed. Chapters 4 and 5, for instance, describe the connection between emotions and their manifestations in the body, particularly the process of mapping measurable psychophysiological signals into an emotional state.

Emotion detection is a complex and multidisciplinary problem that demands knowledge from many different areas. In this thesis, focus is given to the field of games research. This chapter presents and discusses the outcomes of this research, which is focused on creating a non-obtrusive method for emotion detection, particularly in the context of games. The results and contributions of this research are aimed at and discussed in the light of games research; however, they may be useful for scholars in other fields as well. The following sections also present insights obtained during the systematic investigation and development of the proposed method, including a discussion on how they relate to games research and other areas.

10.1 GAME-BASED MODEL FOR EMOTION DETECTION

Generally, the process of detecting emotions using psychophysiological signals relies on mapping the patterns of such signals into an emotional state. As indicated by the literature review conducted in this thesis, a validated way of achieving this is by measuring the changes of psychophysiological signals caused by the interaction between users and emotion elicitation materials. The process typically involves three main parts: emotion elicitation, signal acquisition and the mapping of such signals into an emotional state. Simply put, subjects are exposed to materials that are likely to produce certain emotional reactions, e.g. video and images depicting sad events. This is then followed by observations of how the signals of interest, e.g. HR, change in accordance. Finally, the emotion detection is conducted by a technique that aims to produce a model to map the changes of those signals into emotional states, e.g. machine learning model like neural networks. The literature review presented in this thesis reveals a myriad of different approaches that have been used in each of the previously mentioned parts.

Most previous research focuses on producing a group model, where data from several individuals are used to create a trained machine capable of detecting the emotions of any
other subject outside the training population. Contrary to the established notion that a group model is better, this research investigated the venue of a user-tailored approach. As indicated by previous findings (Bailenson et al., 2008), a model trained on the data of a given person might be better at predicting the emotional state of that person. This is motivated by the fact that people are different in many aspects, including cultural and personal expectations (Goldberg, 1993). Furthermore, it is reasonable to believe that these individual characteristics might be preserved and better accounted for in a method that applies a user-tailored model, instead of a group model, to detect emotions. In this thesis, both the emotion elicitation process and the mapping of psychophysiological signals into emotional states were focused on the notion of the individual instead of the group.

10.1.1 CALIBRATION GAMES AS EMOTION ELICITATION

Previous works have used several different emotion elicitation materials, mainly images and videos, and less often game-related elements. These materials, however, lack a more user-tailored approach for studying the variations of signals. When games are used, emotional states, such as stress and boredom, are often inducted by administering a game with the same particular setup, e.g. high/low difficulty, to all subjects. People respond differently to media according to their personality (Ravaja, 2004), and differ in social, learning and play styles (Goldberg, 1993). A game session labeled as stressful, for instance, assumes that all the subjects have the same expectations and behave similarly, which dilutes the individuality of each person, since some might experience the interaction as not being stressful as intended. Additionally, the analysis usually involves the interaction of subjects with some game levels (from the same game) that feature a constant difficulty scale, which does not contemplate the variations of signals in a context where the difficulty of the game constantly increases in the same game level/session.

The investigation of better game-based emotion elicitation materials was one of the main aspects of this research. In the aim to properly elicit particular emotional states in each user, this research introduced the novel idea of calibration games. As detailed in Section 8.4 (page 61), calibration games are carefully designed to have a difficulty level that constantly and linearly progresses over time without a pre-defined stopping point. At the beginning, the games are highly predictive, without novelties, changes or surprises and with an emphasis on the passage of time during a wait, which leads to an emotional state of boredom (Aart et al., 2010; Koster, 2013; Schell, 2014). The game difficulty is then periodically increased until the subject is not able to cope with the challenges at hand, which happens at different times for different users. The ever-increasing game difficulty leads to an emotional state of stress towards the end of the interaction, accounting for the different expectations and gaming skill of a wide range of users.

Sections 8.5 and 8.6 present a detailed analysis of how the responses to psychophysiological activity, i.e. HR and facial actions (FA), relate to emotional states in a context featuring calibration games. The results show that a calibration game is a valid emotion elicitation material which accounts for personal differences among subjects when it induces the emotional states of stress and boredom. Using the proposed calibration games, made it possible to observe and confirm, with statistical significance, the variations of HR and naked-eye recognizable FA that happened during the interactions with the games, especially during situations that were designed to provoke boredom and stress. These findings were an essential part of the user-tailored method proposed in this thesis, since they proved that calibration games can be used as emotion elicitation material. Another important factor is the nature of the calibration games compared to other emotional
stimuli, e.g. images or videos. The use of images, videos or text as content to produce the emotional stimuli is less likely to produce the reactions of a real gaming session. In a game, users are in charge of actions, which are bound to have consequences. A bad judgment might cause the main character to be hurt, or a correct movement might produce a reward. This feedback loop happen constantly in a game, likely producing emotional reactions in the user. It is plausible to believe that the calibration games present a more sophisticated interaction through their game mechanics, as opposed to the simplistic, one-way interaction between users and images/videos, for instance. Consequentially the use of calibration games is likely to create a deeper emotional connection between users and the emotion elicitation material, resulting in clear and observable changes in psychophysiological signals.

10.1.2 REMOTE READINGS OF PSYCHOPHYSIOLOGICAL SIGNALS

Several works found in the literature rely on physical sensors to acquire the signals used in the emotion detection model. Physical sensors are not convenient since they require a cumbersome setup and might disturb the user experience, i.e. invalidate the use of a finger or hand. The use of remote sensing to acquire psychophysiological signals, a non-obtrusive data collection approach, is mentioned in the literature as a promising solution for this problem. The remote sensing of psychophysiological signals is an essential part of this research, since its objective is to create a method that is able to non-obtrusively detect user emotions. A complete non-obtrusive method for signal acquisition, however, is a complex and challenging problem, particularly in a context involving games. The literature review conducted for this thesis found the main psychophysiological signals that can be remotely acquired and whose data can be used to detect emotional states.

One of the signals that can be generally acquired using remote and non-obtrusive approaches is facial activity. Chapter 4 describes in detail the techniques for facial analysis and the approaches that use them for emotion detection. As mentioned in the chapter, the results indicate that facial analysis is a promising source of information for use in the process of emotion detection. Additionally, the combined use of facial and body features (multimodal emotion recognition) is known to perform better than using either one alone (Zacharatos, Gatzoulis, and Chrysanthou, 2014). Following the findings of previous work, the present thesis used facial activity as an important signal in the emotion detection process. A novel method for the automated analysis of facial cues from videos was developed, as explained in Section 8.8 (page 82). Empirical results of this method show its potential for detecting stress and boredom in players of games. The method is based on Euclidean distances between automatically detected facial points, designed to be robust enough to correctly perform facial analysis even when users naturally interact with games. In such a case, players behave naturally as they play, e.g. moving, laughing and speaking. Evaluations of the method were conducted experimentally using game-based emotion elicitation, which properly contextualized the efficiency of the method in the field of games research. The results, detailed in Section 8.8, confirm the method has the potential to differentiate the emotional states of boredom and stress in players. However, the natural behavior of users during the interaction with the games is a significant factor impacting the process.

