,

Assessment of Fun from the Analysis of Facial Images

Luiz Carlos Vieira

PhD thesis presented
to the
Institute of Mathematics and Statistics
of the
University of São Paulo
as a
partial requirement for the degree
of
Doctor in Sciences

Program: Computer Science
Advisor: Professor Flávio Soareas Corrêa da Silva, PhD

The author has received financial support from CAPES during the execution of this work.

São Paulo, July 2017

Assessment of Fun from the Analysis of Facial Images

This is the version of the thesis with the corrections and changes suggested by the Judging Commission during the defence of the original version of the work, held on 16/05/2017. A copy of the original version is available at the Institute of Mathematics and Statistics of the University of São Paulo.

Judging Commission:

- Prof. Dr. Flávio Soares Corrêa da Silva (advisor) – IME-USP
- Prof. Dr. Ricardo Nakamura – EP-USP
- Prof. Dr. Silvia Carla Rodrigues – UFABC
- Prof. Dr. Cláudia Josimar Abrahão de Araújo (UFABC)
- Prof. Dr. Sara Lúcia Manzoni (Univ. Milano-Bicocca)

Acknowledgements

First of all, I would like to thank my family and friends for supporting my decision to pursue such a difficult and demanding goal. I would also like to thank my advisor, professor Flávio Soares Corrêa da Silva, for all the help, advices and friendship that he provided me during the execution of this work. I thank my professors and colleagues at the University of São Paulo, particularly my dearest peers from the Laboratory of Interactivity and Digital Entertainment Technology (LIDET) for their support and sharing of knowledge, and the professors from the Judging Commission for their commentaries and suggestions regarding this work. Finally, I thank the volunteers that have agreed to play the games in the experiment performed in this work, the Committee of Ethics in Research, and CAPES (*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*) for their financial support during the execution of this work.

This work is dedicated to my lovely family: my wife Laura and my infant son Bruno, who was born just as I was finishing working on this thesis. I love you guys!

Abstract

VIEIRA, L. C. **Assessment of Fun from the Analysis of Facial Images**. 2017. 153 p. Thesis (PhD) - Institute of Mathematics and Statistics, Univerisity of São Paulo, São Paulo, 2017.

This work investigates the feasibility of assessing fun from only the computational analysis of facial images captured from low-cost webcams. The study and development was based on a set of videos recorded from the faces of voluntary participants as they played three different popular independent games (horror, action/platform and puzzle). The participants also self-reported on their levels of frustration, immersion and fun in a discrete range [0,4], and answered the reputed Game Experience Questionnaire (GEQ). The faces were found on the videos collected by a face tracking system, developed with existing implementations of the Viola-Jones algorithm for face detection and a variation of the Active Appearance Model (AAM) algorithm for tracking the facial landmarks. Fun was represented in terms of the prototypic emotions and the levels of frustration and immersion. The prototypic emotions were detected with a Support Vector Machine (SVM) trained from existing datasets, and the frustration, immersion and fun levels were detected with a Structured Perceptron trained from the collected data and the self reported levels of each affect, as well as estimations of the gradient of the distance between the face and the camera and the blink rate measured in blinks per minute. The evaluation was supported by a comparison of the self-reported levels of each affect and the answers to GEQ, and performed with measurements of precision and recall obtained in cross-validation tests. The frustration classifier could not obtain a precision above chance, mainly because the collected data didn't have enough variability in the reported levels of this affect. The immersion classifier obtained better precision particularly when trained with the estimated blink rate, with a median value of 0.42 and an Interquartile Range (IQR) varying from 0.12 to 0.73. The fun classifier, trained with the detected prototypic emotions and the reported levels of frustration and immersion, obtained the best precision scores, with a median of 0.58 and IQR varying from 0.28 to 0.84. All classifiers suffered from low recall, what was caused by difficulties in the tracking of landmarks and the fact that the emotion classifier was unbalanced due to existing datasets having more samples of neutral and happiness expressions. Nonetheless, a strong indication of the feasibility of assessing fun from recorded videos is in the pattern of variation of the levels predicted. Apart from the frustration classifier, the immersion and the fun classifier were able to predict the increases and decreases of the respective affect levels with an average error margin close to 1.

Keywords: Fun, Assessment, Emotions, Frustration, Immersion, Facial Expressions.

Resumo

VIEIRA, L. C. **Avaliação de Diversão a Partir da Análise de Imagens Faciais**. 2017. 153 p. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2017.

Este trabalho investiga a viabilidade de medir a diversão apenas a partir da análise computacional de imagens faciais capturadas de webcams de baixo custo. O estudo e desenvolvimento se baseou em videos gravados com as faces de voluntários enquanto jogavam três diferentes jogos populares e independentes (horror, ação/plataforma e puzzle). Os participantes também reportaram seus níveis de frustração, imersão e diversão no intervalo discreto [0, 4], e responderam ao renomado Game Experience Questionnaire (GEQ). Faces foram encontradas nos videos coletados utilizando um sistema desenvolvido com implementações existentes do algoritmo de Viola-Jones para a detecção da face e uma variação do algoritmo Active Appearance Model (AAM) para o rastreamento das marcas faciais. A diversão foi representada em termos das emoções prototípicas e dos níveis de frustração e imersão. As emoções prototípicas foram detectadas com uma Máquina de Vetores de Suporte (SVM) treinada com bases de dados existentes, e os níveis de frustração, imersão e diversão foram detectados com um Perceptron Estruturado treinado com os dados coletados e os níveis reportados de cada afeto, com o gradiente da distância entre a face e a câmera, e com a taxa de piscadas por minuto. A avaliação foi apoiada pela comparação dos níveis reportados com as respostas ao GEQ, e executada com métricas de precisão e revocação (recall) obtidas em testes de validação cruzada. O classificador de frustração não obteve uma precisão acima de chance, principalmente porque os dados coletados não tiveram variabilidade suficiente nos níveis reportados desse afeto. O classificador de imersão obteve uma precisão melhor particularmente quando treinado com a taxa de piscadas, com uma média de 0.42 e uma Amplitude Interquartil (IQR) entre 0.12 e 0.73. O classificador de diversão, treinado com as emoções prototípicas e os níveis reportados de frustração e imersão, obteve a melhor precisão, com média de 0.58 e IQR entre 0.28 e 0.84. Todos os classificadores sofreram de baixa revocação, causada por dificuldades no rastreamento das marcas faciais e pelo desbalanceamento do classificador de emoções, cujos dados de treinamento continham mais exemplos de expressões neutras e de felicidade. Ainda assim, um forte indicador da viabilidade de medir diversão a partir de vídeos está nos padrões de variação dos níveis previstos. Com exceção da frustração, os classificadores de imersão e de diversão foram capazes de prever os aumentos e reduções dos níveis dos respectivos afetos com uma margem de erro média próxima de 1.

Palavras-chave: Diversão, Avaliação, Emoções, Frustração, Imersão, Expressões Faciais.

Contents

List of Abbreviations

10k	10k US Adult Faces Database
AAM	Active Appearance Model
ANN	Artificial Neural Network
ANS	Autonomic Nervous System
AI	Artificial Intelligence
CK+	Extended Cohn-Kanade Dataset
CNS	Central Nervous System
CSV	Comma-Separated Files
DDA	Dynamic Difficulty Adjustment
FACS	Facial Action Coding System
Full HD	Full High Definition (1920 x 1080 pixels)
GEQ	Game Experience Questionnaire
GMM	Gaussian Mixture Model
HCI	Human-Computer Interaction
HD	High Definition (1280 x 720 pixels)
HMM	Hidden Markov Model
IAPS	International Affective Picture System
IQR	Interquartile Range
KNN	K-Nearest Neighbour
MDA	Mechanics, Dynamics and Aesthetics
ML	Machine Learning
MLP	Multilayer Perceptron
MPEG	Moving Picture Experts Group

NES	Neuroendocrine System
OBS	Open Broadcaster Software Studio
OvO	One-versus-One
OvR	One-versus-Rest
PANAS-X	Positive and Negative Affect Schedule - Expanded Form
PCA	Principal Component Analysis
PGCG	Procedural Content Generation
POSIT	Pose from Orthography and Scaling with Iterations
RBF	Radial Basis Function
SAM	Self-Assessment Manikin
SNS	Somatic Nervous System
STAI	State-Trait Anxiety Inventory
SVM	Support Vector Machine
UX	User Experience
YAML	YAML Ain't Markup Language

List of Symbols

θ	Angle orientation in radians of the sinusoidal carrier of a Gabor kernel
λ	Wavelength in pixels of the sinusoidal carrier of a Gabor kernel
σ	Standard deviation, used in the Gaussian envelope in a Gabor kernel and in SVM kernels such as RBF
γ	Aspect ratio of the Gaussian envelope in a Gabor kernel or the variance parameter used by the RBF kernel in a SVM
ψ	Offset in pixels of the sinusoidal carrier of a Gabor kernel
ξ	Slack variable used by a SVM to achieve soft margin classification
η	Rate used to control the amount of weight adjustment in the learning process of a Perceptron
Φ	Mapping function (which can be referring to the Kernel in a SVM or to the feature function in a Perceptron)
x, y	Feature vector or label of a sample used for training or prediction in a ML algorithm
w	Vector of weights learned or used by a linear classifier
b	Bias learned or used by a linear classifier
k, n, m, t	Length or number of elements, samples or steps in a vector, dataset or time-framed learning process
C	Regularization parameter of a SVM classifier

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Contextualization

Video games (also called digital, electronic or computer games) are undoubtedly one of the most important forms of modern entertainment. It is the industry sector that has experienced the biggest growth of the decade, reaching in 2015 more than 23 billion dollars of revenue just in the United States of America – the world's main market, where 63% of the households have at list one person who plays digital games (Entertainment Software Association, 2016, p.2,13). In the same year, the global consumer spending on that media surpassed 80 billion dollars and placed the video game industry among other important entertainment segments such as music (US$ 95 bn) and cinema (US$ 37 bn) (McKinsey & Company, 2015, p.8).

But this interest is not exclusive to modern humans. It is truly very old, as confirmed by many archaeological evidences available. Tapestries with people playing the game Nard (the precursor of modern Backgammon) and pieces and boards of the game Chaturanga (the precursor of modern Chess) indicate that they were respectively created in Persia and India between AD 500 and AD 800 (Bell, 2010, p.42;53). Boards of the game Nine Men's Morris (a game that shares the origins of Tic Tac Toe/Noughts and Crosses) found carved in limestone blocks in many places in Europe show that this game may have been created by Romans between 100 BC and AD 40 (Berger, 2004, p.15). Depictions in wall paintings (like the one in figure 1.1) and well-preserved boards and pieces confirm that the Egyptians Senet and The Royal Game of Ur – perhaps the most well known ancient games – were already played circa 3000 BC (Piccione, 1980, p.55; Bell, 2010, p.23). Games of the Mancala family are still very popular on a wide territory (Africa, Asia, Caribbean and America), what can only mean that they are indeed very old (Bikić and Vuković, 2010, p.192). As a matter of fact, limestone carved blocks resembling Mancala games were found in Africa and Western Asia and dated to the Neolithic Age (circa 5000 BC) (Rollefson, 1992, p.1;3).

This practise is as old as humankind because it comes from natural interests. The symbolism in the boards, pieces and rules of many of those ancient games is strongly related to religion, war and agriculture, all important aspects of early civilizations. Senet, which means "passing", simulated the stratagems of gods at the passage to afterlife (Piccione, 1980, p.55–56). Boards of Nine Men's Morris had mystical values attributed to the golden ratio properties between circle and square, and thus were used in architecture as well as in divination and worship (Berger, 2004, p.11;17). Chaturanga represented army branches (elephants, ships, horses, infantry and rajahs) and simulated the actions of war by means of different movements, sacrifice of pieces and promotion of pawns (Bell, 2010, p.52–54). Mancala games were also used in divination and still have religious meaning in some places, where they are played to amuse the spirit of the dead awaiting burial (Bell, 2010, p.122). Playing Mancala involves counting and collecting pieces (seeds) in hollows scooped into the board (earth) in a fashion very similar to sowing, what could have served as informal training of important skills required to the production and management of food resources (Bikić and Vuković, 2010, p.188).

Figure 1.1: *Wall painting depicting Egyptian queen Nefertari playing Senet in the afterlife. Reproduced from McDonald (1996, p.58).*

In that sense, playing is not merely a social or psychological trait since other animals also do it (Huizinga, 2008, p.3–4). Instead, it is an innate impulse that helps in experiencing and exercising real life preoccupations such as survival, competition, counting and social organization (Maitland, 2010).

The form by which such natural interests are represented in elements like stories, rules and pieces is what makes games so attractive. Those elements make games *very interesting* by producing pleasant cognitive and physical sensations. Playing is an intrinsically motivated activity that does not depend upon any visible external reward, like money or food (Malone, 1980a, p.3), meaning that people play only because they want to (Sweetser and Wyeth, 2005, p.1; Brown, 2010, p.1). This engagement (i.e. the will to play again) is driven by characteristics of the relationship that emerges from the user interacting with a game, like curiosity, fantasy and challenge (Malone, 1980a, p.162–166; Malone, 1980b, p.49–63), as well as concentration, self-awareness, sense of control, distortion of temporal experience and intrinsic rewards (Csikszentmihalyi, 1991, p.71–93). In fact, that is true not only for games, but also for any enjoyable activity: the satisfaction obtained in performing those tasks is related to the human natural desire to understand the world by interacting with it (Litman, 2005, p.794).

All the pleasure, excitement and enjoyment experienced playing with games is generally known as *fun*. It is an important aspect of life and, consequently, something that designers systematically desire to imbue their products with. Fun is a very obvious requirement to games, after all they are supposed to be fun otherwise people will simply not want to play them (Sweetser and Wyeth, 2005, p.1). But it is also very important for other products – that might be called "serious" (commercial and scientific software systems included). The satisfaction by interacting with a product, that used to be searched by Human-Computer Interaction (HCI) researchers just in terms of utility and absence of physical and cognitive discomfort, now includes other non-utilitarian aspects of the *user experience* related to appeal, preference and emotions (Forlizzi and Battarbee, 2004, p.263; Hassenzahl, 2005, p.31; Brown, 2010, p.74) - all, aspects of fun (Chen, 2007b, p.32). Indeed, the deployment technologies currently available (like on-line mobile markets, for instance) make very easy to have access to many similar products, particularly software systems. In cases in which users have several options to choose from, the use is discretionary or involves sustained activity, like that, easy of use and simplicity is not enough and developers must try and make users *want to* use their systems by making them more fun (Carroll, 2004, p.38).

1.2 Motivation

Fun is very difficult to design because it is an experience essentially made of emotions (Picard, 1995, p.10). Human emotions result from the conscious judgement of events and are complementary to reason in the decision process by making memory of past experiences more relevant, reinforcing intentions and preparing the body for action (Scherer, 2005, p.700–701). People interpret experiences in different levels of emotional details depending upon context of use, past experiences, preferences and expectations (Damásio, 2012, p.19), and this causes emotions to be very ephemeral and hard to guarantee as a result of the interaction with a product (Hassenzahl, 2004, p.47; Chen, 2007b, p.33).

The fleeting character of the fun experience is something that video game designers know by heart. Even though their product is a software system, there are no functional requirements to solve real problems. Instead, video games are supposed to elicit pleasurable emotions in a very similar fashion to what other media (like films and music) do (Fierley and Engl, 2010, p.205). Hence part of the work that game designers are used to do involves the use of patterns (or heuristics) to fulfil "situational needs", like challenges, social interaction and aesthetic preferences, in order to increase the chances that the resulting experiences are pleasurable beyond the fulfilment of pragmatic aspects (Hassenzahl, 2005, p.34–36). Additionally, video games involve more interaction and control than other types of entertainment media, what produces experiences much less linear and makes the distinction between product and experience much more obvious (Schell, 2008, p.11). Consequently, game designers must continuously learn what their users (the players) want (Chen, 2007a, p.11). That is why the other part of the game design process – arguably the most important one – is testing.

Designing for fun requires a two-way communication between designers and users, in order to allow all used patterns to be exercised throughout the entire product development and gain insights into whether or not the aimed experiences are being achieved (Fullerton, 2008, p.248). By creating prototypes of increasingly quality, reviewing ideas, identifying potential users and asking them to test the product, the designer can have valuable feedback on her design choices and apply the needed changes as soon as possible (Fullerton, 2008, p.249). Also, since the emotional experience is much more subjective than satisfaction aspects of usefulness or safety, only the observation or inquiring of people using the product can help identifying obscure aspects of the design and provide opportunities to incorporate good unexpected events discovered by users themselves (Schuytema, 2008, p.30–31). Therefore, the practise of testing is fundamental to help improving the possibilities of a product being fun to broader audiences and contexts. In fact, in the game industry the tests are so important that is common to consider the design choices as hypotheses and the tests with players are experiments (Ambinder, 2009, p.6).

The evaluation of games during such testing sessions has been traditionally done by specialists, either by observing people playing games and inquiring them about their experience with pre and post interviews and by collecting and analysing performance data (Fierley and Engl, 2010, p.206–208). More recent approaches are using physiological measurements collected from biometric sensors to help evaluating the emotional variations that accompany fun experiences (Scherer, 2005, p.709–712; Mandryk *et al.*, 2006, p.3–4; Mauss and Robinson, 2009, p.9–12; Fierley and Engl, 2010, p.206; Nacke, 2013, p.2).

But there are important difficulties with these approaches. Firstly, the presence of an analyst may influence results because the acknowledgement of being observed can temper participants' emotions (Dix *et al.*, 2003, p.328; Isbister, 2010, p.13;15). Also, even though experiments have demonstrated that self-reporting can reduce this influence and facilitate the observation of experiences not foreseen by the designers (especially regarding variations in the context of use) (Jääskö and Mattelmäki, 2003, p.129–130), inquiring is subjective and may disrupt the experience if performed during interaction or not capture the players' real emotional state if performed after the game session is ended (Schell, 2008, p.399). Finally, the use of intrusive sensors applied to the participants' body may also influence results by causing discomfort and easily breaking immersion (the involvement with the game) (Tan *et al.*, 2012, p.2).

In that sense, the use of cameras for capturing facial expressions and extracting data about player emotional variations during game sessions seems to be a promising alternative. First of all, the face is an important channel for the expression of fun. The brain is “hard-wired for facial recognition, just as it is for language”, being an important channel for communicating behavioural intentions (Koster, 2010, p.16). Indeed, many studies performed by psychologists in the last four decades have lead to the general agreement that the expression of emotions in the human face is consistently interpreted among different cultures, being very important to the interaction between humans (Matsumoto and Hwang, 2011, p.1–2; Bettadapura, 2012, p.7–8). Also, facial expressions do not include only signs of emotions. The concentration in children’s faces as they learn new skills, an important aspect of fun, is a good indication of the enjoyment they are feeling (Csikszentmihalyi, 1991, p.47).

Additionally to all that, the current technological state favours this approach because most of the video game consoles and almost all modern game-enabled mobile devices, whether they be smart phones, tablets or notebooks, already have cameras (Tan *et al.*, 2012, p.1). In the domain of digital games, the player attention is consistently focused on the output screen during playing. Even in full body games the analysis of gaze showed that participants observed the action on the screen most of the time (Koštomaj and Boh, 2009, p.151). This condition removes or at least considerably reduces difficulties with Computer Vision algorithms due to partial occlusion of the face (except regarding spectacles, hats or facial hair) and to low resolution (since the player face is usually close to the device), and allows for an non-intrusive capture of data for automated analysis.

1.3 Objectives

The hypothesis that is verified by this research work is that there are many indications of fun expressed by a single person playing a digital game through her face, including not only facial expressions but other cues like the proximity to the camera and the rate in which the person blinks.