Another signal acquired using remote and non-obtrusive approaches is HR and its derivatives. Chapter 6 details the progress that has been made in the remote estimation of physiological signals, particularly the use of rPPG to estimate HR. Despite the potential of rPPG to eliminate physical sensors completely, its use is considerably impacted by the natural behavior of users. As presented in Section 8.7 (page 73), rPPG estimations of
HR are sensitive to noise caused by movement, facial expressions or changes in illumination (e.g. screen activity reflected on user’s face), which are all likely to happen in gaming sessions. These interferences can produce unreliable measurements of the HR signal, resulting in misleading data. Although these challenges are limiting factors, the use of the remote measurement of physiological signals, such as rPPG, has already been applied to emotion detection. Signals, such as HR and HRV, were used to remotely detect stress (McDuff, Hernandez, et al., 2016; McDuff, Gontarek, and Picard, 2014a; Bousefsaf, Maaoui, and Pruski, 2013b), for instance. In the majority of the cases, the subjects are typically instructed to stay still (Rouast et al., 2016), which improves the accuracy of the rPPG technique. In some other cases, however, authors evaluate the accuracy of the HR estimation in scenarios where subjects are instructed to act naturally. Despite the fact that such works present experiments where subjects are told to behave naturally, their accuracy evaluation is based on artificial or simple human-computer interactions. Subjects stare idly at the camera (Zhao et al., 2013; Y. Hsu, Lin, and W. Hsu, 2014), faking an interaction with a computer (Poh, McDuff, and Picard, 2010), working on a task, i.e. creating a website (Monkaresi, Calvo, and Yan, 2014), mentally subtracting numbers (McDuff, Gontarek, and Picard, 2014b), or performing arbitrary movements (Tran, Lee, and C. Kim, 2015), e.g. head rotation in different degrees. In contrast to previous works, this thesis employed rPPG in a context that enabled users to naturally interact with games. Information related to HR is an important physiological indication of the emotional state of users, therefore, the use of rPPG in this thesis to remotely acquire HR data was essential. Extensive evaluations were conducted to establish the reliability of remote HR measurements acquired in situations where users behaved naturally, rather than differently to their usual behavior. The analysis of the accuracy of remote HR estimations clearly established the limitations of the rPPG technique, showing how it is affected by user behavior. One of these identified limitations is the effect of facial occlusion on the rPPG estimations of the HR. The act of resting the chin on the palm of the hand, for instance, a common trait of bored users, significantly affects the process of detecting a face in the videos, thus directly affecting the estimations of HR. Evaluation results of the rPPG technique, as detailed in Section 8.7, have shown an average estimation error within the range that still allows the identification of HR variations caused by emotion elicitation materials, as detailed in Section 8.6. It shows that it is feasible to remotely extract HR and facial data from video recordings of users interacting with games. As a consequence, those signals can be acquired non-obtrusively and used to detect the emotional state of users playing games.

10.1.3 MULTIFACTORIAL EMOTION DETECTION

The literature review supporting this thesis suggests that the mapping of psychophysiological signals into emotional states based on a multifactorial analysis, when more than one signal is used, is more likely to produce accurate results (Kukolja et al., 2014). As detailed in Chapter 7, a combination of signals can reduce the interference and noise caused by signal manipulation, enhancing the accuracy of an emotion detector. Early studies mapping psychophysiological signals into emotional states focused on a multimodal analysis, when more than one modality, e.g. ECG and skin conductivity, are used in conjunction. Such approaches use a wide variety of physical sensors to acquire signal data. As previously mentioned, physical sensors are not ideal, so a completely remote-based approach for data acquisition would be of interest. As mentioned in Chapter 7, a few studies focus on the remote extraction of different user signals, i.e. HR and blinking rate, from a single source, i.e. video recording. In such a
case, a single modality is used, i.e., video, however a set of different signals (factors) is extracted and used in the emotion detection process, i.e., multifactorial analysis. Despite the fact that those studies use non-obtrusive extraction of user signals for emotion detection, they do not use game-focused elicitation materials. The research presented in this thesis used a novel approach to produce a multifactorial analysis, which is non-obtrusive, user-tailored and game-focused. The overall idea, presented in Section 1.2 (page 6), is to remotely extract signals from a given user playing calibration games, then use that data to train a user-tailored neural network, i.e., data from a given user are employed to train a single neural network tailored to that given user. Finally, the signals that a given user exhibits during the interaction with a particular, but ordinary, non-calibration game are remotely extracted and fed into the previously trained neural network. The neural network then outputs the emotional state of that given user in that particular game.

It is important to highlight that this method for emotion detection based on calibration games, remote sensing and a user-tailored multifactorial analysis has not been found in the literature. Furthermore, its feasibility was unknown and the research conducted and reported in this thesis reflects the steps taken to conceive and validate this method for emotion detection. In that light, studies 1 to 4, detailed in Sections 8.5 to 8.8, were conducted to evaluate each component of this novel emotion detection method. They focused on understanding the capabilities and limitations of each component, including their validity in the process, e.g., use of calibration games as emotion elicitation material. Each of the studies contributed to the final assembling of the novel architecture for emotion detection proposed in this thesis. Finally, study 5, detailed in Section 8.9 (page 94), has shown a systematic evaluation of the feasibility of the proposed method in detecting the emotions of users actually playing games. That study was based on the first experiment conducted. It tested a user-tailored neural network trained on data samples from two calibration games of a given subject, which was then used to classify samples from a third calibration game of that same subject. The evaluation included the testing of different user signals, such as HR and facial data combined or used separately. The intent of such an evaluation was to better understand the benefits of a multifactorial analysis, confirming that it indeed produces better estimations of the emotional state of users. The results suggest that the proposed method is feasible and has the potential to non-obtrusively detect the emotional state of users, in a user-tailored fashion during their interaction with games.

10.1.4 USAGE AND VALIDATION

The initial feasibility evaluation of the proposed method, conducted in study 5 (Section 8.9, page 94) was limited by the number of games available for analysis, i.e., three, and the reduced number of subjects. Despite these constraints, the results suggest the feasibility of a non-obtrusive, game-based and user-tailored emotion detector. In order to further test such suggestions, a final validation of the proposed method was conducted in a second experiment with a larger sample size, as detailed in Chapter 9. In the experiment, the previously mentioned calibration games, i.e., Mushroom, Platformer and Tetris, were used as emotion elicitation materials to train a user-tailored model, i.e., neural network. This model was then used to detect the emotional state of each user during the interaction with a fourth game, i.e., Infinite Mario. Following the expectations inferred by study 5, the proposed method was able to identify the emotional state of subjects with a mean accuracy of 61.6%. The results confirm with statistical significance that the proposed method indeed classifies emotional states, achieving an accuracy rate better than chance-level classification.
However, compared to existing works in the literature, the mean classification accuracy of 61.6% achieved by the proposed method is still below the mean classification accuracy achieved by other affective computing studies, i.e. 77.91%. A fair comparison of these numbers, however, is not possible. Each study is conducted in particular situations, using different emotion elicitation materials and different training/testing models. As previously mentioned, the method proposed in this thesis focuses on the individual, not the group, which is a common factor found in the literature. Another important and highly distinctive difference between the present work and existing ones is how data is obtained to train and evaluate the emotion detection model. The method proposed in this thesis uses a completely independent dataset to train the model, which is obtained from the natural interactions of users with game-focused elicitation materials, i.e. calibration games. These games are similar to COTS games, which portray a more real gaming experience. The evaluation of the method was conducted on the Infinite Mario game, which mimics the commercial Super Mario game. Data from this game were never used in the trained model, yet it was able to classify the emotional state of users. As mentioned previously, the evaluation of classification performance on fresh data is a better measure of how well classifiers generalize (James et al., 2013, Chapter 5). Several works focused on emotion detection do not use game-focused materials for training of a model. Commonly they evaluate accuracy by testing samples that are considerably similar to those found in the training dataset, e.g. splitting available data into training and evaluation datasets, which is different from what is presented in this thesis.

Finally, it is relevant to mention that the method for emotion detection proposed by this research had to deal with significant challenges not faced by previous works. It aims to remotely detect the emotions of players without instructing them how to behave during their interaction with games, i.e. natural behavior. Such a configuration presents a set of unique challenges that inherently affect the results of an emotion detection procedure. The remote acquisition of signals, for instance, is considerably influenced by a subject’s movements, which are themselves directly affected by the interaction between the subjects and the games being played. Previous studies that use physical sensors, non-interactive emotion elicitation materials, e.g. images, instructions for subjects to keep still, and highly correlated samples for training and testing any emotion classification model faced less challenging conditions than the method presented in this thesis. Therefore, they were able to obtain data with less noise and ambiguity which is more likely to lead to better results for emotion detection. The research presented in this thesis features unique elements, e.g. calibration games, the natural behavior of subjects and the remote acquisition of signals, which were employed under significantly challenging circumstances. The proposed method is less likely to outperform the accuracy achieved by previous affective computing studies that were more established and were conducted in more controlled contexts, e.g. studies reported by Moghimi, R. Stone, and Rotshtein (2017). However, it is important to emphasize that the systematic evaluation of the method proposed in this thesis during various studies and experiments, repeatedly indicated its feasibility and potential in different statistical analyses.

As a final and less significant note, it can be mentioned that the design of the method for emotion detection proposed in this thesis allows classifications of emotions in real-time, in contrast to some existing approaches that require offline analysis of all available data. Assuming that a user-tailored model has already been trained, the proposed method only needs an initial amount of data points to detect emotional states, i.e. enough data to fill the moving window used for the analysis. After the minimum amount of data has been collected, e.g. 15 seconds of video data, the method can continuously estimate the emotional state of a user. The estimation can be performed with an arbitrary time interval, including for each new frame of the input video feed. The majority of cases, however,
are more likely to employ a longer interval between classifications, e.g. 1 second, which was the value (of the window step) used in the evaluations presented in Section 8.9 and Chapter 9. Nevertheless, the proposed method is able to detect emotions off-line, i.e. analyze a previously recorded video, or on-line, i.e. real-time analysis of a live video streaming session.

10.2 INSIGHTS OUTSIDE GAMES RESEARCH

The contributions of this thesis are aimed at the field of games research, however, techniques and theories from a variety of fields, e.g. computer vision and psychology, were used in the investigation process. Several of the components that constitute the proposed method for emotion detection were studied and evaluated separately, which in themselves produced insights that could be used outside the field of games research.

10.2.1 FACIAL BEHAVIOR AND EMOTIONS

One insight that this thesis presents is related to the exploration of facial activity under stressful and boring situations. As detailed in Section 8.5 (page 63), observations of facial actions during the interaction with games indicate that a neutral face remains for longer periods of time during boring segments of play. Additionally, in the context of the experiment presented in this thesis, the facial analysis at an individual level, as opposed to a group level, produced more information to connect facial activity to stress/boredom. This analysis was based on an experiment with a particular configuration that allowed a better exploration of how facial activity relates to the emotional states of stress and boredom. The results obtained could be used by other scholars or practitioners interested in understanding or exploring the relationship between facial activity and emotions. In that light, the present research also introduces a novel method for the automated analysis of facial behavior, which has been proven to have the potential to differentiate, in real-time, the emotional states of boredom and stress in users. The evaluations of such an automated facial analysis show that the values of facial features detected during boring periods are different from values of the same facial features detected during stressful periods. These results can contribute to further investigations regarding emotions and automated facial analysis. They could also be used as indicators of the emotional state of users in human-computer or human-robot interactions, for instance.

10.2.2 PHYSIOLOGICAL ACTIVITY AND EMOTIONS

Another insight that this thesis presents is related to the remote estimations of HR using rPPG in a context involving natural behavior. The use of rPPG is a promising technique to obtain physiological signals from users/subjects non-obtrusively. Such information has applications in research and industry. As detailed in Section 8.7 (page 73), the evaluation conducted in this thesis regarding rPPG and natural behavior contributes more information related to the reliability of remote HR measurements. Even though the evaluation was performed in a context involving games, the analysis still indicates the limitations of the rPPG technique when applied on users interacting with a computer. As demonstrated, a user’s natural behavior, e.g. movement, affects rPPG estimations to different extents. Such information can guide the use of rPPG in contexts where the mentioned natural behavior is an inherent factor.
Finally, this thesis presents an extensive analysis of how physiological signals, particularly HR, relate to emotional states of boredom and stress. As detailed in Section 8.6 (page 67), this research presents indications that the average HR mean during periods of stress is greater than the average HR mean during periods of boredom in the interaction with games. Similar to the facial exploration mentioned previously, the analysis of the HR activity was performed in an experiment that permitted a more elaborate observation of such activity in boring and stressful interactions. The results of this analysis suggest that changes in HR are a promising indicator of stress, which contributes information to the body of knowledge of physiological reactions and emotions.
CHAPTER 11
SOFTWARE FOR EMOTION DETECTION

During the systematic evaluation of the elements that compose the method proposed in this thesis, several software-based tests were performed. Additionally, all the data collection, e.g. HR estimations and facial actions, was carried out with custom made software. The previously mentioned automated facial analysis, responsible for collecting facial cues from videos, is an example. Each of these code components was incorporated in a software, which can be seen as a partial instantiation of the artifact produced as the outcome of this research. Figure 11.1 shows the software working on a video file. This chapter details each of these components and how they were unified in a single software that can be used to collect data for the remote estimation of the emotional states of a player interacting with games.

![Software developed as a partial instantiation of the proposed method working on a video file. Detected facial points and eye gaze information are superimposed on each frame of the video.](image)

11.1 OVERALL STRUCTURE

The software was mainly developed in C++ using the computer vision library OpenCV (Bradski, 2000). The overall structure of the software is illustrated in Figure 11.2. The system contains six main components, i.e. emotion model, emotion estimator, face de-
FIGURE 11.2: OVERALL STRUCTURE OF THE SOFTWARE AND ITS COMPONENTS.

11.1.1 FACE DETECTOR

The face detector component locates a human face in a given frame read from the input video being analyzed. It performs a face alignment procedure using one of the two available algorithms: Constrained Local Neural Fields (CLNF) (Baltrusaitis, Robinson, and Morency, 2013) and Ensemble of Regression Trees (ERT) (Kazemi and Sullivan, 2014). The face alignment procedure is based on existing implementations of CLNF and ERT provided by OpenFace (Baltrusaitis, Robinson, and Morency, 2016) and dlib \(^1\) (King, 2009), respectively.

The output of the face detector is a vector containing 68 2D points, each one representing a facial landmark. Figure 11.3 shows a visual representation of the mentioned vector and its points overlapped in a face.