So the primary objective of this work is to verify the feasibility of using just this visual information, captured from common off-the-shelf webcams, to assess the level of fun experienced by an individual, without requiring any additional information from other sensors or from the game itself. If this is possible, this will not advocate that other forms of assessment should be replaced or ignored but instead it will contribute with another channel of assessment that potentially has advantages, including low intrusion and low cost. Also, the intended use of the results obtained is to provide tools to help particularly game designers to evaluate their games. It is far from the scope of this work to allow for the creation of self-adapting games – even though the findings obtained might be helpful in that particular effort.

Secondary objectives of this work are:

- Perform robust face detection and tracking, together with effective frustration, immersion and emotion identification, from video images captured from people playing.
- Use this information to perform a dynamic classification of fun in video games played in a personal computer equipped with a web-cam focusing players faces.
- Provide information that can be helpful to game designers in their creation process.

The formulation and scope of these objectives follows this reasoning:

- **On the use of just visual information from the face**.
  Most of the work being performed in the automated assessment of fun in video games uses data collected either from performance metrics programmed into the games or from physiological responses of biometric sensors applied to the body of participants (Nacke, 2013, p.589–590). Data from expressive behaviour like vocal tone, facial expressions and body movement is less used for that purpose. Also, in the domain of video games, the analysis of vocal amplitude and pitch may be more difficult because players don't necessarily are much talkative while playing alone and particularly because games make extensively use of sound effects and music as part of their aesthetic appeal. And body expressions may suffer from a lot of occlusion, considering that the existing cameras in game-enabled devices are naturally focused at the players sitting in front of the screen.

- **On the assessment of just individual players**.
  The relevance of social context for positive experiences and fun has been well documented (Forlizzi and Battarbee, 2004; Lazzaro, 2004; Lazzaro, 2008; Lazzaro, 2010) and it undeniable. However, it poses a much harder problem to consider because multiple faces must be individually tracked and reasoning about the context becomes absolutely necessary, since it is required to known if all observers are indeed players and what are their relations to the game and to each other. Additionally, disregarding its clear importance, the social context is believed to be a secondary aspect because for the experience of fun to be allowed a bond of the individual and the game must be first established (Calvillo-gámez *et al.*, 2010, p.53). That is why the social aspect of fun is intentionally left outside the scope of this research work.

- **On the absence of integration with game logs and metrics**.
  Having a tool that produces information about weather the player is having or not fun from just measurements external to the game is interesting, but the designers would still need to manually match the peaks of fun and boredom to the game design elements in order to understand and explain those "cues" (Nacke, 2013, p.590), most probably by comparing recordings of the gameplay screen to the reports provided by the tool. This naturally requires programming some form of communication between the game and the assessment tool. But instead of proposing an interface *from* the game, it looks better to use an interface *to* the game, that is, an indication about the level of fun being experienced without considering particular aspects of the product. First of all, fun is not just about the balance between challenges and skills (more details will be given in the next two chapters), so this approach concentrates the analysis of fun in its emotional root without much influence of the performance constraints already imbued by the designer in the product's interface. Also, it allows the possible reuse of the solution in other entertainment products, even when there is no direct interaction from the users. Consequently, the integration of the responses provided to the internal measurements of a game are intentionally left outside the scope of this work, even if game designers will still have to match its reports to the game cues manually.

- **On the purpose of producing helpful information to game design**.
  As it will be detailed in the next two chapters, fun is largely dependent upon emotions (Fullerton, 2008, p.258; Lazzaro, 2010). And emotions can only be *inferred* from bodily displays – a definitive measurement of the affective state of a person would require also reading her mental representation of feelings, which is very unlikely to be ever achieved (Scherer, 2005, p.712). It is also unlikely that emotions would be well enough assessed to allow a game to adapt itself to the player emotional state, because the body is constantly influenced by external and internal elements beyond the game control (Nacke, 2013, p.612). In fact, game designers usually dislike Dynamic Difficulty Adjustment (DDA) systems because they take away their control over the user experience, and few commercial games have it implemented (Hunicke and Chapman, 2004, p.91; Chen, 2007a, p.8). So the efforts performed in this work always consider the assessment with intention to support game design, and not to automatically adjust any of the game elements.

1.4 Contributions

The main contributions of this work will be the findings regarding the feasibility of inferring fun from facial images. If the findings are positive, there is also the source code of software system employed in the evaluations, which will be freely available under an open-source license and could be useful for further research and independent video game creation producers. The results found in this thesis might also help further evolutions in the creation of better Dynamic Difficult Adjustment (DDA) mechanisms to allow a game to update itself based on the measurement of fun from the player's face.

There is also the minor possibility of the results and source code provided being useful to other areas that need to assess the hedonistic aspects of fun, in entertainment and publicity, for instance, even though this is not the focus of application originally intended.

1.5 Work Organization

This document is presented as a requirement for the conclusion of a PhD in the course of Computer Science at the Institute of Mathematics and Statistics of the University of São Paulo. The text is organized as follows:

- Chapter 2 (What is Fun?) presents a review on what fun is and what are its psycho-physiological aspects, building the general model used by the rest of this work.
- Chapter 3 (Existing approaches to assess fun) reviews the literature on how fun has been assessed, either in manual or automated fashions, particularly focusing the design of video games.
- Chapter 4 (Data Collection) presents the experiment conducted to obtain the data needed for the creation of the prediction models and the analysis of the results.
- Chapter 5 (Extraction of Features) describes the features required to characterize fun through its component aspects, and how these features were extracted from the videos collected in the experiment.
- Chapter 6 (Towards the Assessment of Fun from Facial Images) describes the computational solutions employed to predict fun through its component aspects: the prototypic emotions, frustration and immersion.
- Chapter 7 (Results and Discussion) presents the results and discuss their implications regarding the objectives of this thesis.
- Chapter 8 (Conclusion) summarizes the work and its results, and also indicates relevant follow-up works that should be performed to improve on what has been obtained.

Chapter 2

What is Fun?

Fun is something hard to define in a few sentences. The main reason is that, like any affective subject, fun intersects many different concepts such as joy, amusement, pleasure, satisfaction and enjoyment, each one with its own nuances. Pleasure, enjoyment and fun are particularly close, and this proximity can be observed in their semantic definitions[1]:

pleasure

enjoyment, happiness or satisfaction (Cambridge Dictionary)

something that is done for enjoyment or satisfaction; sensual gratification; frivolous amusement (Merriam-Webster Dictionary)

enjoyment

the feeling of enjoying something (Cambridge Dictionary)

a feeling of pleasure caused by doing or experiencing something you like; having and using something that is good, pleasant, etc (Merriam-Webster Dictionary)

fun

pleasure, enjoyment or entertainment (Cambridge Dictionary)

someone or something that is amusing or enjoyable; an enjoyable or amusing time; the feeling of being amused or entertained (Merriam-Webster Dictionary)

Although those definitions seem very similar, there are important distinctions that humans intuitively understand because they have personally experienced such feelings many times in their lives, either by spinning with chairs, playing hide-and-seek and walking on parks as children, or by going to theatres, doing hobbies and driving cars as adults. Like in this passage quoted from Dix (2004, p.9):

> An evening quietly sipping wine with friends, slowly watching the breeze fleck the still surface of cool waters, far off the gentle sound of a beck tumbling towards the lake, ducks and swans slowly gather on the water's edge as the sun casts vivid light shows across the distant hills. Enjoyable – yes. Fun . . . ?

At first glance it seems very obvious that the feeling portrayed in this poetically illustrated passage is something lacking the intensity commonly related to fun, because most probably it is about the calm and relaxing moments of a group of people on holiday as described by one person. However, that same scenario could be remembered by another person as fun if it were considered in a different *context* like "that night my friends and I got together by the shore and laughed our heads off telling jokes". If remembered like that, memories will certainly include the specific people present, their excitement, the loudness in their laughs and how fast those precious moments slipped away, differently than an instinctive memory of just physical good sensations captured by the senses. Picking up on the car driving example mentioned before, while most of the time this activity may be only pleasurable for some people, it can become quite fun in especial circumstances, for example when competing with friends in a kart track. The differences are sometimes subtle, but very clear for

[1] as provided by the Merriam-Webster http://www.merriam-webster.com/ and the Cambridge http://dictionary.cambridge.org/ on-line dictionaries

any human with their own preferences and inserted into different situations. This only strengthens the relevance of context in fun, and definitely makes its understanding much harder.

Humans know they had fun when they have experienced that intense, fleeting and desirable moment in which they enjoyed laughing at the comedy play, crying at the cinema or fearing the next descent in the roller coaster, and completely forgot about time and problems. In the domain of interactive entertainment systems, aspects that have been described as contributing to that intense pleasure include attention, play, interactivity, conscious and unconscious control, engagement, and style of narrative (Preece *et al.*, 2002, p.19). Therefore, the rest of this chapter pursues into building a better understanding of fun from such aspects, exploring main subjects such as Flow, Immersion and Emotion, and concluding with a review on how the Human-Computer Interaction and the Game Design areas attempt to research and project fun experiences.

2.1 A Broad View of Fun

2.1.1 Interaction and Attention

When humans have fun there is usually some form of activity involved, like doing sports or playing with toys and games. Playing is definitely the sort of activity that is most easily related to fun, specially in the case of games. After all, games are supposed to be fun, otherwise people will simply not want to play them (Sweetser and Wyeth, 2005, p.1). Games are systems with rules that offer possibilities of action by which people attempt to control the outcomes (Fullerton, 2008, p.89). This means that playing a game is done through some *interface*, which in the case of video games (games played with the aid of computers) is clearly composed of control devices, speakers and graphical displays (Calvillo-gámez *et al.*, 2010, p.51). So *being active* is perhaps the most straightforward way of having fun. However, one might have fun by just watching others playing, that is without having any direct influence in the activity outcomes. Nevertheless, the experience is still due to the *perception* of the outcomes obtained from others. It then seems only natural to assume that whenever someone have fun, the experience can only happens because a person is somehow *interacting* with a task, an object or with other people.

Interaction is a phenomenon of mutual or reciprocal influence that happens when two or more entities communicate with or react to each other (Wagner, 1994, p.8). In the domains of Product Design and Human-Computer Interaction (HCI), which are specifically concerned with the interactions of humans and man-made objects, this mutual influence is seen as being structured in the general form of a feedback loop: a person with goals acts in the environment to achieve them (i.e. provides inputs for a “system”, which may be an object or a person with her own goals), measures the effects of her actions (i.e. interprets feedback outputs from the system to whom she interacts with) and them compares the results with goals, restarting the cycle if judging necessary (Dubberly *et al.*, 2009, p.69,70,75). Humans, as well as other animals, are frequently interacting with objects and with each other, and this behaviour is important because the world is a very large, dynamic and stochastic environment in which identical situations are rare. This condition forces intelligent beings to have to constantly deal with uncertainty and makes the ability of perceiving, reasoning and acting upon changes very important for an effective subsistence (Valiant, 1995, p.3).

The dynamics of interaction foster the emergence of relational properties that are primal to how humans experience the world. In the level of object interaction, humans rely on perceived physical or cognitive cues of possibilities of action, called *affordances*, in order to know how objects can possibly be used (Norman, 2002, p.9). With the help of the data perceived from such sensory cues, humans also build conceptual models from existing knowledge of previous similar interactions and also consider physical or logical constraints and social or cultural conventions that may respectively limit or suggest approaches of use in current context (Norman, 1999, p.39–41; Norman, 2002, p.9–13). In the level of human interaction, people usually share a common environment either in the real world or in the context of product usage (the fantasy world in a video game, for instance). As agents coexisting and working for individual goals, they *always* affect each other's results in either a positive, negative or neutral way, even if they do not acknowledge each other or each other's goals (Garcia and Sichman, 2003, p.282). When the acknowledgement of this *social interference* happens, that is, when the agents get to perceive each other's goals by means of verbal or non-verbal communication and to act upon reasoning on that knowledge, many complex group behaviours arise, like competition, cooperation and coordination, among others (Dubberly *et al.*, 2009, p.75; Garcia and Sichman, 2003, p.285–286).

The way evolution found to fine tune this perception-action mechanism is by making such relational properties very appealing to the nervous system, particularly when novelty is involved. All the information about what is happening outside and inside the organism is represented in the consciousness so it can be evaluated and acted upon by the body (Csikszentmihalyi, 1991, p.24). This content is kept in order by intentions – other bits of information derived from biological needs or internalized social goals – which drive *attention* towards or away from the received stimuli (Csikszentmihalyi, 1991, p.27). But the capacity of information processing inside the conciousness is very limited[2] (Csikszentmihalyi, 1991, p.29), hence any patterns recognized in the sensed data focused by attention are broken into "chunks" of information so they can be hereafter used without the need of much reasoning by the autonomous nervous system (Csikszentmihalyi, 1991, p.29; Koster, 2010, p.30).

Because of this way in which the nervous system works, the human brain dislikes chaos and is constantly trying to understand the patterns it encounters. That is most likely why artefacts and situations like games and toys are commonly very *interesting*. They are abstractions of the real world in the form of a symbolic system with rules and signs that work like patterns triggering this need of understanding when presenting themselves as something different that requires new "chunking" (i.e. the solving of patterns sensed and their transformation into learnt chunks of knowledge) (Koster, 2010, p.34–35). Once the game patterns are known, they become predictable and are no longer as interesting as before.

Still, in order for any new piece of information to be interesting it must have some similarity to existing knowledge. Sensory data that is always unchanged quickly leads to predictability, but noise that conveys a lot of information (in the sense of Information Theory) causes total incomprehension (Schmidhuber, 2010, p.237). In the former case, the processing of information no longer happens in the conscious part of the brain, and become just a repetitive task that does not require attention at all. In the latter case, the amount of information rapidly overloads the consciousness capacities, the brain is no longer able to construct proper conceptual models about the subject of attention, and the information ends up causing confusion and making more difficult the decision on how to act. Therefore, in both scenarios the information is not interesting, but simply boring and undesirable (Schmidhuber, 2010, p.237). Both situations are actively avoided, and hence the attention mechanism helps in filtering what is worthwhile putting effort to.

As consequence, attention is a very singular capacity of the human brain that has much to do to how we experience enjoyment and fun. In a very singular work resulting of decades of interviews and analysis of how people from different cultures, socio-economic conditions, gender, age and jobs experienced their lives, Csikszentmihalyi (1991) formulated a theory he denominated *Flow* by which enjoyment is explained as being the consequence of the proper balancing between challenges and skills that is achieved by people when they are able to focus their complete attention on a given task.

[2] estimations indicate that a person can only process at most 126 bits of information per second, equivalent to paying attention to a theoretical maximum of three conversations simultaneously if everything else is kept away from consciousness

2.1.2 Challenges and Skills

By focusing attention, a person basically retrieves bits of information from memory, evaluates those bits in consciousness and then chooses the right things to do. People who can focus attention at will – something that requires cognitive effort in a way that it can be understood as the spending of *psychic energy* – are known to fully live their lives and thus enjoy it more often (Csikszentmihalyi, 1991, p.33). The reason is that as humans interact with the world they maintain an image of themselves – called the *self* – which is a sum of all memories, actions, desires, pleasures and pains experienced so far. The self is a product of attention because only things considered relevant in past experiences are worthwhile internalizing, but it also helps driving attention in future interactions because the self contains as well mental representations of the entire structure of an individual's goals and their relative importance (Csikszentmihalyi, 1991, p.34). In that sense, the self subsumes personal preferences and attitudinal tendencies, and in fact personality traits like extrovert, high achiever or paranoid can be described by how each person preferably allocates her limited attention (Csikszentmihalyi, 1991, p.33).

The cyclic dependence between attention, self and an individual's goals is connected to the quality of life (Csikszentmihalyi, 1991, p.35). Whenever internal or external information sensed from experiences threatens one's goals, it may disrupts the consciousness order and thus causes a *psychic entropy*: depending on how important are the goals and how severe are the threats, some amount of attention has to be dedicated to eliminate the danger, leaving less space for other matters. If experiences that cause psychic entropy are frequent, the constant disruptions of consciousness order may even weaken the self and make more difficult future investments of attention to goals (Csikszentmihalyi, 1991, p.37). The opposite condition of the psychic entropy is called the *optimal experience* (or Flow), and is the moment when enjoyment is felt in its biggest magnitude with people feeling to be "in the flow", as described by Csikszentmihalyi (1991, p.39) himself:

> When the information that keeps coming into awareness is congruent with goals, psychic energy flows effortlessly. There is no need to worry, no reason to question one's adequacy. But whenever one does stop to think about oneself, the evidence is encouraging: "You are doing all right". The positive feedback strengthens the self, and more attention is freed to deal with the outer and the inner environment.

That improvement on the self is what imposes important distinctions between what are pleasure and enjoyment (Csikszentmihalyi, 1991, p.45). Pleasure is a feeling of contentment achieved when psychic entropy caused by biological or social issues is reduced through the sensing of information, like the taste of food when hungry or the sight of a beautiful (or very private) beach. Experiences involving sleeping, resting, feeding and having sex can bring pleasure and are important to the quality of life, but they do not add complexity to the self by their own. Enjoyment, on the other hand, is related to life events that occur when a person has not only met prior expectations or satisfied a need or desire, but also exceeded those expectations perhaps in an unique way (Csikszentmihalyi, 1991, p.45–46).

Supporting the improvement of the self there are the challenges. Challenges are simply "opportunities for action" contained in any activity and that require appropriate skills to be realized (Csikszentmihalyi, 1991, p.50). Just like the activities and the skills themselves, the challenges do not need to be physical and can be simply mental representations of actions (Csikszentmihalyi, 1991, p.49). Even though people can experience extreme joy for apparent no reason, it is far more common that those experiences happen in a context involving goal-directed activities bounded by rules, thus requiring the deliberate investment of psychic energy (Csikszentmihalyi, 1991, p.49).

This self-improvement mechanism is divided into eight components (or requirements) that contribute to enjoyment (Csikszentmihalyi, 1991, p.49–67; Nakamura and Csikszentmihalyi, 2001, p.90):

- **Challenging activity**
  Enjoyable experiences are usually comprised of a set of challenging activities that require concentration (and thus, the spending of psychic energy) and skills to be performed. Such activities are goal-driven and bounded by a set of rules that must be followed. No matter how difficult the challenges are, there still must be a perceivable chance that they can be completed and goals achieved.

- **Merging of action and awareness**
  Activities in enjoyable experiences completely absorb a person's attention, in a way that there is no psychic energy left to process any other information but what the activity offers. Actions become spontaneous and automatic, and a person does not even realize she is doing them in a way that all seems effortless, despite the amount of concentration required. Nevertheless, this sensation is very fleeting in the sense that any lapse of concentration will simply erase it.

- **Ability of total concentration**
  Due to the high demands of concentration that the challenges impose, as long as the Flow experience lasts people are no longer able to think about unpleasant aspects of real life. They forget about problems and frustrations and completely focus attention on the task at hand.

- **Clear goals**
  The concentration is just possible because the task has clear goals. During the execution of an activity one always needs to know what is possible to be achieved and what are the benefits of success, in order to evaluate if the effort in doing so is worthwhile. Even artists that do not have clear goals at the beginning of a piece need to have a strong sense of what is "bad" and "good" and have to develop their intentions as their work progress.

- **Immediate feedback**
  Similarly to clear goals, the person involved in an activity needs to have immediate feedback on whether her actions are being successful. Feedback that is related to the goals is enjoyable because provides information on the invested psychic energy. The type of feedback may differ according to the activity and people may have preferences towards the type of information they pay most attention to. So valuable feedback is a symbolic message about the success of the interaction towards the goals, that creates knowledge capable of ordering the consciousness and structuring the self.

- **Sense of control**
  Most of the enjoyment of facing challenges is due to the acknowledgement that risks were diminished and difficulties were overcame by one's own actions. In Flow, perfection is attainable at least in principle. To succeed in facing difficult challenges means that enough skills were acquired to control the world and its entropy, and that is very enjoyable. The sense of control partially explains why enjoyable experiences are very addictive: even in activities with random outcomes (like games of chance), people commonly attempt to explain their results in the light of their own choices.

- **Loss of self-consciousness**
  When engaged in enjoyable activities, a person in Flow loses consciousness of her own self because there is not enough attention left to anything but the task demands. This does not mean that she is no longer in control of her consciousness or that her self is completely lost, but just that during the time of the activity there is little opportunity for the self to be threatened by real life worries. The person engaged in the experience feels like she is one with the environment, and paradoxically have her self improved afterwards when she has the chance to resume and evaluate the new skills learned and achievements reached.

- **Alteration in time-perception**
  As with the loss of self-consciousness, people engaged in enjoyable experiences seem to lose track of time passage. A fact is that humans perceive time passage with reference to external references of events like night and day or the progression in clocks and other ordered devices. And, as the attention is completely focused on an activity, there is no much space for following time progress – unless the ability to keep track of time is a skill required to succeed at the task in execution. So after the experience is ended, time can be perceived as to have passed quickly than expected. Indeed, experiments that intentionally distort time perception indicate that people perceive experiences as more enjoyable if time seems to pass surprisingly quickly rather than slowly, even when the experience is clearly unpleasant (Sackett *et al.*, 2010, p.118).

In essence, enjoyable experiences usually have challenges that never overmatch or underutilize one's skills, meaning that people in the Flow are constantly *in a state in which the perceived action capacities match the perceived action opportunities* (Nakamura and Csikszentmihalyi, 2001, p.90). But it is not just that: the challenges and skills must be above the individual's average level, otherwise apathy is experienced. Also, if just one level is smaller than the other the result is yet not a positive experience: high challenge with lower skills quickly turns into worry and anxiety, and low challenge with higher skills may quickly turns into relaxation and boredom. Good experiences keep the levels of challenge and skill constantly above the individual's average, as a person improves herself in the performed activity and searches for more challenging goals (Csikszentmihalyi, 1991, p.75). That is, the challenge-skill state keeps changing from control (high challenges with higher skills) to arousal (high skills with higher challenges) and back to control again (Nakamura and Csikszentmihalyi, 2001, p.94). The figure 2.1 represents graphically the possible Flow states considering the levels of challenge and skills respectively in the vertical and horizontal axes.

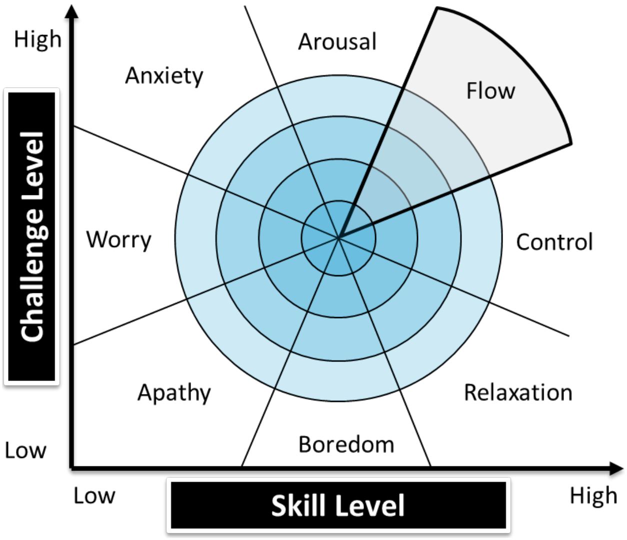


Figure 2.1: *Current model of the Flow state. Based on Nakamura and Csikszentmihalyi (2001, p.95).*

Video games are good in keeping people in the Flow not so much by automatically adjusting the difficulty of challenges, but also by offering information on progress. Via the correct amount of feedback (neither too much nor too low, so to avoid frustration) a player is either rewarded for mastering challenges or presented with enough information on what and how she failed them, thus being able to try again and having more chances to improve skills (Prensky, 2001, p.117).

Back at the beginning of this chapter the differences between pleasure, enjoyment and fun were initially attempted from their dictionary definitions, when it was argued that the differences could be due to the intensity and fleetingness of the emotions felt as well as to the context in which they happened. The Flow theory handles enjoyment and fun as similar subjects, and clearly differentiate them from pleasure in a sense very close to what has previously been intuited. Csikszentmihalyi

(1991, p.46) provides a comparison example involving food, stating that everybody has pleasure in eating but a gourmet enjoys doing it in a different and more complex way, paying attention and, thus, investing psychic energy to the various sensations received. The gourmet's intentions when eating are not just satisfying biological needs, but to discover novel flavours and interesting ingredient combinations. Also, to become fit to the culinary challenges, she needs to have her palate and nose trained. As consequence, she imposes herself some goals and tries to reach them by exploring different possibilities of interaction, so when the goals are satisfied the gourmet surely feels enjoyment *beyond the physical pleasure*. Therefore, enjoyable and fun experiences have in its context goals and challenges to overcome, and the intensity of the feeling depends upon the concentration put into the interaction.

2.1.3 Intrinsic Motivation

Activities that produce challenging experiences are so gratifying that people are willing to do it for its own sake, no matter if they are difficult or dangerous (Csikszentmihalyi, 1991, p.71). But even activities that are goal-driven like video games do not rely only on challenges to produce fun as rewards. There are other relevant aspects that are due to the human natural desire to learn, and that help in making activities intrinsically motivated.

The mechanism of attention previously discussed involves the continuous searching for unknown patterns to decipher, and that simply means being eager to learn new things. The very notion of an active behaviour driven by pleasure from learning has also been studied as *curiosity*, though with two slightly different views (Litman, 2005, p.793–794):

> Curiosity may be defined as a desire to know, to see, or to experience that motivates exploratory behaviour directed towards the acquisition of new information. [...] (it) is often described in terms of positive affective, and acquiring knowledge when one's curiosity has been aroused is considered intrinsically rewarding and highly pleasurable. [...] (however) discovering new information may also be rewarding because it dispels undesirable states of ignorance or uncertainty rather than stimulates one's interest.

In other words, acquiring new information after having the senses aroused (i.e. being previously interested at something and having attention focused on it) does not seem to be the only way of getting pleasure from learning. Organisms will actively search for things that are "yet unexplained but easily learnable" even in the absence of novel or complex stimuli (Litman, 2005, p.794; Schmidhuber, 2010, p.230). This means that fun probably is as much about avoiding boredom as it is about maximizing internal joy by overcoming challenges.

That *exploratory behaviour* is intrinsically motivated because the rewards obtained are not obvious or externally visible things like money, food or social reinforcement, but instead are a positive affect produced by performing an activity just for the sake of doing it (Malone, 1980a, p.8; Schmidhuber, 2010, p.230). In another seminal work, Malone (1980a, p.49–63) separated the characteristics of intrinsically motivated learning activities into three major categories, that he named *curiosity*, *challenge* and *fantasy*. The author based his work not only in Csikszentmihalyi's Flow theory, but also in other important researchers, like Piaget (1952). Indeed, the description of how these categories relate to the intrinsic motivation in learning environments is well made by comparison with Piaget's theory of cognitive development (Malone, 1980a, p.49):

> [...] people are driven by a will to mastery (challenge) to seek optimally informative environments (curiosity) which they assimilate, in part, using schemas from other contexts (fantasy).

According to Malone, curiosity is an important motivator for learning *disregarding the fulfilment of any conscious goals or fantasies*. It only depends upon an environment with the *right level of* *informational complexity*, meaning that the sensed data is novel and surprising, yet not completely incomprehensible (Malone, 1980b, p.165). The best environments for arousing curiosity will be the ones in which the learner have enough knowledge in order to have *expectations* about what will happen, and *sometimes* those expectations are not met (Malone, 1980a, p.60). That partially explains why following narratives (in films or books, for instance) – an experience in which a person takes no particular action towards the achievement of goals – can be described many times as fun beyond the mere experience of physical or cognitive pleasure.

Challenge, on the other hand, is strongly related to the existence of goals, as described by the Flow theory. In learning environments such as games, goals directly represent the means to achieve some intended results and, in consequence, the skills necessary for doing so (Malone, 1980a, p.51). As consequence, goals are mostly interesting when their attainment is *uncertain*. Goals that are either certain to be achieved or definitely impossible are just not worthwhile to be attempted, but when there is some degree of uncertainty and a goal is achieved, this accomplishment is good to self-esteem, makes people feel better about themselves and helps them to learn about their own abilities (Malone, 1980b, p.163).

Fantasy is about "mental images of things not present to the senses or within the actual experience of the person involved" (Malone, 1980a, p.56). Even though modern computer games can provide digital representations of fantastic beings and scenarios, the idea is that fantasy is something just represented inside the mind of a person immersed into such a learning environment – i.e. the make-believing mentioned by Piaget as central for the development of the symbolic representation skills in children (Malone, 1980a, p.59). In that sense, fantasy includes representations of physical objects, people and social situations, either completely possible or impossible (Malone, 1980b, p.164), that help accommodating to an external reality, passively repeat past experiences to achieve emotional mastery or fulfil unconscious wishes and maintain optimal levels of mental arousal (Malone, 1980a, p.5).

There are many important parallels between those three concepts that help in explaining why people enjoy doing activities like playing games. Challenge and curiosity are similar regarding their requirements for making the environment interesting and how uncertainty is diminished. The former aims to achieve an optimal level of difficulty by means of *adjusting learner's abilities*, and the latter aims to achieve an optimal level of complexity by *adjusting learner's understanding*. Challenge reduces uncertainty by helping in perceiving one's own capacity to reach goals, and curiosity does the same by helping in perceiving the state of the world (Malone, 1980a, p.60–61). Also, sensory curiosity adds to challenge by serving as rewards for good performance and increasing the salience of goals (Malone, 1980b, p.166).

In challenge, the best goals are practical or related to fantasy because in that way they are easier to comprehended. When the skills used to interact with the environment have a strong relation to the fantasy involved (meaning that the use of the skill also depends upon information represented by the fantasy), the learning activity is usually more interesting and instructional because the skills work as *analogies* between existing knowledge about the fantasy world and the unfamiliar things that are being learned (Malone, 1980b, p.164).

It was discussed before that for any new information to be interesting it needs to have some similarity with existing knowledge acquired in previous experiences. So, when curiosity matches with existing knowledge, fantasy helps in the construction of conceptual models that are not confusing and hence are useful in deciding how to interact. It does that by evoking past experiences and memories, components of the self. Thus, fantasy may be the characteristic that is mostly related to subjective preferences and expectations. Indeed, it is a very important aspect of interaction since something radically new can only be comprehended in terms of old knowledge (Malone, 1980a, p.59).

As a matter of fact, those characteristics of intrinsically motivated activities result from interaction in both object and human-level interaction. The perceived affordances serve as mental interpretations of things, based on past experiences applied to new perceptions (Norman, 2002, p.219), and the social interference in learning activities like games definitely affects the attention drawing mechanism of individuals from commentaries, coaching or proposals of new challenges made from others (Isbister, 2010, p.13). Also, appeal and engagement seem to be mostly explained through a combination of things contained in those general categories. Studies with children have shown that while attributes like colour have less importance in choices for objects of play, novelty is very important in determining which toys to initially chosen, and complexity (both in terms of construction and possibilities of use) is very important in determining how long the interaction with it will last (Malone, 1980a, p.6). After a toy ceases to be interesting, it might no longer be used. Nevertheless, all the knowledge gained from playing with it, as well as the physical and cognitive sensations remembered to be experienced when that activity was performed, will definitely be employed in deciding what to play in the future.

2.1.4 Immersion and Engagement

The observation that fun is not just about challenges and the achievement of goals is specially clear regarding playful activities. The Categories of Play suggested by Caillois (2001, p.12–26) already described free-form (*paida*) activities that do not have clear goals (like the ones in the Vertigo play, or *ilinx*) or to which fantasy has stronger relevance (like the ones in the Make-believe play, or *mimicry*) and yet are commonly experienced as fun. Some examples of playful activities classified according to those categories are provided by Fullerton (2008, p.92) in table 2.1.

Table 2.1: *Examples of playful activities in each combination of Caillois's categories. Reproduced from Fullerton (2008, p.92).*

	Free-form play (*paida*)	**Rule-based play (*ludus*)**
Competitive play (*agôn*)	Unregulated athletics (foot racing, wrestling)	Boxing, billiards, fencing, checkers, football, chess
Chance-based play (*alea*)	Counting-out rhymes	Betting, roulette, lotteries
Make-believe play (*mimicry*)	Children's initiations, masks, disguises	Theater, spectacles in general
Vertigo play (*ilinx*)	Children "whirling", horseback riding, waltzing	Skiing, mountain climbing, tightrope walking

It turns out that the differences in playful activities occur according to the level of interactivity involved (Fullerton, 2008, p.38; Deterding *et al.*, 2011, p.13; Groh, 2012, p.41), making possible to also classify them in the manners illustrated in figure 2.2. Fantasy and curiosity are very basic characteristics of all forms of play, and the need of challenge starts to appear as there are more possibilities of interaction, specially from object-level to human-level interactions (the Gaming-Playing axis in the figure). Stories and narratives like the ones in books and films involve fantasy play, but can not be changed or manipulated by the person engaged in interaction[3]. Toys also have fantasy but can be manipulated, although without any fixed goal. Puzzles are rule-based systems with fantasy and that can be manipulated, however they have a goal of finding a solution, and so the feedback information starts to become much more relevant. Finally, games include all previous elements with the difference that they have the goal of winning, that is overcoming challenges proposed by the game itself or by other human players (mostly from competition). Indeed, modern researchers also consider an additional axis (the Whole-Parts axis in the figure) in this classification to differentiate whole products (like toys and games) from enjoyable activities composed of several elements (or parts) of game or playful design (like outdoor activities, interactive art installations and gamified[4] tasks) (Deterding *et al.*, 2011, p.13; Groh, 2012, p.40), so any enjoyable experience

[3] even though the figure says "no interaction", there are still the perceptions of text and images by the person following the story, as well as the cognitive effort in focusing attention

[4] gamefication is a very trendy word meaning the use of game design elements in non-game contexts

can be understood and discussed in those forms.

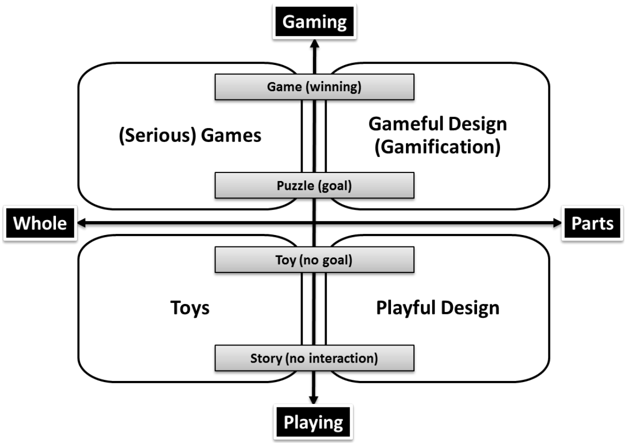


Figure 2.2: *Types of play according to interaction and part dimensions. Based on Fullerton (2008, p.38) and Groh (2012, p.41).*

That means that enjoyable activities certainly are more than just overcoming challenges and achieving internal goals, because they are also about the freedom of choice and the provision of opportunities to use imagination, fantasy, inspiration and social skills in a free form (Fullerton, 2008, p.34). Due to this view, fun has also been studied from other concepts, the most notable one being Immersion.

Immersion is believed to be a "very important experience of interaction" (Brown and Cairns, 2004, p.1297) and as something critical to the enjoyment of playful activities like video games (Jennett *et al.*, 2008, p.644). It is considered to be similar to the concept of Flow regarding the reduction of self-awareness and the distortion in the time perception, and yet as something different (Jennett *et al.*, 2008, p.647). In essence, Immersion describes the feeling of being totally involved by the environment of an interactive system (like video games, interactive narratives and virtual reality environments), sometimes reaching the extreme of being completely transported into a different world and profoundly related to constructional elements like characters and their stories.

In the context of interactive drama, Immersion has been described as being the result of the use of schemas in the interpretation of content (Douglas and Hargadon, 2000). Schema theory is a recurrent topic in the analysis of narratives, as it describes how the perceptions and actions shape expectations and interactions through schemas: data structures representing generic concepts and knowledge that allow humans to understand the world and eventually act upon it (Douglas and Hargadon, 2000, p.153). So, in sum, schemas are conceptual models built from previous interactions and that are part of the self.

So Immersion is a state in which a person is completely absorbed in trying to fit the fantasy and narrative into a single known schema, with pleasure deriving from the recognition of that long-familiar pattern infused with unique or unpredictable elements (Douglas and Hargadon, 2000, p.154–155). That would be the reason why mystery books and horror films are continually appealing even though they usually present minor variations on the basic genre structure. Additional to that, when a narrative content subverts a single schema but still provides familiar alternatives, its audience can work out the conflicts from multiple known schemas and a state of Engagement can be achieved (Douglas and Hargadon, 2000, p.154–155). This state requires much more attention and cognitive effort since its comprehension involves "decision-making, superb eye-coordination, the ability to read character's intentions and predict their actions" (Douglas and Hargadon, 2000, p.158). The pleasure obtained when in this state comes from the feeling of having the skills to solve an unusual and difficult plot, hence this is an immersive state much closer to what Flow describes: the absorption is much bigger to the point of causing the sensation of being a decision-maker or even a co-author.

In the context of video games, Immersion has also been defined from the structured analysis of data collected from interviews with players (grounded theory), resulting in a similar view but focused on the fact that the sense of involvement increases with time as people interact with a game, and also as *barriers* are removed (Brown and Cairns, 2004, p.1298). Barriers are difficulties to immersion from both the human and the system perspectives, like the amount of concentration and the system's construction elements, which need to be removed or resolved in order to *facilitate* the enjoyment in an experience – *but not to guarantee it* (Brown and Cairns, 2004, p.1297). In playful activities like video games, there are three levels of involvement that may be achieved as the interaction unfolds (Brown and Cairns, 2004, p.1298–1300):

- **Engagement**
  This is the first level of immersion, in which a person starts doing a playful activity and thus spending time with it. A first barrier that needs to be removed in order to allow entering this initial state is access: according to preferences and curiosity regarding fantasy, genre and aesthetic characteristics, a person may or may not want to play. Once engaged in playing, feedback becomes an important matter from the very beginning, since the person engaged needs to learn how to act or use objects and controls. In this level of Immersion people are already concentrated and tend to loose track of time disregarding the existence of goals and challenges, because they may want to keep interacting just to satisfy curiosity. However, there is yet no complete lost of self-awareness since the interaction lacks the emotional attachment that occurs in the other levels. Also, as the time progress and the person keeps playing some form of additional reward is required to maintain interest, consistently with the idea of "chunking" knowledge mentioned before in this chapter. Keeping the interest is the second barrier that can lead to the next level of Immersion.
- **Engrossment**
  In this second level of immersion the player becomes more involved with the interaction, mainly due to the construction of the playful activity. That is, due to its structural elements of fantasy, challenges and rules. In a matter of fact, this is probably what paves the path for the optimal experience of Flow to happen, since attention is now almost completely consumed and only the task in execution seems to matter. So, the first barrier here is the provision of initial goals and challenges that are not too difficult. The combination of activity features and the investment of time made by the player cause emotions to be directly affected and the self-awareness to be reduced. Hence people in this level of Immersion become involved not only with the physical aspects of the task but also with mentally represented elements like characters and stories, and start to have their disbelief in the fantastic environment temporarily suspended. The empathy with the fantasy is the last barrier that needs to be broken in order to allow moving to the next level of Immersion.
- **Total Immersion (or Presence)**
  The final state is of total involvement or Presence. It is not just about the sense of being away from the real world, but of being completely inside a virtual word (Calvillo-gámez *et al.*, 2010, p.52), in which "a person's cognitive and perceptual systems are tricked into believing they are somewhere other than their physical location" (Brown and Cairns, 2004, p.1298). Such a state of total involvement is only achievable when all constructional elements, like sounds, graphics, fantasy, narrative, etc, are working together to create a consistent atmosphere and so the perceived actions/responses are not just successful in achieving goals but particularly *meaningful* to the person playing. The player feels like she is really walking, shooting, jumping, dancing or singing along, and that her actions are definitely changing the fantastic world. Consequently, the playful activity is the only thing really impacting her thoughts and feelings. However, since the activity meaning is mostly a mental representation, the experience is as ephemeral as it is intense.