11.1.2 FACE ANALYZER

The face analyzer component uses a frame of the input video and the information related to facial landmarks provided by the face detector. The face analyzer uses that data to extract information regarding facial activity, e.g., eye area. The face analyzer orchestrates a list of independent analyzers, each one responsible for extracting specific activity patterns, e.g., eye area. Figure 11.4 shows the user interface regarding the data provided by the face analyzer.

Available analyzers extract information regarding eye (including eyebrow activity), mouth (lips/mouth activity), facial center of mass (mean position of all detected landmarks),

\(^1\)http://dlib.net
distance among facial landmarks, face instability (including measurement of movement/rotation of the face), head movement, FACS facial action units (based on the implementation of Baltrusaitis, Mahmoud, and Robinson (2015)), and eye gaze tracking (based on the implementation of Wood et al. (2015)).

11.1.3 SIGNAL ESTIMATOR

The signal estimator component works similarly to the face analyzer, however, it uses input video frames and located facial landmarks to estimate physiological signals, e.g. HR. It contains different estimators, each one responsible for estimating a single signal. There are two signal estimators available, both estimate HR by using different techniques. These estimators use ICA-based rPPG techniques (Poh, McDuff, and Picard, 2010; Poh, McDuff, and Picard, 2011) implemented in Matlab. Figure 11.5 shows the user interface regarding the data provided by the signal estimator. Available information provided by the signal estimator include the photoplethysmographic signal, estimated HR, and the ROI used in the estimation.

11.1.4 REPORT MANAGER

The report component aggregates the information produced by other components, generating a CSV report file as output. The report files contain information regarding the video, e.g. time, as well as estimated signals and extracted facial activity. The file generated by the report manager is used to exchange information between the software and third-party systems, e.g. software for statistical analysis. The report file is also used by the two components related to emotion modeling and estimation, as explained in the next section.
11.1.5 EMOTION MODEL AND ESTIMATOR

Finally the emotion model and the emotion estimator components are responsible for training the user-tailored model and applying it, respectively. Both components were developed in R using the caret package for machine learning. The components work on the report file produced by the report manager.
CHAPTER 12
ETHICS AND PRIVACY

Several contributions of the research presented in this thesis have implications that concern ethics and privacy. Technology has become an essential part of modern life; therefore, the advances it brings to different fields, including human-computer interaction and games research, should be guided by ethics and privacy. This chapter provides an overview and discussion regarding how the research and the technology proposed in this thesis touches issues related to both ethics and privacy.

12.1 ETHICAL USE OF TECHNOLOGY

Mason (1995) mentions that the facts of an ethical situation can be summarized by four factors. The first is the identification of the moral agent, which is the one causing the technology-induced change. The second relates to the available courses of action that the moral agent can undertake. It is not always possible, or even viable, to choose more than one course of action. Consequently, it must be selected according to the best interests of all parties involved. Additionally, a course of action is bound to have consequences, which can be irreversible. In that light, the third factor to emerge is the delimitation of the results that are expected to occur if each act is taken. A proper delimitation of results makes it clear for the involved parties how to measure the impact and implications of an act. Finally, the fourth factor is the identification of the stakeholders who will be affected by the consequences of the acts.

One of the main goals of the technology developed in this thesis is a non-obtrusive form of emotion detection. Given that a person has agreed to have a user-tailored model of him/herself created, i.e. play the calibration games while being filmed, any moral agent, i.e. researcher or company, is then able to use such data freely and unrestrictedly. After the model has been trained, the person used to train said model can be indefinitely surveyed in a context of gaming. Once trained, the model can be easily transferred to another moral agent, e.g. another institution or company, and used at a later time. Even though the proposed method is constrained by a gaming context, it can still be widely used. If the person in question, who is the stakeholder of the process, is not properly and clearly informed about the identity of the moral agents and the delineation of the results expected from the use of his/her model, an ethical issue may exist.

An ethical issue is said to arise whenever one party in pursuit of its goals engages in behavior that materially affects the ability of another party to pursue its goals (Mason, 1995). One could claim that sharing a person’s user-tailored model among institutions/companies is not materially affecting that person. Additionally, people are more prepared to accept potentially invasive technology if they consider that its benefits outweigh potential risks (Ladd, 1991). However, one of the moral agents might be a game development company that uses the model to detect the emotional state of a person, in order to maximize the sale of in-game goods. In that case, the act could materially affect the person, which would clearly be an ethical issue if the person was never made aware of
such a possible use of his/her model. As previously mentioned, the facts of an ethical situation must be clear, otherwise obscure information about courses of action, delimitation of results and even the identity of the moral agents could lead stakeholders into making poor judgments regarding ethics and privacy.

Another implication of non-obtrusive technologies is how it influences the ability of a user to decline the propagation of any information. In the context of games research, for instance, if a subject answers a questionnaire about a game being played, it is completely plausible to assume that the subject could deliberately lie about the answers. Subjects might even decline to answer a particular question about emotions, if they feel uncomfortable, for instance. If the method proposed by this thesis is used to detect emotional states and one assumes that the subjects have previously agreed to the creation of a user-tailored model of themselves, then they do not have the option to decline to answer a query about emotions. A researcher might have a previously trained model of a subject, e.g. from an old experiment, which can be used again for the same subject, however in a different context. The method proposed in this thesis can be adapted to be trained on data from a group instead of an individual, i.e. group model instead of a user-tailored model. In that case, the trained model could be applied to any person (or subject), without the need for them to play the calibration games. It is plausible to believe that such a configuration of the method could be used by companies to survey a player’s emotional responses to a particular game. A company could, for instance, apply the method to online videos, e.g. “Let’s play” videos on YouTube, to gather unsolicited emotional data. If the videos are freely available, does it mean such use of the method is ethical? Was the person in the video thinking about having his/her emotions automatically detected by a software when he/she made the video?

The technology proposed by this thesis has several limitations and constrains, however, it can be extended and improved to broaden its accuracy and usage. It has moral and ethical implications that should be discussed by all stakeholders involved, making the facts of any ethical situation completely clear and understood by all affected.

12.2 PRIVACY AND PERSONAL DATA

Discussions about privacy in the field of human-computer interaction are common and there is a clear indication that HCI tools must not invade a user’s privacy (Pantic and Rothkrantz, 2003). Any tool’s capacity to monitor and concentrate information about somebody’s behavior must not be misused. The technology presented in this thesis significantly relates to privacy. As defined by Culnan (2000), privacy is the ability of the individual to control the terms under which personal information is acquired and used. When a system collects personal information, which is the case of the method in this thesis, information privacy becomes an issue. E. F. Stone et al. (1983) define information privacy as the ability of the individual to personally control information about one’s self.

In the context of this thesis, information privacy relates to how a person controls the digital data collected from him/herself, e.g. video recordings and the user-tailored model. As previously mentioned, the technology presented in this thesis has moral and ethical implications, which leads to information privacy implications. Privacy is extremely contextual, based in the specifics of by whom, for what, where, why, and when a system is being used (Ackerman and Mainwaring, 2005). In that sense, individuals monitored and analyzed by the technology presented in this thesis might have divergent opinions regarding information privacy. Some individuals might believe the use of such technology is beneficial and could be used to enhance their gaming experience, for instance.
On the other hand, some individuals might oppose the use of such technology, due to concerns about information privacy. Awad and Krishnan (2006) show that consumers using online shopping websites who desire greater information transparency are less willing to be profiled. In contrast, users that do want a more personalized experience when shopping online are more willing to be profiled. One possible solution for such a problem, which is a recommendation by Awad and Krishnan (2006), is the utilization of mechanisms that account for both types of clients, the ones willing to be profiled to increase service personalization and those that are not.

The method proposed in this thesis is based on the analysis of video recordings. Normally, users are not concerned about a video recording beyond the issue of the usage of their personal image. The present research, however, uses several techniques to collect additional data from those video recordings, including facial analysis and HR information. When being filmed during the interaction with a set of calibration games, a person might not be aware of the amount of information that is actually being collected. How such data are stored, processed and used is a matter of information privacy. As previously mentioned, people are more prepared to accept potentially invasive technology if they consider that its benefits outweigh potential risks (Ladd, 1991). Users constantly decide and account for the trade-off between the benefit of a solution and the privacy implications of such act. Nguyen, Rosoff, and John (2016) show that when privacy was evaluated against usability, convenience, and speed, the concern for privacy was relatively high. However, when compared to cost, concern for privacy was relatively low. This suggests that people have a clear trade-off between price/cost and privacy. As Awad and Krishnan (2006) demonstrate, some consumers of online shopping are willing to give personal information in exchange for better services. Consequentially, users might be willing to be filmed and analyzed by a software, to have a personalized emotion detector to improve game experience, for instance, specially if they see benefits in any existing trade-off analysis taking place. What needs to be clear to those users, however, is the kind of data being collected, by whom and how it will be used.

Technology is not neutral, when it comes to privacy, and it can increase or reduce the extent to which people have control over personal data (Bellotti and Sellen, 1993). The technology presented in this thesis, if used in misleading ways, contributes to reducing the control people have over personal data. Ensuring users know what is happening, which can be achieved with clear information practices, is paramount. Langheinrich (2001) reveals the principles that guide system design, based on a set of fair information practices common in most privacy legislation in use. The author highlights the following principles: notice, choice and consent, proximity and locality, anonymity and pseudonymity, security, and access and recourse. The principles of notice, choice and consent are essential to the technology presented here. Making users notice what is happening and what is being collected, as well as allowing choice and consent of the process, is the bare minimum to ensure privacy.

12.3 ETHICAL AND PRIVACY IMPLICATIONS OF THIS RESEARCH

One can both discuss and speculate about the ethical and privacy implications of the technology developed and presented in this thesis. Sections 12.1 and 12.2 present a more formal and academic view of ethics, privacy and technology. This section presents a less formal and more personal discussion regarding the matter.

The technology presented in this thesis can help researchers and game developers to pro-
duce better games by tailoring experiences at a user level. The main goal is to help those actors and enhance the user experience, which will, however, inevitably result in several ethical issues. The quick pace in which technology progresses produces ever smaller devices for data collection, for instance, cameras with infrared and depth sensors that capture data in the living room of players. The remote acquisition of physiological data applied to emotion detection, such as the method described in this thesis, allows dangerously easy access to the psychological profile of users. Even though companies and researchers mean no harm to subjects and customers, the unethical use of any technology is facilitated when the process is completely remote and unobtrusive. A significant amount of personal information can now be acquired remotely from any users in any location, including HR and facial activity. Such information will certainly be used for a wide variety of aims, including personal and commercial gain. Differences in culture might play a big role in the acceptance and adoption of such technologies. In a society where the individual and limited regulation are valued, a technology that promises a tailored experience will likely be accepted and not seen as an invasion of privacy. However, the misuse of such technology might not even be anticipated or understood by all. For instance, it is quite common to accept that a particular interaction with an entity is being recorded, e.g. a phone call to a customer service center. What are the consequences of video recording a given meeting and later using the video data to infer emotional states? Is it fair or ethical to use such technology during a job interview, to measure the reactions of a candidate on the basis of his/her emotional state when salaries are discussed, in order to reach an amount that maximizes the company’s profit? The emotional exploitation of individuals becomes a plausible idea when no sensors need to be used to detect their emotional state, only an ordinary camera.

The measurement of players’ emotions in a game could lead to pleasant and personalized experiences. However, it could also lead to issues involving induced behavior to enforce malicious behavior, for instance. A company selling items can significantly benefit from any information regarding the emotional state of a user. If a particular user is more likely to spend money on game items or expansions when he/she is stressed, then the game can work to keep such a player stressed more than other players. As previously mentioned, users are likely to accept a technology they regard as beneficial. The issue rises when users do not recognize the problems with a technology or the negative effects it might have. Consequentially, users could be tricked into believing that a technology is good and works on their behalf, when the situation or the company’s intentions could be the complete opposite. In a world where all actions are guided by good will, there would be no need for reflections regarding the technology presented in this thesis. However, good intentions are not always clear to different actors, therefore, the technology presented in this thesis must be thoroughly discussed, regulated and monitored. Technological progress without regulations could lead to the mass exploration of users without their consent. New technologies, especially those using the remote acquisition of psychophysiological signals aimed for emotion detection, allow a new paradigm of data collection. Users might not even be aware of what can be remotely collected from them and used to infer emotions and induce a certain behavior. It is possible to believe that users could be acting in a certain way, in the belief they are following their own free will, when in reality they are being manipulated on the basis of their psychophysiological reactions. This is a situation that should never happen during the ethical uses of any technology.

My personal opinion as a researcher regarding the work described in this thesis is that its benefits outweigh its problems. New tools and experiences can be created from a better understanding of what the user is feeling during the interaction with a system. If used with privacy and ethics in mind, the technology proposed here can help the development of brand new categories of software and hardware that are not disconnected from their
users emotions. Instead, they are aware of them and can work on the behalf of users in that regard. Regulations and transparency for data collection and its use are paramount to ensure the ideal use of this new technology.
CHAPTER 13
LIMITATIONS AND CRITIQUE

One potential limitation of the work presented in this thesis is the nature of the calibration games. Even though they serve the purpose of emotion elicitation materials, they were designed and developed as ordinary games. Along the process, several decisions were made concerning different aspects of each game, which inevitably affected the end result. These decisions had an impact, for instance, on the genre of each calibration game, as well as its graphical appearance and the level of complexity of the mechanics. A calibration game should induce a state of boredom at the beginning of the interaction, thus users should easily understand its mechanics in order to perceive the game as boring without a long exposure. It entails that the game mechanics must be easily understandable, preferably without much text or tutorials. Users should not spend a considerable amount of time learning the game, otherwise the concepts that induce boredom might be misunderstood and the desired emotional state would not be induced. Additionally, all calibration games should not allow users to deliberately control the mechanics’ pace, since it was a key factor that was automatically controlled to induce stress towards the end of the session. Those constraints led the design of the calibration games towards more casual, 2D game mechanics. Even though games with similar characteristics exist, the proposed calibration games lack 3D content or a more complex interaction similar to those found in AAA COTS games, for instance. The genre/mechanics selected for the calibration games likely hinder several other genres and mechanics that could potentially be used as calibration games as well. The nature of the calibration games presented in this thesis does not cover the wide range of possible game types that exists, which limits its reach.