Besides having a strong temporal characteristic, Immersion is also a multidimensional phenomenon because people experience it in different manners according to their own preferences and moods, to the characteristics of the games played and to information external to a particular interaction, like peer influence, game reviews and other sociocultural references (Mäyrä and Ermi, 2011, p.100;101). Thus, Mäyrä and Ermi (2011) have argued that immersion has Sensory, Challenge and Imaginative dimensions, in a way very consistent with the characteristics of intrinsically motivated tasks defined by Malone.

Their model represents a *particular* interaction between a game, characterized in terms of its constructional elements like interface, rules, content and Playability (or easy of play)[5], and one or more players, characterized in terms of their expectations, schemas, emotions, motivations, sensations and skills in a particular social context. Immersion is what provides the meaning for all actions and perceptions that happen during that interaction and it is what allows for the emergence of experiences. It happens in three dimensions, Sensory, Challenge-Based and Imaginative, each with its own feedback loop. The Sensory Immersion is related to how the audiovisual attributes of the game interface, like impressive 3D graphics, large screens and stereophonic systems can keep players curiously engaged even if they are not very used or interested in playing games. The Challenge-Based Immersion is related to how the game rules and playability can keep players interested in go on playing, not only because there are new challenges or goals but because the players have to improve their motor and mental skills in order to overcome or achieve them. And finally the Imaginative Immersion is related to how the game content, like aesthetics, characters, narratives and fantasy, can keep players interested by creating empathy and offering the chance to use imagination. The experience with games always have effects in all of these dimensions, although one is usually stronger depending on the type of game.

Due to its proximity with the concept of Flow, there are still many arguments about what Immersion is and how it can be defined. But many researchers share the belief that Immersion (at least in the two initial levels before Total Immersion mentioned above) provides a sub-optimal experience that does not guarantee fun *but is still valuable* as microflows (figure 2.3). Immersion differs from Flow in the sense that the former is required so the latter can be ever achieved as an *extreme experience* (Douglas and Hargadon, 2000, p.158; Brown and Cairns, 2004, p.1300; Jennett *et al.*, 2008, p.646; Nacke and Lindley, 2008, p.82; Calvillo-gámez *et al.*, 2010, p.53; Mäyrä and Ermi, 2011, p.92–94).

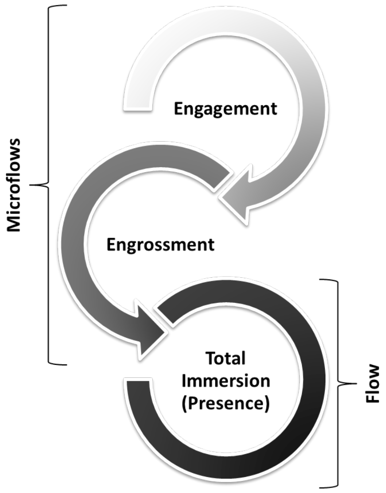


Figure 2.3: *Evolution of the immersion. Based on Brown and Cairns (2004, p.1298–1300).*

2.1.5 Emotions

From all that has been discussed so far, an important observation is that fun is a subjective matter. A big part of its subjectivity is due to the fact that things like immersion and fantasy depend upon

[5] Playability is discussed in more details in section 2.2 (A Design View of Fun)

past experiences, preferences and current mood, and that challenges and goals have distinct appeal to people with different abilities and skill levels. But the subjectivity of fun is also due to its strong relation to human emotions. Emotions are an essential part of entertainment. Classic sports and games have win-lose states that elicit strong emotional and ego-gratification responses, and this is also a big reason for their attraction (Prensky, 2001, p.117).

The study of human emotions is very old, and still today there are different points of view regarding their origin and function. The major theoretical perspectives that inspired contemporary researchers can be classified in four branches (Cornelius, 2000): the Darwinian, the Jamesian, the Cognitive, and the Social Constructivist. The Darwinian perspective is derived from the work of Darwin (1872 *apud* Cornelius, 2000, p.3–4), in which emotions are believed to be the expressions of important *communicative and survival functions* that have evolved in humans and other animals as the species suffered natural selection. The Jamesian perspective is associated with the work of James (1884 *apud* Cornelius, 2000, p.4–5), in which emotions are believed to be not just external expressions of internal functions, but mainly the result of *perceived* bodily responses to the environment that regulate action tendencies. The Cognitive perspective has origins in the Hellenistic philosophy (Cornelius, 2000, p.5–6) in which the central assumption is that thought and emotion are inseparable because emotions, as well as the physiological responses, are the result of the conscious judgement (appraisal) of internal or external events as bad or good. Finally, the Social Constructivist perspective was originated from works on anthropology and sociology that started being applied to psychology in the early 1980's (Cornelius, 2000, p.6–7). Differing from the other branches regarding the primary biological origin, the Social Constructivist perspective understands that emotions are the product of culture and hence emerge from the appraisal of social content to serve particular purposes established by cultural rules (Cornelius, 2000, p.7).

The Social Constructivist perspective is the one that diverges the most from the others, particularly in its disbelief regarding the existence of universal forms of emotional expression, experience and physiology (Cornelius, 2000, p.7). Theorists from the Darwinian and Jamesian perspectives share a belief that the same emotion expressions would be seen across different human cultures, and perhaps also in mammals since they share evolutionary past. In fact, contemporary researchers – most notably Ekman and Friesen (1971) – have demonstrated through years of empirical research that facial expressions of emotion are universally recognized among different human cultures, specially regarding the basic (or prototypic) emotions of happiness, sadness, anger, fear, surprise and disgust. So, even though most modern researchers have acknowledged the role of culture in regulating *emotional displays*, they all seem to accept that there are common grounds for the expression of emotions in humans (Cornelius, 2000, p.5).

The Cognitive perspective is nowadays the most dominant among the four, being commonly referred to as Appraisal Theory and including under its umbrella successful attempts of integrating aspects from the Darwinian and Jamesian perspectives (Cornelius, 2000, p.7). In the past, emotions were considered to be obstacles to good decisions because emotional behaviour was seen just as the opposite of rational thought. But experiences and emotions are inseparable, because interaction has many affective consequences (McCarthy and Wright, 2004, p.17; Hassenzahl *et al.*, 2010, p.353). The human perceptions produce “singular coloured versions of the world as opposed to objective data”, and as consequence actions are not just driven by utility but also by values, needs, desires and non pragmatic goals that are unique to each situation (McCarthy and Wright, 2004, p.85–86). Therefore, it is now largely accepted that emotions occur in many levels at the body and have an important role altogether with reason in influencing behaviour (Picard, 1995, p.2; Plutchik, 2001, p.347; Scherer, 2005, p.706; Damásio, 2012, p.128).

According to the Appraisal Theory, emotions in humans are episodes of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the conscious evaluation (appraisal) of external or internal stimuli as relevant to major concerns of the organism (Scherer, 2005, p.697). The table 2.2 shows the relation between the functions, organismic subsystems and processing components involved in the expression of complex emotions. The Central

Nervous System (CNS) is composed of the brain (with the cortex and the limbic systems) and the spinal cord, which communicates with the Somatic Nervous System (SNS) to transmit impulses to the skin and the skeletal muscles, with the Autonomic Nervous System (ANS) to transmit impulses to the internal organs and muscles, and with the Neuroendocrine System (NES) to transmit impulses to the glands (Hiller-Sturmhoefel and Bartke, 1998; Bogart and Ort, 2007). As it can be seen, the CNS plays a major role in the expression of emotions, and that is why the bodily changes that accompany such experiences are not just physical but also mental.

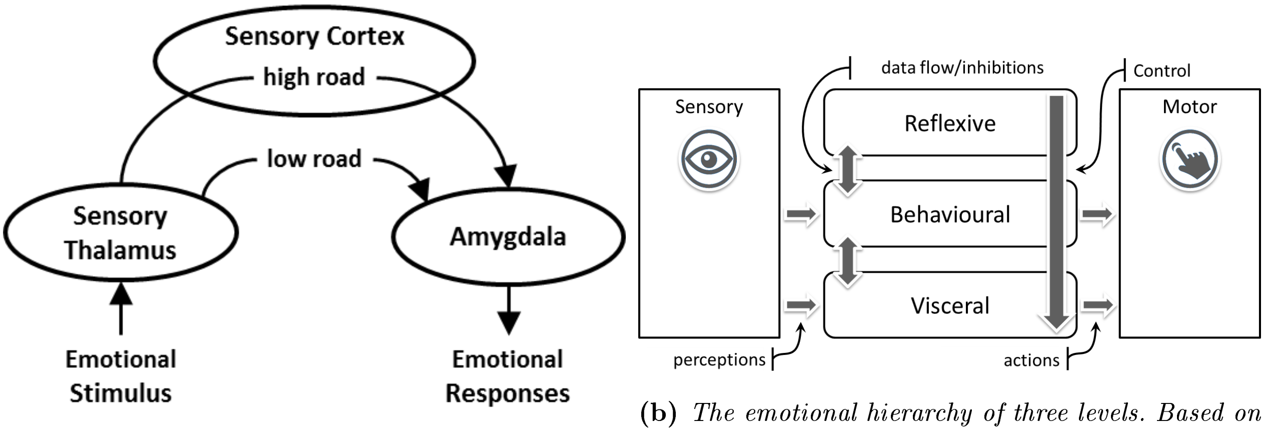
Table 2.2: *Relationships between organismic subsystems and the function of components of emotion. Reproduced from Scherer (2005, p.698).*

Emotion function	Organismic subsystem and major substrata	Emotion component
Evaluation of objects and events	Information processing (CNS)	Cognitive component (appraisal)
System regulation	Support (CNS, NES, ANS)	Neurophysiological component (bodily symptoms)
Preparation and direction of action	Executive (CNS)	Motivational component (action tendencies)
Communication of reaction and behavioral intention	Action (SNS)	Motor expression component (facial and vocal expression)
Monitoring of internal state and organism-environment interaction	Monitor (CNS)	Subjective feeling component (emotional experience)

In the CNS, the limbic system (including the thalamus and amygdala) is the brain structure centring memory, attention and emotions, since all external and internal sensory information pass through it *before and after* conscious analyses in the cortex (Picard, 1995, p.2). However, this is not the only part of the brain involved in the experience of emotions. Many studies with patients that suffered damages in their cortex frontal-lobe indicate that the inability to feel emotions impairs the ability to make decisions, and this is a strong evidence that emotions are just as important to the subsistence as it is rationality (Damásio, 2012, p.66–83).

There are two basic ways (or "roads", as illustrated by figure 2.4a) by which emotions are handled by the CNS: primary and secondary. The primary emotions are understood from the Darwinian and Jamesian perspectives. They can be seen as both innate and acquired behavioural dispositions (Damásio, 2012, p.131). Innate dispositions are simply "hard-wired" unconscious responses to general perceptions that are a product of evolution. They can be adapted when the emotional estate caused by an experience is also unconsciously acknowledged (or "felt"), allowing the formation (acquisition) of new autonomous dispositions for specific objects or situations – what is commonly referred to as having "sentiments" towards things (Branco, 2006, p.29; Damásio, 2012, p.131;140). Primary emotions, both innate and acquired, are completely handled by the limbic system (the "low road") and control visible and non-visible changes in the whole body, like involuntary facial, gestural and vocal expressions, and visceral, muscular, skeletal and glandular alterations (Damásio, 2012, p.131). Since they are primitive and unconsciously triggered, they produce very quick responses that help humans and other animals to rapidly detect and act upon the presence of suitable mate or eminent danger, for instance (Damásio, 2012, p.129). The bodily changes that accompany primary emotions serve to allow communication of intentions (posture, vocal and facial displays, very important for social interaction) and to prepare the body for action (Scherer, 2005, p.698; Damásio, 2012, p.132).

The secondary emotions are understood from the Cognitive perspective. They can be seen as the result of conscious evaluations (appraisals) of objects, people and situations, as well as the internal emotional state of the body, which occur at the brain cortex (the "high road") (Scherer, 2005, p.697–698; Damásio, 2012, p.132–133). This reflective ability is mainly a human characteristic, because in this case the whole process *starts* from the deliberated consideration of a current situation and



(a) *The different neural roads from perception to emotion. Based on from Branco (2006, p.28).*

(b) *The emotional hierarchy of three levels. Based on Norman (2005, p.22).*

Figure 2.4: *The basic mechanism of emotional processing.*

the possible consequences of actions in order to form mental representations of an experience. This higher level process is what allows, for instance, the elicitation of strong emotions in humans without any real external perceptions, that is, from just the remembrance of especial people or situations from the past. After this conscious deliberation, the cortex unconsciously reacts to the cognitive representation of the experience by producing involuntary and automatic bodily responses and eliciting and reinforcing the acquired dispositions and *feelings* (Damásio, 2012, p.135). Feelings are subjective experience components of emotions that reflect the cognitive patterns of appraisal as well as the motivation and somatic responses that underline emotional experiences (Scherer, 2005, p.699).

The inducement and perception of the bodily changes relevant to the secondary emotions are expressed through the same neural structures of the primary emotions, that is the limbic system (Damásio, 2012, p.135), and that is why both emotion roads "feel the same". Nevertheless, emotions are more complex in humans because the way the two roads interact. Norman (2005, p.21–24) presented this view as a hierarchy of levels (figure 2.4b), renaming the innate and acquired primary emotions respectively as Visceral and Behavioural levels and the secondary emotions as Reflective level. The Visceral level is the most basic one, totally primitive and reactive, straightly related to the expression of bodily changes. It is triggered by internal and external perceptions, but can be *inhibited* by the Behavioural level from unconscious automated behaviours learnt from past experiences. And, by its turn, this second level can also be further *inhibited* by the Reflective level due to conscious considerations done by the person experiencing the emotion. Most importantly, information from the top most level is transmitted down to the other levels in this same hierarchy, so new behaviour dispositions can be acquired by skill training and one can intentionally control physical and mental bodily changes.

The emotional aspects of fun happen in all Visceral, Behavioural and Reflective levels, consistently with the previous studied components of intrinsic motivated activities (Norman, 2005, p.23–24). The emotions experienced in the Visceral level are directly connected to sensory perception and arousal, hence being about physical pleasure obtained from the satisfaction of basic needs. The emotions experienced in the Behavioural level are connected with the execution of well learnt routines and tasks and the achievement of difficult goals, hence being about curiosity, learning and the development of skills. And the emotions experienced in the Reflective level are connected to the study and interpretation of things and the pleasure obtained from that, being thus about fantasy, immersion and engagement. And again, a fun experience like riding a roller coaster is complex set of events, being related to all of these levels interconnected. In one level people may feel excited about the speed and the fall, but in another they may also be trilled by the enhanced self-image achieved after they complete a fearful task that many others are not willing to do (Norman, 2005, p.24).

Emotions are expressed through this iterative pattern instead of in one straight direction from perception to physical arousal and unconscious reactive action, because otherwise they would easily take over any rational thought and no human would be able to have self control in face of strong emotionally eliciting situations. Another view of the emotion experiencing process is as a dynamic chain of events originated from both external (sensory perceptions) or internal (memories, dreams and bodily changes) sources of unexpected or unusual information, that are continually handled by the different organismic systems until a state of physical and mental equilibrium is restored (Plutchik, 2001, p.347; Scherer, 2005, p.697–698). Plutchik (2001, p.348) describes this feedback loop process with the stage blocks in figure 2.5 and with some examples (table 2.3) of emotions being experienced and helping in determining behaviour. Joy, an emotion commonly felt when a person is having fun, is illustrated in the situation of gaining a valued object (like receiving a gift from a friend, or earning money from a slot machine or a poker game, for instance).

When a person gains a valued object (*stimulus event*) the emotional process starts with the conscious evaluation (appraisal) of the new situation, which forms a mental image of "having a desired item" (*inferred cognition*). Then, this image is unconsciously matched against similar situations experienced in the past, inducing the same bodily changes (*physiological arousal*). The unconscious acknowledgement of the bodily changes updates the mental image of the situation with a subjective good sensation (*feeling of joy*). By themselves, the bodily changes and the subjective feeling are strong enough to produce the intention to repeat or retain whatever caused them (*action impulses*), and thus the person may react by involuntarily communicating her intentions or by voluntarily doing once again the thing that provided the object (*overt behaviour and displays).* In that way, she can keep gaining more of that object (*effect*).

The stimulus event in this example is the gain of a valued object, but it could also be any abstract desired reward, like a clever comical sound clip intimately related to the preferred fantasy in a video game or a symbolic badge in a social network demonstrating that a difficult practical goal was achieved. Thus, it serves for a general understanding of how enjoyable experiences elicit positive experiences.

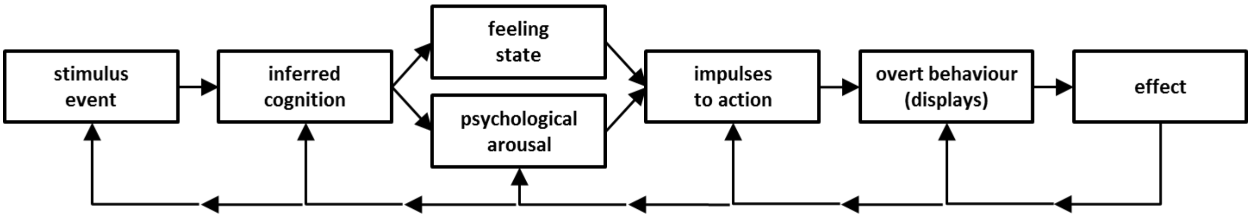


Figure 2.5: *The complete mechanism of emotional processing as a feedback loop of sensory information and action. Based on Norman (2005, p.22)*

All physical and mental changes that happen throughout the stages of the process are continually acknowledged by the person, either consciously or unconsciously, updating the subsystems' states until an global equilibrium is achieved (Plutchik, 2001, p.347; Damásio, 2012, p.134). For instance, even after the physiological arousal and feeling of joy are experienced, a person may consciously choose to ignore the action impulses due to new inferred cognitions regarding the social context or conflicting objectives. Also, the acknowledged bodily changes may become too physically or mentally exhaustive or the effect of having gained the valued object may end up re-establishing the condition that triggered the stimulus (that is, the person becomes satisfied). Indeed, it is hypothesized that emotions serve to mark information as more or less relevant, helping attention to converge to possible best courses of action based on experience, and thus allowing for taking decisions even in the presence of great uncertainty or a big number of options (Lazzaro, 2010; Damásio, 2012, p.163). In other words, "emotion is cognition's silent partner without which choices are simply impossible" (Lazzaro, 2010).

Table 2.3: *How chain of events re-establish equilibrium with the experience of some example emotions. Reproduced from Plutchik (2001, p.348).*

stimulus event	cognition	feeling state	overt behaviour	effect
threat	"danger"	fear	escape	safety
obstacle	"enemy"	anger	attack	destroy obstacle
gain of valuable object	"possess"	joy	retain or repeat	gain resources
loss of valuable object	"abandonment"	sadness	cry	reattach to lost object
member of one's group	"friend"	acceptance	groom	mutual support
unpalatable object	"poison"	disgust	vomit	eject poison
new territory	"examine"	expectation	map	knowledge of territory
unexpected event	"what is it?"	surprise	stop	gain time to orient

The cognitive part of the emotional process, that is, the feelings, are also classified in terms of a three-dimensional space formed by valence, arousal and tension (Scherer, 2005, p.718). Valence reflects the attractiveness of the feeling, ranging from negative (unpleasant) to positive (pleasant). Arousal reflects the intensity of the feeling, ranging from very calm (or sleepy) to very exciting (or energized). And Tension reflects the potency or control of the feeling, ranging from relaxed to tense (Scherer, 2005, p.718). Since it is difficult to consistently identify the third dimension (tension) from the second one (arousal), most researchers use only the first two dimensions combined in a circular structure to map all possible feelings in this space (figure 2.6).

Positive valance is usually preferred and considered as naturally pleasurable. However, both positive and negative valences are important for the experience of fun because they compose optimal emotional patterns consistent with the ones in real life (Ravaja *et al.*, 2006, p.344). Fear and anger, for instance, are two emotions that involve negative-valenced feelings. The consequence of conquering fear or overcoming enemies after reaching emotional equilibrium – which is important for the subsistence in real threatening conditions – is made stronger in entertainment and playful contexts because it is accompanied by the acknowledgement that it was experienced in a safer environment, detached from the risks of real life (Ravaja *et al.*, 2006, p.344). This partially explains why experiences with horror films, games and roller coasters, for instance, are considered as pleasurable despite the elicitation of negative feelings such as those.

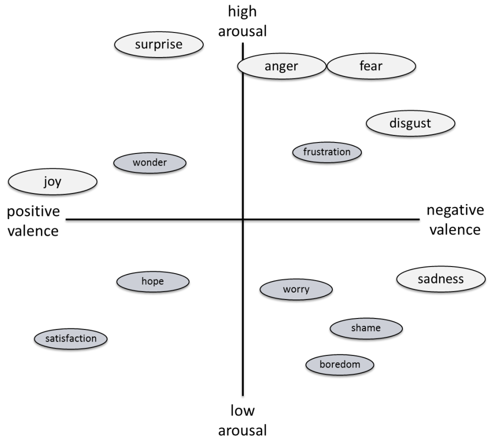


Figure 2.6: *Illustration of emotions in the discrete and dimensional spaces. Based on Scherer (2005, p.720)*

Consequently, one difficulty of this dimensional description is that classifications of feelings using valence and arousal may be less conclusive than directly using prototypic emotions, since in cases like fear and anger (both negative-valenced, high arousal) the feelings are very close in the bi-dimensional space and yet are related to very different emotions (Scherer, 2005, p.718). So usually the use of discrete and basic terms are preferred to describe emotions when discussing or studying them, since even users are more accustomed to refer to their feelings in terms of prototypic emotions (Scherer, 2005, p.719).

In conclusion, when games are played, emotions have five roles (Fullerton, 2008, p.258). First of all, players *enjoy the sensations* created because the physiological changes are pleasurable. Emotions *focus attention*, hence helping progression. Frustration, for instance, is an emotional state that arises as a response to perceived difficulties to reach a goal, that can either result in anger and disappointment or inspire and motivate if an individual's abilities matches the challenge ahead (Canossa *et al.*, 2011, p.61). They also *affect performance* by facilitating repetitive behaviour through the experience of negative and positive feelings, hence also *reward and motivate learning.* Finally emotions aid in *decision making*, by turning the consequences of two options (equally good from a mere rational perspective) easier to be compared.

2.2 A Design View of Fun

2.2.1 User Experience

In the Interaction Design, product designers and HCI researchers have been traditionally concerned with the achievement of design objectives taken from the user's perspective, what is called *Usability.* These design objectives include utility (provision of features needed), effectiveness and efficiency (good results and their easy achievement), safety (prevention of serious errors and easy recovering from them), learnability (easy learning on how to use a product), memorability (easy remembering on how to carry out tasks) and satisfaction (freedom of discomfort and positive attitudes towards a product) (Preece *et al.*, 2002, p.14–17).

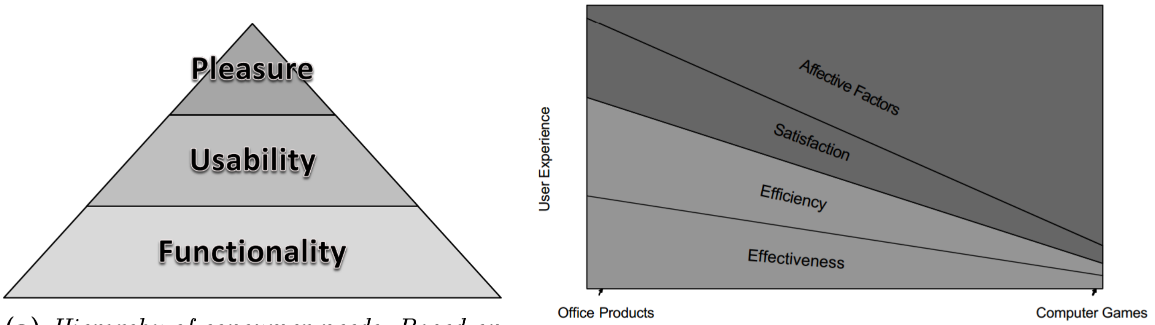
However, for more than two decades there has been a shared understanding that satisfaction is not just about the absence of discomfort and positive attitudes due to the achievement of practical goals. It should also include non-utilitarian aspects of the interaction, such as pleasure, appeal, preferences and emotions (Preece *et al.*, 2002, p.19; McCarthy and Wright, 2004, p.5; Calvillo-gámez *et al.*, 2010, p.51). The reasons are that actions and results are respectively shaped and interpreted by unique values and feelings (McCarthy and Wright, 2004, p.83–85), and also that users take functional features and quality for granted (Hassenzahl, 2005, p.31), immediately searching for more (like for aesthetic pleasure and emotional benefits) once the utility of a product is acknowledged (Jordan, 2002, p.6).

User experience (UX) is the subjective relationship between user and product that includes those non-utilitarian aspects as a newer view on Interaction Design (Calvillo-gámez *et al.*, 2010, p.45; Bernhaupt, 2010, p.4). The design objectives in UX are making products enjoyable, fun, entertaining, motivating, aesthetically pleasing, supportive of creativity, rewarding and emotionally fulfilling (Preece *et al.*, 2002, 18), and this is sought by focusing attention on human factors that consider the user as more than just a physical and cognitive system component and the product as more than a tool (Jordan, 2002, p.7; Forlizzi and Battarbee, 2004, p.261).

Usability and UX are not opposed concepts. Usable products will not necessarily be pleasurable, but if a product is not at all usable it is unlikely that it will provide any positive experiences or allow for immersion and engagement (Jordan, 2002, p.6; Brown and Cairns, 2004, p.1300). Indeed, it has been shown that in order for people to allow themselves to enjoy using a product, they first need to achieve high levels of efficiency, effectiveness and satisfaction regarding their own goals, which requires them to at least be able to easily understand how the product works (Bentley *et al.*, 2002, p.230). This is strongly consistent with the feedback requirement in Flow, discussed earlier in this chapter. Thus Usability, in its more operational and specific character, can be seen as a key component to positive experiences, and in its turn UX can be seen as an aggregation of flexibility in favour of more meaningful experiences and the fulfilment of the needs of different users (Preece *et al.*, 2002, p.19–20). This dependency relation has been compared to Maslow's (1943 *apud* Jordan, 2002, p.) "Hierarchy of Human Needs", indicating that the *consumer needs* (figure 2.7a) evolve in a similar fashion: users of a product first need to crave satisfaction regarding any functional aspects (i.e. having the need or desire to interact with it), to then achieve satisfaction regarding easy of use (i.e. being able to interact with it) and finally reach pleasure (i.e. enjoying interacting with it).

Nevertheless, the frontiers between those "consumer needs" are not straight lines, and a prober balance needs to be searched. It is very common that some utilitarian objectives are *intentionally unmet* when designers particularly aim for UX, and in a matter of fact some combinations of utilitarian and non-utilitarian objectives are not even possible or desirable (Preece *et al.*, 2002, p.19). For example, products like puzzles and video games are built to not be easy to use or to require more effort than really necessary, and products like control systems simply can not (or, at least, should not) be designed to be simultaneously safe and fun. Also, fantasy and genre preferences have been reported to strongly affect how much feedback is *perceive as required* from a given activity, corroborating the idea that users are more willing to put physical or cognitive effort into figuring out what needs to be done when using products for entertainment (Bentley *et al.*, 2002, p.230). Additionally, utilitarian aspects that add value to satisfaction in first interactions may become detrimental with the use, eventually leading to frustration (Hassenzahl, 2005, p.33). For example, an interaction with an automated teller machine using big sequences of small steps is always easier to understand, but this attribute becomes less relevant when the user gets more experienced with the product or is in a condition of time pressure.

As consequence, the continuum between how important are utilitarian and non-utilitarian design objectives to characterize satisfaction is very fuzzy, depending on the product's purpose and the situation of use. The figure 2.7b illustrates how Bentley *et al.* (2002) distinguish the relevance of some of the Usability and UX factors between the extremities that they called *Office Products* and *Computer Games*. The idea is that efficiency, effectiveness, satisfaction (considering the absence of discomfort and the achievement of both external and internal goals) and affective factors (the rest of the non utilitarian design objectives related to pleasure and satisfaction) are relevant to any kind of product, though their relevance changes as the product's purpose – as intended by the designer *and* perceived by the user – moves along this axis. Utilitary aspects are naturally more important for the satisfaction in "serious" products, and pleasure and fun for entertainment products.



(a) *Hierarchy of consumer needs. Based on Jordan (2002, p.6).*

(b) *Notional divisions of the User Experience. Based on Bentley et al. (2002, p.239).*

Figure 2.7: *How utilitarian and non-utilitarian aspects of experience intersect in products.*

Thereby, the HCI community recognizes that both utilitarian and non-utilitarian aspects of a product need to be combined, and more importantly, that positive experiences can not simply be guaranteed: it is only possible to design *for* an experience (Hassenzahl, 2004, p.47; McCarthy and Wright, 2004, p.8). As argued before, people may have different internal goals and experience ephemeral emotions as they interact with products, and that is why the design of positive experiences is seen as attempts to promote general needs (Hassenzahl, 2004, p.4) and to provide increasingly meaningful choices as the users interact with the products (Chen, 2007b, p.33). These conclusions can be observed in the very formal definition of UX provided by the ISO (2010, p.1) standard on Human-Centred Design for Interactive Systems:

> **2.15**
> **user experience**
> person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service
>
> Note 1 to entry: User experience includes all the users' emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviours and accomplishments that occur before, during and after use.
>
> Note 2 to entry: User experience is a consequence of brand image, presentation, functionality, system performance, interactive behaviour and assistive capabilities of the interactive system, the user's internal and physical state resulting from prior experiences, attitudes, skills and personality, and the context of use.
>
> Note 3 to entry: Usability, when interpreted from the perspective of the users' personal goals, can include the kind of perceptual and emotional aspects typically associated with user experience. Usability criteria can be used to assess aspects of user experience.

There are many models of UX that follow this complementary view of utilitarian and non-utilitarian aspects of experience emerging from context of use. A simple but useful model is the one proposed by Jääskö and Mattelmäki (2003). It combines different qualities that are most relevant to the user experience into two groups: the ones directly related to the product, like appearance and user interface, and the ones connected to the human-product relationship but dependent upon socio-cultural, time, physical, function or market contexts. The model is described as being useful in understanding and articulating the different aspects of experience as a whole, but particularly helpful in defining and comparing existing user data (Jääskö and Mattelmäki, 2003, p.127).

Another very comprehensive model of UX was proposed by Hassenzahl (2005). In this model the experience with a product is characterized from two distinct points of view, one from the designer and another from each user in a specific situation. When the designer creates a product, she chooses and combines features (content, presentation style, functionality and style of interaction) to convey a particular, *intended*, high-level description (also called *gestalt*) summarizing the product functions, strategies of manipulation and aesthetic elements. This description is composed of both pragmatic and hedonic attributes (Hassenzahl, 2005, p.34–36). The pragmatic attributes are related to the satisfaction of internal or external goals by the direct manipulation of the product, and the hedonic attributes are related to the creation of psychological well-being by the stimulation of curiosity, the development of skills and self image, and the evocation of fantasies and memories. When individuals come in contact with the product, they perceive its features and its attributes biased by their own objectives, mood, preferences and past experiences with it and with similar products, then constructing an *apparent* character.

In the apparent character, the pragmatic and hedonic attributes of the product may be perceived by the user as being either weak or strong, depending on the natural purpose of the product (the designer may have put more emphasis in one or the other aspect) and also on the context of use. Products that are primarily pragmatic are intrinsically linked to external or internal goals and are called ACT, and products that are primarily hedonic are intrinsically linked to ideals, preferences and memories and are called SELF (figure 2.8). The combination of strong pragmatic and hedonic attributes would be a desirable product, but they are unlikely to be in balance. Also, the appreciation of SELF products is much more stable (Hassenzahl, 2005, p.37). An example given by the author is that a simply and plain car originally bought to go to work may loose its utilitarian appeal after the user moves to a new apartment closer to the office. However, a luxurious sports car bought for personal reasons is not likely to have its appeal diminished by this type of specific change because it is intimately linked to the user's self.

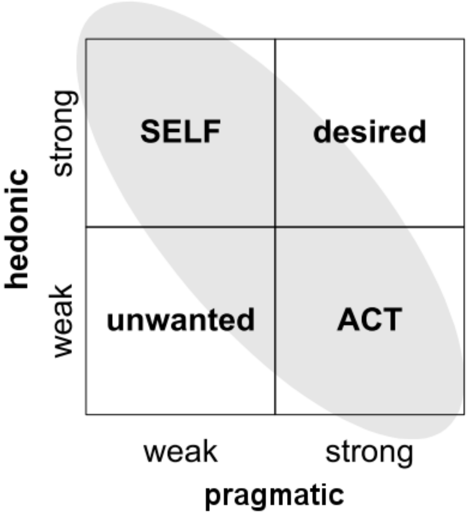


Figure 2.8: *Product characters that emerge from pragmatic and hedonic attributes. Based on Hassenzahl (2005, p.37).*

A third model that is worthy mentioning is the one proposed by Forlizzi and Battarbee. This model is based on an interaction-centred perspective, as opposed to the individual consideration of either the product or the user, in which the experience is seen as "a totality, engaging self in relationship with object in a situation" (Forlizzi and Battarbee, 2004, p.262). Therefore, the model encompasses three types of user-product interactions which yield three types of experiences in a given context of use.

The types of interaction are called Fluent, Cognitive and Expressive (Forlizzi and Battarbee, 2004, p.262), and are related to the different forms by which the user-product relationship happen. They have a strong connection to the expression of affection, being indeed a classification very close to the Visceral, Behavioural and Reflective levels of emotion used by Norman (2005) and discussed in the previous topic. The Fluent interactions are the most automatic and well learned, that do not compete for attention and hence allow the user to focus in the consequences of activities. The Cognitive interactions involve direct attention and effort, and can result either in knowledge or confusion depending on the matching between perceived possibilities of action and previous experiences. They cause changes in the self of the user by the development of skills and acquisition of knowledge, and often cause changes in the context of use as well by impacting other people's experiences. Finally, the Expressive interactions are the ones that help the formation of strong relationships with a product or some aspect of it. Through these interactions, users can create a better fit between themselves and the products by investing effort in changing it or by creating personal stories that are memorable.

These user-product interactions unfold in particular contexts and yield different types of experience, called Experience, An Experience and Co-Experience (Forlizzi and Battarbee, 2004, p.263). Experience is the plain result of interactions that happen all the time as users engage with products and other people. It is a "stream of conscious self-talk" that takes place as users assess their goals relatively to the products, the environment and other people surrounding them. An Experience is a particular meaningful and memorable experience, that can be articulated or named. It is characterized from a number of interactions and emotions, having a beginning and an end and providing a sense of completion that inspires further emotional and behaviour changes. And finally a Co-Experience is about user experience in social contexts. They are experiences created together with other people when sharing attention. The final interpretation of the experience as positive or negative depends upon group agreement, making it a very complex result. Direct communication is not always required, since even the mere presence of others may affect individual outcomes.

When individuals interact with products their experiences dynamically flow between fluent, cognitive and expressive interactions. And, as time passes, smallest experiences are forgotten and only larger experiences, extremely emotional and that are connected to others, are remembered (Forlizzi and Battarbee, 2004, p.265). This *scalability of experience* is strongly related to the concepts of immersion and engagement previously studied. Users need to attain fluency with a product early on to ensure they will continue using it and not abandon in frustration, and this means to easily learn controls and be rewarded from start. Over time, the interaction should enable cognitive effort to flow in order to provide the chances for long-term emotional and behavioural responses. And, finally, the product should foster co-experiences through individual expression of results and mutual assistance. The perception of the use in social contexts may also create the opportunity for new experiences, and the cycle begins again, so the user experience also dynamically flows between experience, an experience and co-experience. The table 2.4 shows a summary of the types of interactions and experiences in this model, given examples of situations related to them.

Table 2.4: *Summary of the UX model as it relates to the design of interactive systems. Reproduced from Forlizzi and Battarbee (2004, p.263).*

Types of User-Product Interactions	**Description**	**Examples**
Fluent	Automatic and skilled interactions with products	• riding a bicicle • making the morning coffee • checking the calendar by glancing at the PDA
Cognitive	Interactions that focus on the product at hand; result in knowledge or confusion and error	• trying to identify the flushing mechanism of a toilet in a foreign country • using on-line algebra tutor to solve a math problem
Expressive	Interactions that help the user form a relationship to the product	• restoring a chair and painting it a different colour • setting background images for mobile phones • creating workarounds in complex software
Types of Experience	**Description**	**Examples**
Experience	Constant stream of "self-talk" that happens when one interact with products	• waling in the park • doing light housekeeping • using instant messaging systems
An Experience	Can be articulated or named; has a beginning and an end; inspires behavioural and emotional changes	• going on a roller coaster ride • watching a film • discovering an on-line community of interest
Co-Experience	Creating meaning and emotion together through product use	• interacting with others with a museum exhibit • commenting on a friend's remodelled kitchen • playing a mobile messaging game with friends

From what has been presented, designers understand that meaningful experiences have both an utilitarian and a hedonic character, which occur in the same form of a feedback loop as emotions. Past interactions affect the future experiences creating changes in the general sentiment towards products or classes of products that may be either positive or negative (Calvillo-gámez *et al.*, 2010, 51). Therefore, a sequence of bad experiences (even if just related to aesthetic preferences) can be harmful for engagement even to the most useful of products. In the other hand, sequences of positive experiences make engagement stronger and do not seem to have negative effects to Usability. For instance, it has been shown that elements of non-aggressive humour, in the form of jesting error messages or fantastic supporting characters, may increase the satisfaction with a product without causing advert effects like distraction or lack of seriousness (Morkes *et al.*, 1999, p.419).

Indeed, the aesthetics of the user interfaces are known to have effects on how users perceive the product's Usability, because the real utility of a product is the result of the perceptions users have regarding the task difficulties and their own abilities to address them (Preece *et al.*, 2002, p.144). Also, being happy is known to help improving health and well-being in the sense traditionally aimed from ergonomics (Carroll, 2004, p.39). That is why Flow, immersion, beauty and aesthetics are considered important dimensions of the user experience (Bentley *et al.*, 2002, p.229; Fierley and Engl, 2010, p.207). Because of those reasons fun is becoming a very important aspect of user experience, if not an almost direct synonym of satisfaction in the context of Interaction Design, as the motivator for discretionary and sustained use (Carroll, 2004, p.38; Marcus, 2007, p.49; Goodman *et al.*, 2012, p.23).

2.2.2 Game Design Models and Heuristics

In the domain of Game Design, the user experience is considered to also depend upon additional product attributes such as rules of play (mechanics), storytelling techniques, aesthetics and even technology (particularly in the case of video games), that are not totally handled by traditional Usability or UX and yet are important for the experience of fun with games (González *et al.*, 2009, p.2; Schell, 2008, p.41). They are so relevant to the player experience and intimately related that games are commonly described as being formed from those basic elements (the so called *elemental tetrad*, illustrated in figure 2.9) (Schell, 2008, 41–43):

- **Mechanics**
  Mechanics is the set of procedures and rules of a game that describes its goal and how players are allowed or not to try to achieve it, and an important element of challenge. It strengthens story and makes players immersed in the world defined by aesthetics, requiring technology to be constructed.
- **Story**
  Story is the sequence of events that occur in a game, and an important element of fantasy. It turns mechanics meaningful to players, makes aesthetics emerge at the right moment and cause the most impact to the experience, and is conducted by technology.
- **Aesthetics**
  Aesthetics is how the game feels to the senses, and an important element of curiosity and immersion. It emphasizes mechanics, reinforces ideas of story and explores the capabilities of technology to dazzle the senses.
- **Technology**
  Technology is the medium throughout all interactions occur. It allows aesthetics to take place, mechanics to happen and story to unfolds.