In that light, it could be argued that the calibration games proposed in this research only induced emotions elicited from the specific genres/mechanics that were selected. The use of 2D, casual foundations for the calibration games, could have conveyed a message of “old games” to a segment of subjects/users, which would likely impact their emotional reactions. On the other hand, the use of a more complex 3D game with sophisticated mechanics, e.g. Counter Strike, is likely to require a certain level of gaming skills from participants. In such a case, it would impact the interactions of subjects that are not very familiar with gaming, which was the case for some participants in the heterogeneous groups presented in this research. As mentioned previously, individuals have different cultural views and expectations, thus, a more complex game would make it even harder to balance the design of a game with its intended purpose of inducing boredom and stress. The game Infinite Mario was used in the validation process of the proposed method mainly due to its characteristic, e.g. easy to understand and play. Additionally, it allowed more control over the content generation associated with its mechanics, therefore, boring and stressful levels could be easily developed for the experiment mentioned in the thesis. It is plausible that the emotion classification results obtained with Infinite Mario could be generalized to similar games, especially since Super Mario influenced a range of platformer games. However, as previously mentioned, the use of another 2D, casual game for the validation could limit the generalization of the results.
Another limitation of this research concerns the accuracy obtained by the method in the classification of emotional states of boredom and stress. As presented in Chapters 9 and 10, the method achieved an accuracy of 61.6%. Even though this classification rate has statistical significance that proves the method performs better than random guessing, such a performance is still too low for commercial or even academic use. In its current state, the proposed method could not be used as the only tool to detect the emotional state of users, due to its noise. Additional measurements should accompany the proposed method to ensure a proper evaluation of the emotional context of subjects/users, e.g. questionnaires. However, the proposed method could still be used as an insight mechanism to analyze large amounts of video footage in an automated way, for instance. Despite the best efforts invested in this research to design an accurate emotion detector, the complexity of the task and the amount of man-power available limited the exploration process. Instead of aiming for a perfect tool, the research presented in this thesis focused on designing and rigorously evaluating each part of the proposed method. Such an approach is expected to eventually guide the construction of a more sophisticated emotion detector in the future.

It is important to highlight the technical limitations associated with the remote acquisition of physiological signals. The rPPG technique used in this research, as detailed in Section 8.7, is appropriate to deal with the natural behavior that users exhibit during their interaction with games. However, this technique was likely affected by other factors not scrutinized by the research in this thesis. For instance, the 15 seconds long duration of each analysis window used for the estimation of HR may have affected the results. The ideal length of the window (called window size) is not agreed upon in the literature (Rouast et al., 2016). In general, it depends on the characteristics of the rPPG technique being applied as well as the hardware configuration, such as camera framerate (Roald, 2013). The statistical nature of ICA, part of the selected rPPG employed in this research, demands longer video samples to produce accurate results. The longer the video, however, the higher the chances of subject motion, which increases noise. A trade-off between the duration of the video segments and the estimation accuracy could be better investigated. Another factor is that the experimental setup used an external light source to minimize noise caused by changes in illumination, which should narrow the estimation error to causes as subject movement and/or facial activity. Nonetheless, other factors probably could have impacted the estimation accuracy, such as facial hair, e.g. beard and fringe, use of glasses, and skin color. The results obtained with this research were achieved in a laboratory-like environment with controlled light source, which limits the generalization of the conclusions. As detailed in Chapter 6, a subject’s movement and changes in illumination are significant challenges to the estimation accuracy of rPPG techniques. The use of a controlled light source, however, was deemed necessary to concentrate efforts on the remote detection of the emotional state, not on the noise caused by different illumination patterns.

Finally, a significant limitation is that the method proposed in this thesis can only detect two emotional states, i.e. stress and boredom. As detailed in Chapter 3, there are different models and theories about emotions. Previous works focused on emotion detection commonly classify the six basic emotions proposed by Ekman and Friesen (1971), i.e. happiness, surprise, sadness, fear, anger and disgust. There are also a significant number of works that measure emotional states in terms of Russell’s Arousal-Valence space (Russell, 1978). The method described in this thesis relies on the emotion elicitation provoked by the concept of calibration games, which are designed to be user-tailored materials that account for the differences among users in the emotion elicitation process. These games are, by design, limited to inducing only boredom and stress. It is plausible that other emotions are also elicited by the calibration games, e.g. happiness.
and anger. However, the very idea of constantly and endlessly increasing the difficulty of the games to account for the subjects’ individualities in the emotion elicitation process, e.g. different gaming skills and cultural expectations, directly limits the emotions that can be reasonably tracked without interrupting the gameplay. As demonstrated by the statistical tests performed in the studies conducted on the data collected from experiment 1, the emotional state of subjects is assumed to be boredom at the beginning and stress at the end of the calibration games. Inducing a state of happiness in a player, for instance, is a complex task that depends on several components, including cultural factors and gaming preferences. The same reasoning can be extended to other emotions, such as anger, fear and disgust. Even for commercial games, significant resources are invested to ensure a game is properly balanced to be able to please a wider range of users, however, there is still no guarantee that this will happen. Consequentially, the focus put on detecting only boredom and stress in the method presented in this thesis is a constraint created to counterbalance the differences that exist among subjects, regarding their different gaming, cultural and emotional profiles. However, it is important to mention that emotional states of stress and boredom are still relevant to the industry or studies in HCI and game research. Flow theory is commonly used in game research to model players emotions, which is directly connected to stress and boredom. These emotional states can be described as a function of a current player’s skill level and the level of challenge he/she faces in the game. Such important emotions can help both researchers and game designers in the study of the interaction between players and games.
CHAPTER 14

CONCLUSION

Questionnaires and physiological measurements are the most common approaches used to obtain data for emotion estimation in the field of HCI and games research. Both approaches interfere with the natural behavior of users, which affects any research procedure. Initiatives based on computer vision and the remote extraction of user signals for emotion estimation exist, however they are limited. Experiments of such initiatives have been performed under extremely controlled situations with few game-related stimuli. In those experiments, users had a passive role with limited possibilities for interaction or emotional involvement, compared to game-based emotion stimuli, where users take an active role in the process, making decisions and directly interacting with the media. Previous works also focus on predictive models based on a group perspective. As a consequence, a model is usually trained from the data of several users, which in practice describes the average behavior of the group, excluding or diluting the key individualities of each user.

In that light, there is a lack of initiatives focusing on non-obtrusive, user-tailored emotion detection models, in particular regarding stress and boredom, within the context of games research that is based on emotion data generated from game stimuli. This thesis aims to fill that gap, providing the HCI and the games research community with an emotion detection process that can be used to remotely study users emotions in a non-obtrusive way within the context of games.

14.1 FULFILLMENT OF RESEARCH OBJECTIVES

As detailed in Section 1.2, a set of research objectives to support the overall aim of this thesis has been identified. Each one of these objectives is detailed below, along with the conclusion reached from their fulfillment.

O1: identification of the main concepts, theories and signals associated with the psychophysiological profile of users and their emotions within the field of HCI, particularly regarding games research. The outcome of this objective is a definition of stress and boredom within the context of this research, as well as the identification of the psychophysiological signals that are commonly applied to emotion detection.

The literature review detailed in Chapters 3, 4, and 5 presents the theoretical background related to psychophysiological signals and emotions. Different theories for the modeling and study of emotions in a variety of contexts, including those related to games, have been proposed. For this thesis, focus has been given to those theories based on human physiology connecting emotional reactions to psychophysiological signals, e.g. HR and facial activity.