An important observation needed at this moment is that the word “aesthetics” can have different meanings in the Game Design, referring to the sensory phenomena (visual, aural, haptic and embodied), the appreciation aspects shared with other forms of art, or the broad expressions of player experience (as pleasure, emotion, sociability, etc) (Niedenthal, 2009, p.2). The meaning used in this chapter and intended in the elements just described is the first one (sensory phenomena).

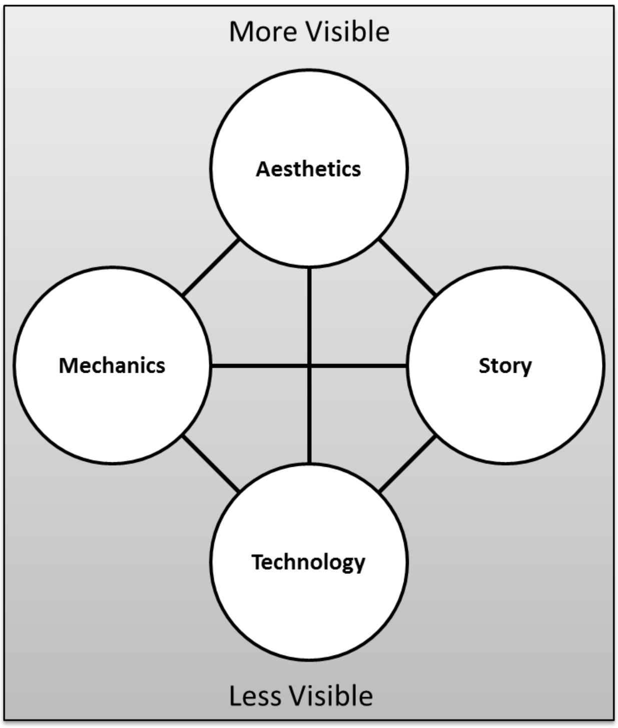


Figure 2.9: *One way of breaking down the basic elements that form a game. Based on Schell (2008, p.41).*

As consequence of this view, much of the discussion of user experience in the area has been performed using the terms *Gameyplay* and *Playability*. Those terms are not clearly defined, being ambiguous and many times used indistinctly to refer to all the experiences players have with games.

Gameplay is commonly described as the style or pattern of interactions that emerge during a session in which a specific user plays a game alone or with others, characterizing the experience in terms of the game mechanics (Nacke *et al.*, 2009, p.1; Mello and Perani, 2012, p.162). It refers to the “gaming process”, or the path of choices taken by the player as she plays a game and tries to achieve goals (Mello and Perani, 2012, p.162). This is indeed the term most naturally used in non-scientific publications, like magazine reviews and tutorial videos, to refer to game trailers, main plots or quality ratings[6], and in scientific publications to refer to sessions of a game.

On the other hand, Playability refers to the qualities of a game that determine how “playable” it is, characterizing the experience in terms of all game elements of interface, mechanics, fantasy and challenges (Nacke *et al.*, 2009, p.1; Mello and Perani, 2012, p.162). So it refers to the design choices that make easier or harder for fun to happen in the interaction with the game, in a very similar fashion to the complementary use of Usability and UX described earlier in this chapter.

In other words, Playability is an extension of Usability and UX with specific methods or “game systems” focused on the technical and learning aspects of games to help improving their design, whereas Gameplay is the intended fun experience (Nacke *et al.*, 2009, p.1;2). As mentioned before, the terms are both related to the design and evaluation of the player experience, but with Playability being the “hidden structure through which the player’s participation on game environment allows the activation of the Gameplay” (Mello and Perani, 2012, p.162). A formal definition of Playability is provided by González *et al.* (2009, p.2):

> Playability represents the degree in which specific player achieve specific game goals with effectiveness, efficiency, flexibility, security and, especially, satisfaction in a playable context of use.

[6] according to Webopedia (http://www.webopedia.com/TERM/G/gameplay.html) and Collins English Dictionary (http://www.collinsdictionary.com/dictionary/english/gameplay).

The "systems" employed to guide the achievement of Playability are very similar to the UX models presented before, but in the area of Game Design they commonly include heuristics. Heuristics are simply an aggregation of rules defining key aspects of design (Preece *et al.*, 2002, p.26; Brown, 2010, p.80), and are a tool that has been traditionally used for judging the compliance of recognized Usability principles in software interfaces. They are drawn upon past experience and used in two ways: during design to choose among different alternatives and during evaluation to find and justify interface problems (Preece *et al.*, 2002, p.26). The most well-known set of heuristics was proposed by Nielsen (1994 *apud* Preece *et al.*, 2002, p.27), including ten Usability principles:

- **Visibility of system status**.
  Always keep users informed about what is going on, through providing appropriate feedback within reasonable time.
- **Match between system and the real world**.
  Speak the users' language, using words, phrases and concepts familiar to the user, rather than system-oriented terms.
- **User control and freedom**.
  Provide ways of allowing users to easily escape from places they unexpectedly find themselves, by using clearly marked 'emergency exits'.
- **Consistency and standards**.
  Avoid making users wonder whether different words, situations, or actions mean the same thing.
- **Help users recognize, diagnose, and recover from errors**.
  Use plain language to describe the nature of the problem and suggest a way of solving it.
- **Error prevention**.
  Where possible prevent errors occurring in the first place.
- **Recognition rather than recall**.
  Make objects, actions, and options visible.
- **Flexibility and efficiency of use**.
  Provide accelerators that are invisible to novice users, but allow more experienced users to carry out tasks more quickly.
- **Aesthetic and minimalist design**.
  Avoid using information that is irrelevant or rarely needed.
- **Help and documentation**.
  Provide information that can be easily searched and provides help in a set of concrete steps that can easily be followed.

Heuristics are very helpful in the design of games because they serve as simple questions or guidelines that can be easily interrogated during the creation and evaluation of all aspects of Playability (Federoff, 2002, p.15; Brown, 2010, p.80). Since fun is a more subjective matter than pragmatic issues of Usability, heuristics have always seemed a good way of describing the experience in games. In fact, heuristics have always been used by game designers, even if just in an informal way (Brown, 2010, p.85).

The development of formal heuristics that included Playability qualities started with the intrinsic motivational aspects of Curiosity, Fantasy and Challenge proposed by Malone. The author himself helped extend his characterization of fun to include Control (feeling that outcomes are determined by actions), Competition (comparison of skills), Cooperation (opportunities to work with others) and Recognition (acknowledgement of the purpose of interface elements regarding mechanics and story) (Malone and Lepper, 1987).

A couple of years later, Federoff (2002) collected individual heuristics from the Game Design literature available at the time and divided then into three groups: Interface, Mechanics and Gameplay. Interface heuristics are related to the devices through which players interact with the game, Mechanics heuristics are related to the ways the player is allowed to perform actions like move through the game environment, and Gameplay heuristics are related to the problems and challenges that players must face to try to win the game (Federoff, 2002, p.12). Interface and Mechanics heuristics are closer to the classic Usability principles, and Gameplay heuristics are more specific to the domain of games. By observing and interviewing a team of game developers as they work, the author compiled a final list of forty heuristics in the form of *expert advices*, like for instance (the rest of the heuristics are available at the original reference):

> (Game Interface) Use sound to provide meaningful feedback
> (Game Mechanics) Feedback should be given immediately to display user control
> (Gameplay) There should be variable difficulty level

Sweetser and Wyeth (2005) also attempted to organize the many heuristics employed so far by game designers into a single and more concise model. In their research they found out that the existing heuristics overlapped the concepts of the Flow Theory, adding up to specific features of games like immersion and social context. So, they have summarized them all in a model named GameFlow and composed of principles separated in eight core elements adapted from the aspects of Flow, with the game being the task to be completed (Sweetser and Wyeth, 2005, p.4): Concentration (ability to concentrate on the task), Challenge and Skills (perceived skills matching challenges and both exceeding a certain threshold), Control (exercise of sense of control over actions), Clear Goals (existence of clear goals), Feedback (provision of immediate feedback), Immersion (deep but effortless involvement, with reduced concern for self and sense of time), and Social Interaction (opportunity to interact with others). The Social Interaction element is as important as the others, even though it is not directly derived from Flow. It is importance is due to the observation that people may even play games they dislike just to do it with others (Sweetser and Wyeth, 2005, p.4). The heuristics proposed are also in the form of expert advices, like for instance (the rest of the heuristics are available at the original reference):

> (Concentration) Games should provide a lot of stimuli from different sources
> (Challenge) Challenges in games must match players' skill levels
> (Player Skills) Players should be able to start playing without reading the manual
> (Control) Players should feel a sense of control and impact onto the game world (like their actions matter and they are shaping the game world)
> (Clear Goals) Overriding goals should be clear and presented early
> (Feedback) Players should receive immediate feedback on their actions
> (Immersion) Players should become less aware of their surroundings
> (Social Interaction) Games should support competition and cooperation between players

After that, Schell (2008) elaborated a very comprehensive and professional set of heuristics. According to the author, the good game design happens when the designer views the game from as many perspectives as possible, like "lenses"(Schell, 2008, p.xxvi). So, he created one hundred cards (the lenses), each one composed of a small number of questions the designer should ask herself about the game during the whole creation process. Those questions are not direct advices, but they included well known practises that help the design freely brainstorming and explore the huge number of requirements without having to worry remember them all. Even though card tools like this are mainly helpful for the creative process of design, they can also aid in the iterative design process. They are portable collections of ideas that can be taken to testing sessions to help in interviews, and can be also shared with users to collect more directed impressions on the experience (Baldwin, 2011).

The Game Design Lenses and their questions are the result of the development of a holistic understanding about the satisfaction in games regarding different aspects of experience, players, interface, game, process and designer. Consequently, there are lenses addressing subjects like curiosity, fantasy, challenge, goals, mechanics, aesthetics, story, technology, perception of time, attention, Flow (as a separated lens), perception of possibilities of actions, emotions, business motivations, architectural qualities, and so on. Many lenses are indeed redundant. For instance, there are specific lenses for Flow, Goals, Skill and Challenge (Schell, 2008, p.122;149;153;179). But that redundancy is good in the sense that it provides different opportunities for constant refining the design choices, through increasing levels of detail in the questions. These are some samples of the lenses (the rest of them are available at the original reference):

> (#18: The Lens of Flow) Does my game have clear goals? Are the player's skills improving at the rate I had hoped?
> (#25: The Lens of Goals) Are my goals concrete, achievable and rewarding?
> (#27: The Lens of Skill) What skills does my game require from the player? Which are dominant?
> (#31: The Lens of Challenge) Can my challenges accommodate a wide variety of skill levels?

Parallel to the development of those heuristics, researchers have been proposing specific models for the design of games. Though, they are largely based on the knowledge learned from studying such heuristics. One model was proposed by (Hunicke *et al.*, 2004) and called MDA after its components of Mechanics, Dynamics and Aesthetics. Mechanics, as previously described, is related to the particular constructional elements of a game (rules). Dynamics is related to the run-time behaviour of the mechanics, acting on players inputs and each other's outputs over time (system). And Aesthetics, in accordance to a more comprehensive meaning of the player experience mentioned before, is related to the emotional responses evoked in the player ("fun").

The model formalizes the *consumption* of games by each individual component, supposedly making it easier for the game designer to focus efforts in each part using the most appropriate approaches while keeping in mind that all constructed artefacts are intended to evoke those emotional responses in the player. Instead of having heuristics as phrases or rules, the Aesthetics component describes a vocabulary defining the experience intentions in a more precise way than "fun", clearly based on the aspects of intrinsic motivation and immersion. The terms of this vocabulary are presented in table 2.5.

Table 2.5: *Game aesthetics. Reproduced from Hunicke et al. (2004, p.2).*

Sensation Game as sensory pleasure	**Fellowship** Game as social framework
Fantasy Game as make-believe	**Discovery** Game as uncharted territory
Narrative Game as drama	**Expression** Game as self-discovery
Challenge Game as obstacle course	**Submission** Game as pastime

Another model of fun was created by Lazzaro (2004) based on four elements derived from emotions. According to the author, emotions are responsible for the experience of fun because all other components come from them. Games may have many rules regarding challenges, goals and the ways to achieve them, but there certainly are simple games with just a few or a single rule (the children game of tag, for instance) with still elicit a lot of emotions that players enjoy (Lazzaro, 2010). So the game designer's role is most of all to create the opportunity for the experience of emotions.

Her model was created from the observation of many emotions that resulted from gameplay, noticed in the facial gestures, bodily language and verbal comments of participants in experiments in which they played their favourite games. The emotions observed altogether with the related bodily and behavioural responses are listed in table 2.6.

Table 2.6: *Emotions experienced during play. Reproduced from Lazzaro (2004, p.6).*

Emotion	Common Themes and Triggers
Fear	Threat of harm, object moving quickly to hit player, sudden fall or loss of support, possibility of pain
Surprise	Sudden change. Briefest of all emotions, does not feel good or bad, after interpreting event this emotion merges into fear, relief, etc.
Disgust	Rejection as food or outside norms. The strongest triggers are bodily products such as feces, vomit, urine, mucus, saliva, and blood.
Naches/Kvell (Yiddish)	Pleasure or pride at the accomplishment of a child or mentee (Kvell is how it feels to express this pride in one's child or mentee to others).
Fiero (Italian)	Personal triumph over adversity. The ultimate Game Emotion. Overcoming difficult obstacles players raise their arms over their heads. They do not need to experience anger prior to success, but it does require effort.
Schadenfreude (German)	Gloat over misfortune of a rival. Competitive players enjoy beating each other especially a long-term rival. Boasts are made about player prowess and ranking.
Wonder	Over whelming improbability. Curious items amaze players at their unusualness, unlikelihood, and improbability without breaking out of realm of possibilities.

From those observations and her past experience designing games, the author summarized fun in games as being the product of four key elements (Lazzaro, 2008, p.258):

> People play games in four ways. They enjoy the opportunity to master a challenge and to fire their imaginations. Games also offer a ticket to relaxation and an excuse to hang out with friends."

In each key, different emotional responses are alternated in a way that fun can is achieved by reaching emotional equilibrium. For instance, having fun by overcoming challenges requires players to be first frustrated to the point of almost being ready to quit, when then they suddenly succeed. There is a huge phase shift in the body that goes from feeling frustration to feeling very good (Lazzaro, 2008, p.259).

The first key is called Hard Fun (*Fiero*). It provides the opportunities for challenge, mastery and feelings of accomplishment. The actions are driven by goals, obstacles and strategies, leading to emotional responses alternating from frustration to relief. The second key is called Easy Fun (*Curiosity*). It inspires exploration and role play. The actions are driven by exploration, fantasy and creativity, leading to emotional responses alternating from curiosity, surprise, wonder to awe. The third key is called Serious Fun (Excitement). It induces changes in how players feel, think and behave, or make difference in the real world. The actions are driven by repetition, rhythm and collection, leading to emotional responses alternating from excitement to relaxation. And the fourth key is called People Fun (Amusement). It provides the excuse to hang out with friends. The actions are driven by communication, cooperation and competition, leading to emotional responses alternating from friendship, amusement to admiration.

2.3 So, what is fun?

In essence, fun can be conceptualized by a combination of utilitarian and hedonic aspects. The former is directly related to performance, being about interesting physical or cognitive challenges and the ways to overcome them, as well as the recognition of interesting but yet unknown patterns in the world. The latter is related to affect, thus about the emergency of empathy with objects and characters and, specially, the experimentation of emotions. Both of them are related to the gradual focus of attention, either to focus on tasks at hand or awe from sensory and imaginative information. The utilitarian aspect of fun does not exist without the hedonic aspect. First of all, performance requires involvement, so the needed skills can be experienced and learnt. But also because emotion is necessary for actions (Lazzaro, 2010). Consequently, fun can be defined by two *essential dimensions*, as it is summarized by figure 2.10: immersion, as the increasing level of involvement with the task, that subsumes all the sensory, cognitive and emotional systems, and the emotions experienced by performing the task, which can be considered pleasurable despite their general likeness in other contexts.

Among the emotions that are important for the experience of fun there are the six prototypic emotions and also the pleasurable emotions named *Fiero*[7], *Naches/Kvell*[8] and *Schadenfreude*[9]. But even though these pleasurable emotions can be distinctly recognized by players experiencing them, they seem to have roots on the prototypic emotions. For instance, the bodily changes produced by *Fiero* are very similar to the ones that accompany extreme Anger, while *Schadenfreude*'s bodily changes have been shown to not differ from Joy's (Boecker *et al.*, 2015). Therefore, the prototypic emotions are arguably the most relevant emotions to the experience of fun.

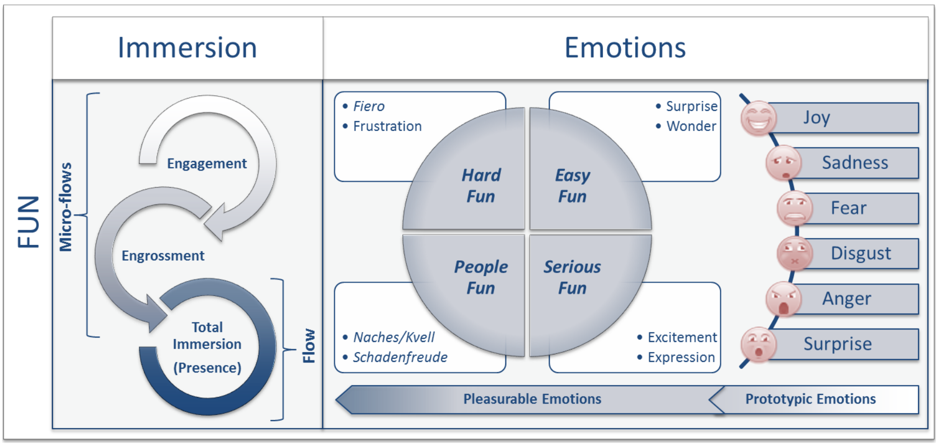


Figure 2.10: *Fun as the combination of immersion and emotions*

[7] from the Italian: personal triumph over adversity or adversary

[8] from the Yiddish: pleasure from pride/bragging at the accomplishment of a mentee

[9] from the German: gloat over the misfortune of others

Chapter 3

Existing approaches to assess fun

The intention of this chapter is to study the existing efforts in assessing fun, particularly in an automated fashion. Since video games are the focus of this research work, it feels natural to start this study with an overview on the practices currently used in the area. Besides introducing the subject of this chapter, it serves as a proper continuation from the previous chapter.

3.1 The Evaluation of Fun in Games

As seen in the previous chapter, positive experiences with interactive systems can not simply be guaranteed by any design choices, due to the constitutional subjectivity character and context dependency of experiences. Product attributes can only try to avoid bad experiences, while provide the means by which pleasurable ones can occur. That is why the core of the Interaction Design process, in an user centric approach, is the constant evaluation of the product with the active participation of users. This iterative process involves four basic activities, that inform one another (Preece *et al.*, 2002, p.12):

1. Identifying needs and establishing requirements.
2. Developing alternative designs that meet those requirements.
3. Building interactive versions of the designs so that they can be communicated and assessed.
4. Evaluating what is being built throughout the process.

Evaluation, that is, testing the designs regarding Usability and User Experience (UX) objectives, is the heart of the interaction design because it focus on ensuring that the product is usable and potentially pleasurable (Preece *et al.*, 2002, p.12). This is achieved by observing users, talking to them, interviewing them, testing them using performance tasks, modelling their performance, asking them to fill in questionnaires, and even asking them to become co-designers (Preece *et al.*, 2002, p.13).

In the video game industry, the one particularly interested in the design of interactive systems that are fun to use, tests are performed with distinct purposes depending on the stage of the development process. For instance, they are used in design review to internally experiment and discuss ideas, in quality assurance to verify and remove bugs and in marketing assessment to estimate interest and sales (Fullerton, 2008, p.248). *Playtesting* refers to the type of tests that are performed in a larger scope, namely throughout the entire development process, in order to understand the kind of experiences players will have during play sessions and to evaluate the game potential for fun (Fullerton, 2008, p.248; Bernhaupt, 2010, p.6). Designers use a lot of prototypes of different fidelities to evaluate their products throughout the steps enumerated above. In the game design, Playtests can be seen as prototypes of the fun experience (Schell, 2008, p.392).

Playtesting used to be performed in a very informal manner (Nacke *et al.*, 2009, p.2), with most developers simply testing their games themselves or with the help of friends and family. And that approach is known to bring problems to the resulting experience. Designers are too close to the game to the point of having distorted opinions, and friends and family don't want to hurt designers' feelings thus providing untruthful predispositions towards the experience (Schell, 2008, p.393). Nowadays, it is more common for even small developers to involve participants representative of the intended audience and to employ methods borrowed from the Human-Computer Interaction (HCI) for both quantitative and qualitative comparisons of players' behaviours. Such methods include focus groups, semi-structured interviews, observation and video coding, data hooking (automatic collection of performance measurements), questionnaires and heuristic evaluations of Playability (Fullerton, 2008, p.256–257; Nacke *et al.*, 2009, p.2; Bernhaupt, 2010, p.6). Among those, heuristic evaluation, observation, inquiries and data hooking seem to be the most used methods ((Fierley and Engl, 2010, p.206–208); Bernhaupt, 2010, p.6–7).

Expert analysis are common practises borrowed from the HCI, with Heuristic Evaluation being the most employed method. Besides serving as golden rules to try to predict what would be useful and interesting for the users during the design conception (as studied in the previous chapter), heuristics can be employed by several expert evaluators to role-play as users and critique the product regarding the principles stated by their rules (Preece *et al.*, 2002, p.409; Dix *et al.*, 2003, p.324). This technique does not have a direct involvement of real users, so archetypes called *personas* are usually employed to formalize the targeted audience of a game(Brown, 2010, p.81). However, even with good heuristics and well-defined personas it is difficult to foresee real bottlenecks in experience, since it strongly depends upon the context in which the product is used. Hence heuristics are used more frequently in informal manners to aid in other evaluation methods (Brown, 2010, p.), particularly in the preparation of interviews and in the production of inquiring questionnaires.

Observation is a method in which specialists watch users as they interact with the product to collect data and analyse the experience. The interaction can be observed and analysed as they are occurring, either in close or through a glass window, and from previous video recordings taken from the product's screen and the user's face and hands (Schell, 2008, p.397). The observation can be carried on either in a controlled laboratory, in which users perform the same pre-planned task, or in a field environment, in which the product is used more freely and in real conditions (Preece *et al.*, 2002, p.361–364; Dix *et al.*, 2003, p.327–329).

Laboratory studies allow the analysis of interaction details and make the recording and comparison of results a lot easier due to the controlled environment, hence being traditionally useful for the quantitative evaluation of Usability aspects. However, these tests are more artificial and make more difficult to observe real variations in qualitative aspects of user behaviour, since they are strictly directed and controlled. On the other hand, field studies allow the observation of real scenarios and may give useful insights into the emotional aspects of the experience. But these tests are usually more difficult to execute, requiring the designer to move herself and her equipment into the field where the tests are subject of interruptions and other forms of disturbing noise. Also, field studies have the additional difficulty that the conditions and events triggering interesting experiences are not easily reproduced (Preece *et al.*, 2002, p.362).

Field studies are usually preferred for Playtesting games because disturbances can indicate flaws in the intended design that are useful for its improvement (Dix *et al.*, 2003, p.328), for example showing the exact moments of Gameplay in which players loose interest or immersion or the reasons why aesthetic aspects are not interesting or appealing. Those types of tests are also useful when including multiple participants, since the social aspect is very important to entertainment. In an open environment, like the players' homes, play sessions can start and stop quickly and informally, and the presence of others – even if just for watching – certainly have impact in the player's interests, learning curves, perceptions of responses and moment-to-moment experiences (Schell, 2008, p.395; Isbister, 2010, p.13). Even so, there is usually a balance between laboratory and field studies, based on analysis of the loss of contextual information in laboratory against the increased costs and difficulties of field studies (Dix *et al.*, 2003, p.329).

The user experience always has a subjective and interpretative element arising from emotions and context, so the behaviours involved in experiencing fun can not be observed from outside the framework within they exist (Wood *et al.*, 2004, p.514). Once a researcher gets to observe the user experience she becomes part of the phenomenon that is being studied, a sort of Heisenberg uncertainty principle (Dix *et al.*, 2003, p.328). Therefore, no matter where the location is, observation has to be executed carefully to avoid tampering the experience as much as possible. The presence of an observer may be dangerous to a test particularly by infecting the user with the designers emotional investment in the product or by interrupting the experience for the collection of data (Schell, 2008, p.397). Also, if users easily acknowledge that they are being observed (what is more common in laboratory environments), they may feel uncomfortable, afraid of making mistakes, or to make noises that would disturb others or embarrass themselves (Schell, 2008, p.394).

That is why inquiring techniques are also very used to evaluate the user experience. An example of inquiring technique is the Think-Aloud protocol (Preece *et al.*, 2002, p.365), that requires users to say out loud everything that they are thinking, planning and feeling as they interact with the product. The designer involvement is minimum, practically without interruption of the user activity. However, that approach has also its problems. Users may still feel embarrassed to do it, and it is not easy for them to keep speaking for long, because as the demands of the activity increase there is less freed attention to be used for vocal externalization of thoughts and feelings (Preece *et al.*, 2002, p.367). The most employed alternative is to use questionnaires and interviews taken before and after participants interact with the product, so that information can be gathered directly from the users (self-reported) about expectations and their fulfilment without possibly disrupting the natural play patterns by asking questions during the Gameplay (Schell, 2008, p.399). The drawback of these approaches is that by the time the test session is finished the user mind is no longer in the exact same state as it was when the product was in use. There is a difficult trade off in the choice of interrupting or not users as they interact to ask questions. The general agreement is that fun requires immersion, and hence most game designers are in favour of only interrupting players if they are doing something really surprising that it is not understandable (Schell, 2008, p.399).

In the attempt to make easier the evaluation while avoiding interruptions, automated recordings are also largely employed. Data hooking involves the logging of timestamped quantitative numerical values programmed in auxiliary devices or directly in the game (sometimes even collected via Internet) so to completely avoid the presence of the designer. The captured data commonly includes performance measurements like number of hits and misses, total kills and “head shots” accounted during the play sessions and called *Gameplay metrics* (Nacke *et al.*, 2009, p.2). Such measurements are objective and numerical, can be captured in large numbers and are easily mapped to specific points in games, being thus helpful in assessing how playable a game is regarding its utilitarian aspects (Nacke *et al.*, 2009, p.2), like efficiency in performing tasks or achieving goals and easy of understanding mechanics. However, by themselves they are not enough to explain non-utilitarian aspects of fun (Nacke *et al.*, 2009, p.2), like immersion and emotions.

With the evolution of technology related to biometric sensors and digital cameras, it is now also common to collect data from psychophysiological measurements like skin conductance, cardiovascular and respiratory activity and muscle contractions (Mandryk *et al.*, 2006, p.3–4; Nacke, 2013, p.2), as well as non-verbal behaviours of bodily movement, vocal tone and facial expressions (Scherer, 2005, p.709–712; Mauss and Robinson, 2009, p.9–12; Fierley and Engl, 2010, p.206; Tan *et al.*, 2012, p.2). The valence and arousal dimensions of feelings are believed to covariate with such bodily changes (Scherer, 2005, p.718), hence the collection of this type of data may be helpful in improving the evaluation of fun with games specially if combined with other methods, like self-reported questionnaires and Gameplay metrics (Nacke *et al.*, 2008; Nacke *et al.*, 2009; Nacke, 2013; Tan *et al.*, 2012).

With this overview concluded, the rest of this chapter presents some of the existing efforts in automating the evaluation of fun from its different composing aspects: attention, Flow, immersion and emotions. The evaluation of those aspects are not always totally automated nor targeted to evaluate fun, but they provide a good review on what has been attempted so far.

3.2 Assessing Attention

Attention has been traditionally assessed from performance tests with users, by measuring reaction time or number of omissions related to surprising (alertness) or frequent (vigilance) stimuli, change of focus (divided attention) and classification (visual scanning and pattern filtering), or by direct observing behavioural cues as users perform activities (Lamar and Raz, 2007, p.290–291; Stanley, 2013, p.8–11). It does not seem to exist self-assessment questionnaires inquiring users if they “felt” to be attentive, most probably because this is something that simply does not make much sense: whenever questioned, users would most probably always respond to be aware of what they just did. Overt attention, the one that is associated with external stimuli and involves the movement of a sensory organ to capture the stimuli data (Stanley, 2013, p.7), can be assessed by observing display behaviours related to bodily posture (leaning towards devices), facial expressions and eye movements (Stanley, 2013, p.16–18).

When humans watch each others’ faces, they pay much attention to the facial expressions and the gaze. They both can indicate many important things for communication, like intentions, active listening and pondering a point (Isbister, 2006, p.145–146), but gaze in particular can also indicate the direction of the focus of a person’s attention. Following gaze is an important behaviour for survival, helping in foreseeing dangers in the environment (Isbister, 2006, p.148). Indeed it has been already shown that head and gaze are both naturally directed towards focus of visual and auditory attention (Bidwell and Fuchs, 2011, p.4; Stanley, 2013, p.34). Leaning the body or moving the head also provides clearance regarding sensory data. It can indicate curiosity and increasing interest due to the user being in the state of Flow – when the movement is an attempt to receive more of the desired stimuli – or confusion and frustration from not being able to understand the information – when the movement is an attempt to resolve the source of frustration (Stanley, 2013, p.18).

One work that attempted to automate the assessment of attention in interactive systems from the analysis of facial expressions was proposed by Levialdi *et al.* (2007). Using as test-bed a video chat system without voice (just text messages and the visualization of peers’ faces), they combined the analysis of facial expressions with the counting of the number of mouse clicks and keystrokes in a given time interval, believing that attention is the result of being active. The absence of voice was intentional to avoid having facial muscle movement due to speech. Using a Naïve Bayes classifier and the tracking of facial features from a simple webcam, they first detected seven emotional states (neutral, happiness, sadness, anger, fear, surprise and disgust) to form a 7-component vector with the probabilities of the facial expression in the current video frame representing each prototypic emotion. Together with the 2-component vector of the number of mouse clicks and keystrokes in the defined period of time, they classified the attention in three levels: low, medium and high. Classifiers were trained from labelled images they collected and classified among users, after interviewing them about the emotions and levels of attention they experienced. Their results are argued to demonstrate that there is a correlation between the active expression of emotions and interaction with the system in the level of attention involved.

Another relevant work was made by Stanley (2013), attempting to assess attention from bodily posture and gaze. Using Microsoft Kinect[1], the author collected bodily posture, gaze direction and speech activation and compared it to performance data collected from the execution of attention-competing tasks performed accordingly to well-established methods in psychology. The data recorded from Kinect included head distance (the proximity of the head to the screen), bodily

[1] http://www.xbox.com/en-US/kinect

orientation (the difference between left and right shoulder depth), head pose (the angle of the face relative to the location of the sensor), head position (the position of the head in relation to the sensor), forward and side body leans (angle and direction of back lean in relation to the hip and to the Y and X axis of the sensor) and talking (measurements whether the individual is talking during any particular moment). The comparison of the behavioural data to the performance data indicated that there is a strong correlation between eye gaze and attention, but that bodily posture requires multivariate measurements for a relatively good degree of predictive power. Also, there was no correlation between speech and attention, but that is most probably because the tasks used in the tests were entirely visual, requiring no auditory stimuli - what would not be the case with video games.

Eye blinking, pupil diameter and eye movement are other possible sources of information regarding attention. The spontaneous eye blinking rate decreases during high-attention tasks in order to maximize stimulus perception, but also when the "mental tension" is low due to task completion, serving as a relief mechanism (Chen and Epps, 2013, p.113). It is also related to the levels of dopamine (a chemical neurotransmitter) activity, which occurs in the reward and pleasure centres of the brain, being also reported to decrease when individuals are exposed to attractive (that is, potentially more pleasurable) visual stimuli (Walla *et al.*, 2011, p.4; Aouaki, 2013, p.7;15). The eye blinking rates are co-related to the variations in heart-rate and skin conductance (Walla *et al.*, 2011, p.4) and are also activated in startle response, what suggests that it is representative of attention to both visual and acoustic stimuli (Walla *et al.*, 2011, p.5). The increasing in pupil diameter has also been shown to mark attentional shift (Laeng *et al.*, 2012, p.22). However, changes that reflect variations in cognitive activities are relatively small compared to changes due to other things, like light reflexes (Chen and Epps, 2013, p.112), what makes pupil responses potentially not as robust as eye blinking to lightning variations. The eye movement is characterized in terms of fixation and saccade. Fixation is the stationary position of the pupil on a region of interest, and saccades are the rapid eye movements from one position to another. They occur in turns when eyes are viewing a scene, and can indicate the degree of importance of elements (Chen and Epps, 2013, p.113).

One work that uses these methods was proposed by Chen and Epps (2013). They attempted to measure attention from pupil changes, eye blinking and movement in terms of cognitive load, but comparing the effect of emotional responses. The cognitive load was induced by arithmetic tasks and the emotional responses by displaying images labelled to variations of valance, arousal and dominance, from the International Affective Picture System (IAPS) public database (Lang *et al.*, 2005). Pupil dilation and position were tracked with a commercial application, but blinks were processed from video recordings with scripts developed in MATLAB due to inaccuracy of the tracker. Fixation and saccade measurements were extracted from the pupil positions recorded from the tracker, and subjective ratings of difficulty and emotion were collected from participants in a self-reporting manner. Using as features the zero crossing count of pupil size and the position and the cumulative numbers of blink, fixation and saccade, the authors performed a preprocessing analysis of variance and concluded that pupil diameter average was the feature most significant to the emotional arousal and cognitive loading. Then, they trained a Gaussian Mixture Model (GMM) classifier, and confirmed in the tests that pupil size and blink number increased with more difficult tasks, and that pupil size also increased with higher arousal regardless the valence of emotions. The authors argue that this suggests that the cognitive load dominates emotion in eye features. Inaccuracy problems involved too much head movement and frequent changes in distance from screen, and also light reflex noise - which significantly affect the classification due to the strong relevance of the pupil size feature.

Other similar works are presented in section 3.4 (Assessing Immersion), since even though they have some relevance to the evaluation of attention their intended purpose is more related to the assessment of immersion.

3.3 Assessing Flow

Flow has also been assessed through questionnaires. A notable example is the questionnaire created by Jackson and Marsh (1996) to measure Flow in sports and physical settings. It is a self-report set of thirty-six Likert scored questions asking participants, for example, on weather they felt challenged but with enough skills to meet the challenge, or if they knew what they wanted to achieve. But there has been many attempts to automate the assessment of Flow, specially in games due to the need for implementing Dynamic Difficulty Adjustment (DDA), that is automatically increase or decrease the difficulty of challenges, and Procedural Content Generation (PCG), that is automatically generate levels and scenario. In these cases, the assessment is based on the comparison of challenges and skills in terms of player and game performances (Hunicke and Chapman, 2004, p.92; Michael and Chang, 2013, p.5).

A representative proposal of this approach was made by Spronck *et al.* (2004). Their research aimed in balancing the difficulty of challenges by adjusting the tactics of enemies controlled by Artificial Intelligence (AI) from reinforcement learning. The game AI scripts are formed from weighted rules, with the probability of a rule being selected and executed being proportional to its weight. The rule database is adapted by updates in the rules' weights, performed according to the success or failure associated with the rule execution in encounters with real players. A fitness function for the AI-controlled characters is composed of performance indicators relative to victory/defeat, death/survival, amount of remaining health and of damage done to the enemies after each encounter. A reward or penalty is added or subtracted from the rule's original weight depending on the value of the fitness being over or bellow a break-even parameter. The authors indicated the controlling the maximum weight of a rule, specially ignoring it if its weight value exceeds that maximum configured, allows for the best challenge balancing. The idea is that the AI tactics can gradually improve with success (player failure) to a point closer to optimal, when then those best rules will be temporarily ignored with the AI using weaker tactics. The adjustment that result from subsequent losses (player success) will eventually put back in use the strongest rules, keeping the challenge level high but manageable.

Another work was made by Hunicke and Chapman (2004). They measured the amount of damage taken in a first-person shooter game in order to try to predict repeated inventory shortfalls: moments in which the players available resources (in this case, the health level) fail to meet the immediate demands. Using inventory theory they have modelled the probability distribution of damage and characterized shortfalls in the health level when the cumulative probability of damage exceeds the initial level (that is, when the demand in the "inventory" surpasses the "supply"). When shortfalls are detected in the player's near future, indicating that the she is "in need", the DDA system can take actions to easy the difficulty of challenges or to provide help in the form of additional resources (like more ammunition or medical packs).

An a particular famous research project was developed by (Shaker *et al.*, 2010). They used a freely available clone of the classic Nintendo's platform game Super Mario Bros and collected from many players three types of data: scenario features, including number of gaps and average gap width, among others, Gameplay metrics, including statistics on the number of jumps, directions changes and deaths, among others, and player experience self-reports, including impressions on three emotional dimensions, fun, challenge and frustration. They performed a feature selection procedure, and then used the number of gaps, average gap width, gap placement and number of player direction switches to construct a Multilayer Perceptron (MLP) to predict fun, frustration and challenge, obtaining respectively 64%, 85% and 70% accuracy. The level generation mechanism of the game was dynamically adapted by searching in the space of these controllable features the values that maximized the trained MLP output value.

Even though a general balance between challenges and skills is a requirement for Flow in a macro level, it is not always necessarily for all smaller experiences composing the Gameplay. In other words, performance does not exactly mirrors Flow (Chen, 2007a, p.12). For example, players enjoying suicidal stunts in racing games should not be considered as poorly-skilled, and hence as not having fun, just because of death counting (Chen, 2007a, p.12). It is postulated that physiological measurements such as heart rate and skin conductance may provide better predictions on whether an user is in the state of Flow (Bentley *et al.*, 2002, p.238).

One example is the research work made by Nacke and Lindley (2008). In their work, the authors customized levels of a commercially famous game to include intentional design characteristics of boredom, immersion and flow. The boredom level included linear walking, weak opponents, repeating visual and auditory patterns and limited options of actions, intending to elicit a condition in which skills are higher than challenge, but both bellow the optimal threshold. The immersion level included complex and exploratory environment, several opponents, varying auditory and visual patterns, many options for actions, and a narrative framing, intending to elicit a condition in which the Gameplay seems curious and appealing. And the flow level included combats with increasing difficulty, challenge mechanics that required concentration, and "cool down" spots that provided sparse amounts of health and ammo, intending to elicit the development of skills followed by the increasing in challenges.

Experiments were conducted with the different game levels while recording physiological responses taken from facial muscle activity and skin conductance, and the Game Experience Questionnaire (GEQ) (IJsselsteijn *et al.*, 2013) was used to collect the subjective scores of fun in terms of immersion, flow, challenge, and positive and negative affects for qualitative comparison. The comparison of self-report responses with the physiological recordings showed that there are statistically significant differences between the game levels, with players experiencing higher arousal (from skin conductance) and positive valence (from muscle activity) in the flow level. The boredom and immersion levels had very similar responses though, indicating that they are more difficult to differentiate using these measurements. Also, the authors conclude that the high-arousal and positive-valence observed in the flow level is a link of gradual challenges in competitive environment to the experience of positive emotions.