O2: identification of existing computer vision techniques that can be employed to remotely extract the identified psychophysiological signals of users via the analysis of videos. The investigation includes the analysis
of how existing techniques are being applied to emotion detection. The set of signals to be remotely extracted is based on the results of objective O1.

The remote sensing of psychophysiological signals is an essential part of this research. Chapter 6 details the progress that has been made in the remote estimation of physiological signals, particularly the use of rPPG to estimate HR. The rPPG technique, proposed by Poh, McDuff, and Picard (2011), has been selected as the most appropriate for the remote estimation of HR in the context of this thesis. The selection is motivated by the statistical nature of ICA, an important component of the rPPG technique. The use of signal filtering via ICA allows the method to better deal with noise caused by motion, which is common in a context involving games and natural behavior. Additionally, several computer vision techniques for facial detection have been studied and a novel method for the automated analysis of facial cues from videos has been developed, as explained in Section 8.8. The empirical results of this method demonstrate its potential for detecting stress and boredom in players of games. The method is based on Euclidean distances between automatically detected facial points, designed to be robust enough to correctly perform facial analysis even when users naturally interact with games. Additionally, the analysis of user behavior, which was focused on facial actions, indicates that a neutral face remains for longer periods of time during boring periods. Finally, for the context of this thesis, facial analysis at an individual level produced more information that connects facial activity to emotional states of stress and boredom.

O3: investigation of the feasibility, accuracy and challenges of applying the identified computer vision techniques, regarding the extraction of the signals, within the context of computer games. This objective also encompasses the analysis of the behavior of players during gaming sessions and how it affects the technique.

Extensive evaluations were conducted to establish the reliability of remote HR measurements. The analysis of the accuracy of remote HR estimations clearly established the limitations of the rPPG technique, showing how it is affected by user behavior. Evaluation results of the rPPG technique, as detailed in Section 8.7, have shown the average estimation error of the technique in the context of this research. The error, however, lies within the range that still allows the identification of HR variations caused by emotion elicitation materials, as detailed in Section 8.6. The evaluations of the identified computer vision techniques have shown that it is feasible to remotely extract HR and facial data from video recordings of users interacting with games for the purpose of classifying emotional states.

O4: investigation and validation of the concept of a game-based calibration phase as an emotion elicitation source able to provide data to fit a user-tailored predictive model. The result of this objective is to design and validate a set of calibration games that can trigger the emotional responses required for the analysis of the remotely obtained signals and detection of boredom/stress levels by the model.

This research introduces the novel idea of calibration games. As detailed in Section 8.4, calibration games are carefully designed to have a difficulty level that constantly and linearly progresses over time without a pre-defined stopping point. This design of an emotion elicitation material takes into account the different expectations and gaming skills of a wide range of users, making the process more focused on the individual rather than the group. Sections 8.5 and 8.6 present a detailed analysis regarding how the effects of psychophysiological activity, i.e. HR and facial actions, relate to emotional states in a context featuring calibration games. The results indicate that a calibration game is a
valid emotion elicitation material which indeed induces emotional states of stress and boredom.

**O5:** proposal of a user-tailored, multifactorial model that uses the identified physiological and non-physiological signals, the computer vision technique and the calibration data to detect the current stress and boredom levels of a person while he/she plays any video game.

The knowledge obtained from the investigation of the previously mentioned objectives culminated in the final design of the proposed method for emotion detection. Such method, which is non-obtrusive, user-tailored and game-based, was evaluated in the first experiment, as detailed in Section 8.9. Despite the small sample size of this first experiment, the results suggest the feasibility of a user-tailored, multifactorial model to detect emotional states of boredom and stress.

**O6:** experimental validation of the proposed emotion detection process through an experiment involving a commercial off-the-shelf game.

Finally, the proposed method was validated in a second experiment using a larger sample size, as detailed in Chapter 9. The game Infinite Mario, similar to the commercial off-the-shelf game Super Mario, was used in the process. The results show that the proposed method was able to identify the emotional state of subjects with a mean accuracy of 61.6%.

### 14.2 ANSWERING THE RESEARCH QUESTION

The fulfillment of the previously mentioned research objectives culminated in the answer to the research question that guided this thesis. As presented in Chapter 1, the research question is:

> "How can the emotional state of players during the interaction with games be remotely detected on a user-tailored basis with the utilization of an ordinary camera and games as emotion elicitation sources for calibration?"

The process of fulfilling each research objective and consequentially answering the proposed research question involved a series of systematic evaluations conducted to understand the relation between psychophysiological signals and emotions. Based on a literature review, facial behavior and physiological signals, i.e. HR, were selected as indicators of the emotional state. The results of the research presented in this thesis show that individualities can be detected regarding facial activity, e.g. the increased number of facial actions during the stressful part of games. Regarding physiological signals, the findings are aligned with and reinforce previous research that indicates a higher HR mean during stressful situations in a gaming context. The results also suggest that changes in HR during gaming sessions are a promising indicator of stress.

All previously mentioned research objectives and the findings related to them culminated in the answer to the research question: the creation of a non-obtrusive, user-tailored and game-based method for emotion detection. The approach uses remotely acquired signals, namely HR and facial actions, to create a user-tailored model, i.e. trained neural network, able to detect emotional states of boredom and stress of a given subject.
The approach is composed of two phases: training (or calibration) and testing. In the training phase, the model is trained using a user-tailored approach, i.e. data from subject $S_a$ playing 3 calibration games (Mushroom, Platformer and Tetris) are used to create model $N_a$. The calibration games are a novel emotion elicitation material introduced by this research. These games are carefully designed to present a difficulty level that constantly and linearly progresses over time without a pre-defined stopping point, inducing emotional states of boredom and stress. The result of the training phase is a user-tailored model, i.e. model $N_a$, which is a trained neural network aimed for use on subject $S_a$. Finally, the testing phase is conducted in a game session involving subject $S_a$ playing any ordinary, non-calibration game, e.g. Super Mario. During the testing phase, the signals of subject’s $S_a$ are remotely acquired and fed into the previously trained model $N_a$, which then outputs the estimated emotional state of subject $S_a$ for that particular testing game.

The feasibility of the proposed method was evaluated in two distinct experiments. In the final evaluation of the method, the previously mentioned calibration games, i.e. Mushroom, Platformer and Tetris, were used as emotion elicitation materials to train a user-tailored model, i.e. neural network. This model was then used to detect the emotional state of each user during the interaction with a fourth game, i.e. Infinite Mario. The proposed method was able to identify the emotional state of subjects with a mean accuracy of 61.6%. The results confirmed with statistical significance that the proposed method indeed classified emotional states, achieving an accuracy rate better than chance-level classification.

14.3 CLOSING REMARKS

The proposed method for the remote detection of emotions has been conceived on the basis of established theories and it has been carefully evaluated in experimental setups. As mentioned in Chapter 10, the process of detecting the emotions of users is a complex task that involves theories and contributions from different fields. The results presented in this thesis prove the method is feasible, however, existing limitations prevent its wider use by researchers and companies, in its current configuration. Nevertheless it is a solid initiative to move away from questionnaires and physical sensors into a non-obtrusive, remote-based solution for evaluation of user emotions.
CHAPTER 15
FUTURE WORK

This thesis presents the conception, design and evaluation of several elements that are orchestrated to produce a non-obtrusive, user-tailored game-based emotion detector. Due to time and resource constraints, several courses of action were selected in favor of others. They can be further investigated to improve the proposed method or to better understand the relationship between psychophysiological signals and emotions. This chapter describes possible ideas for future work that can extend the foundations laid by this thesis.

Initially, further research could be invested on the concept of calibration games. In this thesis, only three of those emotion elicitation materials were developed. As previously mentioned, they were 2D, casual games with particular genres and mechanics. Different types of calibration games could be explored, including 3D variations in different genres, e.g. first person shooters (FPS) or strategy games. Additionally the existing calibration games proposed along with the method, i.e. Mushroom, Platformer and Tetris, could be refined and better investigated. During the debriefing sessions that followed the second experiment, several subjects mentioned their impressions regarding the calibration games. Some participants, for instance, highlighted the speed at which the difficulty of some games increased, e.g. Tetris. A fast increase of the difficulty level is not part of the design of any calibration game, since it is likely to induce stress in the subject in a short time period which might be insufficient for remote analysis, i.e. signals acquisition. The duration of each calibration game could also be further investigated. On average, subjects spent 6.4, 4.7 and 5.8 minutes playing the Mushroom, Platformer and Tetris game, respectively. No investigation was conducted regarding the ideal duration of a calibration game. Short calibration games allow quick data collection, a desirable quality for the speedy production of the user-tailored model. It also mitigates effects related to a subject’s fatigue or emotion recall when answering the questionnaires about stress/boredom at the end of each game.

The acquisition of psychophysiological signals could also be improved on many fronts. For this thesis, only two signals were used, i.e. HR and facial actions. Even though the former produces different information about the face, e.g. eyebrow, eye and mouth activity, more signals could be investigated. The literature review presented in this thesis found several signals that could be acquired in a remote fashion and used for emotion detection. The most notorious of those signals is HRV, which is widely mentioned as an indicator of stress. Another possible signal is related to the eye, including eye tracking, saccade time and blinking. Even though eye tracking is strongly connected to the activity at hand, e.g. movement pattern of the eyes is highly correlated with the game mechanics being played, while blinking and eye saccade time are less influenced by the game mechanics. Consequently, they could easily be integrated and used as emotional indicators. However, the addition of a new signal to the proposed method requires careful evaluations and adaptations. Any new signal needs to be evaluated in the context of emotion elicitation, i.e. calibration games. How such a signal changes in face of induced emotional states produced by calibration games is a key aspect to be understood before
it can be added to the proposed method. Following such investigation, an analysis regarding how the signal is affected by a user’s natural behavior is also needed. It would establish the accuracy level of the signal’s acquisition, such as the evaluation of rPPG estimations of HR presented in this thesis.

The addition of any new signal to the proposed method is also intrinsically connected to further research on the machine learning techniques used to create the user-tailored model. In this thesis, neural networks trained using random search were used, however different methods could be employed. The literature mentions the use of SVM and many more machine learning models. Further research could identify better machine learning models, possibly different techniques for different users, maximizing the idea of user-tailoring the process of detecting emotions. Another possible research idea is to explore how a home environment, e.g. living room, affects a user’s natural behavior and how it impacts the remote estimations of the signals. As described in the thesis, all estimations of HR, for instance, were performed with an external light source, which is an unlikely home setup. A living room containing a game console could be dark, significantly impacting the usefulness of the proposed method. Further investigations of that topic could highlight the limitations of rPPG techniques when applied in a home environment, for instance, and how those limitations could be mitigated.

Another topic for further exploration is the difference between a user-tailored and a group model. The method proposed in this thesis advocates a user centered design, where individualities of participants are likely to be preserved. In this thesis, little research was invested on the use of a group approach, where data from a group of subjects are used to produce the model. A group-oriented design allows the method to be trained once and used on a variety of different users without them having to play the calibration games. Further research on that front would highlight how efficient a user-tailored approach actually is compared to a group approach. One example of such an investigation is the middle ground between a user- and a group-tailored approach. In such an approach, calibration games are used as emotion elicitation material to enhance the readings of individual traits when training a group model. Calibration games are designed to elicit particular emotional states at an individual level, it is therefore expected that a group model trained on several of those enhanced individual manifestations will produce a better emotion classifier. Works in the literature commonly use the same emotion stimuli on all subjects, e.g. an invariant image or game, consequentially subjects may perceive that in an emotionally differently way, e.g. some subjects might not be affected at all. These subjects might not contribute to the training of a group model, because their emotional perception was not properly captured. Observing and capturing such individualities is precisely the aim of a calibration game, which could help produce better group models.

Finally, this research gathered a significant amount of data related to games, psychophysiological signals and emotions from a heterogeneous group of subjects. This data range from HR information, i.e. acquired with a physical sensor, to in-game actions, e.g. jumps in Infinite Mario and movements in Tetris. Further analysis could be performed on such data to better understand the relation between in-game actions and emotions based on psychophysiological signals. Previous studies focused on relating facial activity to emotional states during interactions with Infinite Mario (Shaker, Yannakakis, and Togelius, 2011). Such analysis could be further researched with the addition of physiological data, for instance. Additionally, the data collected by this research can be segmented according to the gaming profile of subjects, e.g. categorize subjects according to reported gaming skills or weekly hours devoted to gaming. This could produce insights regarding the differences between casual and “hardcore” players in relation to changes in psychophys-
iological signals, accuracy of remotely detecting emotional states, body behavior and self-reported levels of stress and boredom.


IEEE International Conference on Computer Vision Workshops. IEEE, pp. 354–361. DOI: 10.1109/iccvw.2013.54.


Boussefa, Frédéric, Choubeila Maaoui, and Alain Pruski (2013a). “Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate.” In: Biomedical Signal Processing and Control 8.6, pp. 568–574. DOI: 10.1016/j.bspc.2013.05.010.

Datcu, Dragos et al. (2013). “Noncontact automatic heart rate analysis in visible spectrum by specific face regions.” In: Proceedings of the 14th International Conference


Garde, Anne et al. (2002). “Effects of mental and physical demands on heart rate variability during computer work.” In: *European journal of applied physiology* 87.4-5, pp. 456–461.


Haan, Gerard de and Vincent Jeanne (2013). “Robust Pulse Rate From Chrominance-Based rPPG.” In: *IEEE Transactions on Biomedical Engineering* 60.10, pp. 2878–2886. DOI: 10.1109/tbme.2013.2266196.


Koster, Raph (2013). Theory of fun for game design. ” O’Reilly Media, Inc.”.


Tran, Duc Nhan, Hyukzae Lee, and Changick Kim (2015). “A robust real time system for remote heart rate measurement via camera.” In: 2015 IEEE International Conference on Multimedia and Expo (ICME). IEEE. Institute of Electrical and Electronics Engineers (IEEE), pp. 1–6. DOI: 10.1109/icme.2015.7177484.


PUBLICATIONS IN THE DISSERTATION SERIES


Laxhammar, Rikard (2014). *Conformal anomaly detection: Detecting abnormal trajectories in surveillance applications.*


Holgersson, Jesper (2014). *User Participation In Public e-service Development: Guidelines for including external users.*