3.4 Assessing Immersion

Immersion has also been traditionally assessed with the use of questionnaires. Read *et al.* (2002), for example, used direct observation and self-reports to assess immersion in the interaction of children and computer systems. They first measured expectations and their fulfilment by using visual signs (the Funometer and the Smileyometer) to inquire the equivalent of "how much fun will this be?" and "how much fun was that?". The authors argue that the comparison between the states measured before and after the interaction is a good approximation of how appealing the activity is. They have also assessed engagement using direct observation of display patterns like smiles, frowns, laughs, concentration signs (fingers in mouth and tongue out), boredom signs (ear playing and fiddling), bouncing and positive or negative vocalization, and measured remembrance by applying post-experience questionnaires inquiring how willing the children are to perform tasks again. They argue that high returnability is due to the human natural remembrance of fun experiences, and help indicating and qualifying engagement.

An important observation is that the existing literature does not always clearly differ the assessment of attention and immersion, particularly because the later requires the former and engagement is sometimes seen as a stronger attentive state. Just as with the case of attention, it seems that automated immersion assessment can be more precisely achieved from the analysis of facial expressions and gaze. In the studies of immersion performed by Brown and Cairns (2004, p.1298) with video recording of gamers whilst playing, there were found little to no indications of whole-body behavioural associated with immersion in playing.

A very comprehensive research work on the automated assessment of immersion was performed by Jennett *et al.* (2008). Their work is founded on both concepts of Flow and immersion, considering their intersection regarding temporal dissociation, reduction of self-consciousness, sense of control and emotional involvement. The authors conducted three experiments using a proprietary eye tracking system, the well-known State-Trait Anxiety Inventory (STAI) (Spielberger *et al.*, 1970) and the Positive and Negative Affect Schedule - Expanded Form (PANAS-X) (Watson and Clark, 1994), and an immersion questionnaire they have created with thirty-three Likert scored questions related to emotional involvement (empathy to characters and story), transportation (suspension of disbelief), attention (distractability by other thoughts and awareness of external events) and control (easy of using controls and interacting with the world). The experiments investigated immersion from the perspectives of the task performance, the movements in the user's eyes and the pace of interaction, in which the questionnaires were used to qualitatively compare the objective measurements to the subjective feelings self-reported.

Their first finding is that the level of immersion seems to have a direct effect to the late performance of execution in external tasks, in the sense that the more immersed a person feels while playing a game the more time she takes to perform activities not related to the game. Or, in other words, the longer she takes to re-engage into the "real world". Another finding is that the participant's eye movements decreases significantly over time when she feels immersed, most probably as a result of focusing attention on visual components of the game. And the third finding is that the increasing pace of interaction elicits strong negative affects and high anxiety, although the activity is still perceived as immersive. The reason is argued to be due to challenges becoming "provoking": they induce more pressure to win because the player may have developed a certain level of proficiency that made her emotionally charged. So, winning would return the emotional state to equilibrium, as studied in the previous chapter. Also, self-paced interactions, in which the speed was controlled by the player herself, were not reported to be as immersive as fast-paced ones, but they were still reported to be enjoyable. This indicates that serene or repetitive tasks might be in Flow, requiring a lot of attention, but without being necessarily emotionally charged.

Bidwell and Fuchs (2011) also worked in the automated assessment of engagement of students in a learning environment. From video recordings of cameras fixed in the classroom and commercial face tracking software, they estimated the students' head direction, position and orientation in each second, and then manually labelled the patterns found regarding the gaze target (among whiteboard, teacher and other student) and eight discrete behaviours: engaged (actively involved, either individually or in group), passively attending (just involved in visual or auditory attention), transition (not paying attention due to other activities, like arranging materials, for instance), non-productive (not engaged, when looking around or rocking in the chair), inappropriate (actively involved in other tasks, like fiddling, interrupting, etc.), attention seeking (seeking for social attention from others, like making noise, talking aloud, etc.), resistive (physically or verbally resisting instructions) and aggressive (physically or verbally being attacking others). Those patterns were used to train a Hidden Markov Model (HMM) to automatically classify gaze target sequences into a probable behaviour. Training used 10% of the collected data and tests the other 90%. The results indicate that there is some correlation of sustained gaze and engagement, since the model is able to correct predict it 80% of the time. But, less attentional states are very poorly predicted, which suggests that other behavioural measurements are necessary.

3.5 Assessing Emotion

The emotional aspects of experience also have been traditionally assessed through self-reports, and there are indeed a vast number of tools available for that. Two most famous ones are the PANAS-X (Watson and Clark, 1994) questionnaire, already mentioned in this chapter, and the Self-Assessment Manikin (SAM) (Bradley and Lang, 1994). The PANAS-X is an evolution of the PANAS (Watson *et al.*, 1988) questionnaire (with the same name) which was composed of two sets of ten Likert scored questions used to measure the valence of both positive and negative affects. The extension included measurements on eleven specific affects: fear, sadness, guilt, hostility, shyness, fatigue, surprise, joviality, self-assurance, attentiveness and serenity. The SAM is a 1-5 pictorial measurement of all three dimensions of affect, valance, arousal and dominance, using figures of an humanoid drawn from frown to happy (valence), excited to relaxed (arousal) and smaller to big (dominance). The STAI (Spielberger *et al.*, 1970), also already mentioned, is intended to assess high arousal, being composed of twenty Likert scored questions to measure the state anxiety (related to the execution of an activity, or how an individual feels "at this moment") and the trait anxiety (related to a personal trait, or how an individual "feels in general"). There is also the Geneva emotion wheel (Scherer, 2005, p.724), named after its origin and shape, which graphically represents sixteen prototypic emotions in the four quadrants of the valence-arousal dimensions: pride, elation, happiness and satisfaction, in the positive-high quadrant; relief, hope, interest and surprise, in the positive-low quadrant; anxiety, sadness, boredom and shame/guilt, in the negative-low quadrant; and disgust, contempt, hostility and anger, in the negative-high quadrant. Each emotion has four circles of increasing size placed away from the wheel's centre, which can be easily marked by a respondent to graphically indicate her emotional state in terms of valence, arousal and discrete prototypic emotion.

A comprehensive automated measurement of emotions would require to assess all processing components, including the subjective mental representation of the feeling. This is something unlikely to be achieved, so there is no way to really know the emotional state of a person than to ask the individual to report on the true nature of her feelings (Scherer, 2005, p.712). However, it is possible to *infer* the emotional state from the bodily displays of physiological response patterns and expressive behaviour that accompany the experience of emotions (Scherer, 2005, p.709). However, this inference is always subject to noise due to the fact that the body is constantly under the effect of other stimuli sources independent from the product being evaluated. So, measures of bodily display should not be used alone, but in conjunction with other measures like self-reports (Nacke, 2013, p.613).

The physiological responses are produced by the Autonomic Nervous System (ANS). The ANS has a general-purpose nature, so its activity is not only related to the experience of emotions, but also to other bodily functions like digestion, homoeostasis, effort and attention, among others (Mauss and Robinson, 2009). That is why some of the physiological responses was also mentioned in the previous topics, used to help assessing attention and immersion. The indices of ANS activation are based on electrodermal (sweat gland) and cardiovascular (blood circulatory system). Electrodermal responses are quantified in terms of skin conductance level, and cardiovascular responses are quantified in terms of heart rate, blood pressure, total peripheral resistance, cardiac output, pre-ejection period and heart rate variability (Mauss and Robinson, 2009, p.211-212).

The relation of such individual measurements and discrete emotions, like fear, anger, joy, etc., has been showed to be highly inconsistent, so the ANS responses are best seen as indicators of broader values such as arousal, reflecting a level of affective state in that dimension rather than in its discrete emotional basis (Mauss and Robinson, 2009, p.212). The ANS responses also operate independently, and some of them – like blood pressure and heart rate – may also map to valence (Mauss and Robinson, 2009, p.213). The skin conductance, for example, increases linearly with the rated arousal of emotional stimuli, but it is independent of valence and of any specific prototypic emotion targeted (Mauss and Robinson, 2009, p.212). Nevertheless, the combination of multiple ANS measures is reported to yield better predictions of discrete emotional states (Mauss and Robinson, 2009, p.213).

Expressive behavioural responses were first studied by Darwin and described as the primary evolutionary function of communicating the emotional state to others. Modern theories also relate this displays to dispositions intended to prepare the body for action (Mauss and Robinson, 2009, p.217). The behavioural responses include patterns of facial and vocal expressions as well as whole-body movements (Scherer, 2005, p.709; Mauss and Robinson, 2009, p.217), and are commonly accessed through recording devices (such as audio and video recorders) or muscle activity (electromyography)

(Nacke, 2013, p.598).

Vocal expressions are measured in terms of sound amplitude (loudness) and pitch (fundamental frequency). Arousal and pitch are directly associated, and emotions with high levels of arousal, like fear, joy and anger, are linked to higher-pitched vocal samples (Mauss and Robinson, 2009, p.217). However, there is no relation of either amplitude or pitch to valence, and hence is very difficult to differentiate emotions that are closer in arousal but distant in valence, like anger from joy, just from the vocal expression (Mauss and Robinson, 2009, p.218).

Facial expressions are measured in terms of skeletal muscle activation (electromyography) (Nacke, 2013, p.597) or visual inspection of the movement of permanent features such as eyes, eyebrows and lips and the textural changes in transient features such as lines, wrinkles and furrows (Bettadapura, 2012, p.10). It is believed that facial expressions are the displays most closely tied to the organism behaviour (Mauss and Robinson, 2009, p.218), and that the expression of at least the six prototypic emotions are consistently recognized cross-culturally, even though their display is regulated according to specific cultural norms (Ekman and Friesen, 1971; Isbister, 2006, p.50; Matsumoto and Hwang, 2011, p.2; Bettadapura, 2012, p.8; Mauss and Robinson, 2009, p.218). The region of eyes and mouth, particularly the eyebrows and the mouth corners, are the facial elements most relevant to the expression of emotions (Bettadapura, 2012, p.10), being capable of reliably indicate the valance of a person's emotional state (Mauss and Robinson, 2009, p.219; Nacke, 2013, p.598) and provide good accuracy in the indication of the prototypic emotions (Bettadapura, 2012, p.10). Besides the fact that the facial displays are dependent upon cultural context, another important caveat of this type of assessment is that spontaneous expressions are usually subtle, making more difficult the measurement in nature (Bettadapura, 2012, p.9).

Whole-body expressions have not received too much attention as an indication of emotion, but are usually measured from electromyography, accelerometers or visual inspection of pose. Certain emotional states are supposed to have distant bodily behaviour signatures, and particularly pride and embarrassment are respectively linked to expansive and diminutive bodily postures (Mauss and Robinson, 2009, p.219). The scarcity of works in this area is probably due to the difficulty in recognizing whole-body expressions, since the configuration of the human body has much more degrees of freedom than the voice and the face (Schindler *et al.*, 2008, p.1239) and different emotions have many similar patterns of bodily motion (Isbister, 2006, p.173).

There are many research works that explored those possibilities to assess emotion. One of them is the work of Schindler *et al.* (2008). The authors explored the classification of prototypic emotions from bodily poses, using an approach inspired in the human visual cortex processing of images. Their training data is from a dataset of fifty actors posing whole-body standing-up expressions labelled to the six prototypic emotions plus a neutral pose. The features are extract in a sequence of steps. First a set of six frequencies and eight orientation Gabor filters (composing thirty two responses) is applied to the image. The choice of this filtering is due to it be an standard approximation of the primary visual cortex cells. The Gabor responses are then maximized (that is, the strongest response determines the output), so to make them more robust to noise. And finally, a Principal Component Analysis (PCA) is performed to extract the "eigen-image". The values of the eigen-image are then used to train a Support Vector Machine (SVM). Tests were performed using 10-fold cross-validation, and the classifier obtained an overall accuracy of 82%, with disgust and anger being the two most difficult emotions to predict (both around 70% accuracy). The authors noticed that the faces were not masked away from the images due to danger of removing important bodily expression information (faces are sometimes covered by hair or hand in expressions of sadness and fear), but they do not believe the faces contribute much to the accuracy of the classifier because in the dataset used the face areas are too small in comparison to the image sizes.

Yannakakis *et al.* (2007) also worked in assessing emotions, but with a specific purpose of evaluating fun in games. They used two test-bed games played in a matrix of tiles by pressing buttons with the feet and observing light displays. Those games are played by stepping over the tile buttons, and therefore are very physically demanding. The authors recorded heart rate signals while children played the games, and collected self-reports about their judgement of fun afterwards using the Funometer tool already mentioned in this chapter. The features collected from heart rate signals were average rate, variance of the signal, maximum and minimum rates, difference between maximum and minimum rates, correlation coefficient between recordings in a parametrized time window, and entropy of signal. They used an Artificial Neural Network to perform several steps of feature selection, ending up with average rate and correlation coefficient between recordings as the features that lead to best classification results, around 80% of correct predictions of fun from the measurement of heart rate signals.

Tan *et al.* (2012) investigated weather sufficient facial expressions are elicited when games are played, and if those expressions can be robustly captured to help assessing the emotional states of players. In tests with voluntaries playing two mainstream commercial games, they tracked face landmarks in video recordings using a deformable fitting algorithm. Then, they classified the probability of six prototypic emotions based on an Artificial Neural Network (ANN) trained from the local responses of a Gabor filter in each facial landmark. Unfortunately, the authors didn't provided further details on the source of training data. After the participants played the two games, they filled the GEW questionnaire. With statistical analysis of the data collected and the emotions detected, as well as visual inspections over the recorded facial expressions, the authors observed that a good variety of facial expressions other than neutral were exhibited with rich variance. Comical scenes accurately followed the elevated detection of happiness, and anger also increased over time according to self-reported frustration in figuring out puzzles. Participants were not instructed about how to behave in front of the camera, so failure in face tracking and emotion detection happened due to eventual head movement and hand occlusion during cinematic scenes, which were also self-reported as being the less engaging moments. The authors also observed that the presence of others considerably increased the occurrence of facial expressions in the players.

In a different domain, McDuff *et al.* (2013) worked on the prediction of likeness and desire to re-watch commercial ads, based on the amount of smiles detected in the viewers' faces. Using the Internet, they recorded videos of volunteers watching famous comical commercial ads and collected their answers to three questions regarding their liking, familiarity and rewatchability of the videos. Smiles were detected using facial features and a decision tree was used to produce the probabilities of smiles in each frame, resulting in a one-dimensional set of values indicating the smile intensity along the duration of each video. These tracks were filtered with a low pass filter to reduce noise, and then divided into twenty windows from which the peak values were used as the feature vector. They used a Linear Dynamic Conditional Random Fields classifier and tested the training using a leave-one-commercial-out approach, obtaining an accuracy of 80%. Most errors were due to the inaccuracy of the smile classifier, but some were reported to indicate people who showed no smile activity and yet reported liking the commercials.

Chapter 4

Data Collection

In the previous chapters it has been argued that fun is characterized by different aspects, among which immersion and emotions are most relevant. It was also argued that the human face, as an important medium of verbal and non-verbal human communication, displays many cues that are related to these aspects. So it has been hypothesized that estimated measurements of immersion and emotions may be reliably obtained from processing digital images with the faces of players, which in turn could help in assessing fun in playing digital games.

In order to verify this hypothesis, efforts have to be made to:

1. Capture images of volunteers playing a predefined set of games and the self-reported feedback on their feelings regarding the games played.
2. Extract from these images the features required to represent the modelled states.
3. Train models capable of classifying the designed states from those features.

The capture of volunteer images is required to create a database that can be used for the extraction of features and for training the model classifiers. The extraction of proper features is hence very important, and strongly dependent on the quality of face detection.

4.1 Planning and Set-up

The collection of data was performed in an experiment in which adult volunteers were asked to play a game while having their faces recorded by a common, off-the-shelf webcam. The volunteers were also asked to self-report their feelings regarding the game played, so the responses could be used to create the models and also to evaluate their predictions.

The experiment was carefully planned to preserve the dignity, identity, and health of the participants. A project proposal was submitted to be evaluated by a Committee of Ethics in Research (CEP, *Comitê de Ética em Pesquisa*) and by the National Health Council (CONEP, *Conselho Nacional de Saúde*) via *Plataforma Brasil*[1], the Brazilian unified and national database of researches involving human beings. The experiment was conducted only after the due approval of the submitted proposal.

4.1.1 Selection of Games

Since different games might induce different emotions (for instance, whilst fear isn't much experienced in a puzzle game, this feeling is certainly expected when playing a horror game), three games

[1] http://aplicacao.saude.gov.br/plataformabrasil

of different genres were employed in an attempt to maximize the variability of the data captured without encumbering the experiment set-up. So the option was to have a horror game, a puzzle game and a casual, more comical, game in order to possibly obtain a large set of emotions. Additionally, the following criteria were also important to the choice of games to use:

- The games should be free or have a free playable version or demo.
- The games should be popular, but not mainstream.
- The games should be acknowledged as fun by a community.
- The games should be as independent as possible of language.
- The games should not discriminate race, sex, colour, sexual orientation or religion.

The first item is needed for the game to be used in this work, but also intends to allow the easy reproduction of the results described in this thesis. The second item reduces the chances that the players might have played the games before, thus avoiding bias related to pre-built knowledge, skills and judgement. The third item intends to guarantee that the games are potentially fun for broader audiences, hence avoiding low-response bias due to poorly designed games. The fourth item also avoids bias due to difficulties not directly related to the game design. And the fifth item helps to assure that no harm or discomfort would be caused to the players.

After a comprehensive search on online markets such as Steam[2] and independent game communities such as IndieDB[3], the following games were chosen: Melter Man, Cogs and Kraven Manor. They are briefly described in the following.

Melter Man

Melter Man (figure 4.1a) is a 2015 independent 2D platform game in which the player has to clean a toy factory from toys that went berserk, by melting, vacuuming and shooting the material the toys are made of. On the IndieDB website[4] the game is ranked 2,289 of 39,461 (top 5.80% best games in January 2017) and tagged with the labels “adventure”, “comedy” and “platformer”. The game is only available in English, but its interface is very little dependent on text: it has small graffitis on the walls with single-word instructions and drawings of the relevant keys/buttons in the keyboard/joystick. Also, the player does not need to know or comprehend the “toy factory taken by berserker toys” story in order to be able to play it.

A free early access version, fully functional but with reduced number of levels, was downloaded from the game website[5]. The number of available levels was enough for the programmed duration of the experiment.

Cogs

Cogs (figure 4.1b) is a 2009 independent puzzle game in which the player has to build a great variety of machines from sliding tiles in a 3D Steampunk ambience. On the IndieDB website[6] the game is ranked 1,474 of 39,461 (top 3.73% best games in January 2017) and tagged with the labels “antiquity”, “family” and “puzzle”. The game interface is deeply dependent of text, due to the

[2] http://store.steampowered.com/
[3] http://www.indiedb.com/
[4] http://www.indiedb.com/games/melter-man
[5] http://www.melterman.com/downloads/
[6] http://www.indiedb.com/games/cogs

description of goals, instructions for each different machine and menu options. However, the game is available in English and Portuguese (among other languages).

A free demonstration version, fully functional but with reduced number of levels, was downloaded from the game website[7]. The number of available levels was enough for the programmed duration of the experiment.

Kraven Manor

Kraven Manor (figure 4.1c) is a 2014 independent 3D horror game in which the player has to explore an old, dark and inhabited manor where its owner and fifty workers disappeared mysteriously. The game was awarded with two prizes (Best Gameplay and Best Visual Quality) at the Intel University Games Showcase. On the IndieDB website[8] the game is raked 613 of 39,461 (top 1.55% best games in January 2017) and tagged with the labels "adventure" and "horror". The game interface is deeply dependent on text due to the need to follow the game narrative and understand the instructions. However, the game is available in English and Portuguese (among other languages).

A free demonstration version, fully functional but with reduced number of rooms, was downloaded from the game's entry in the IndieDB website[9]. The number of available rooms was enough for the programmed duration of the experiment.

(a) *Melter Man* **(b)** *Cogs* **(c)** *Kraven Manor*

Figure 4.1: *Screenshots taken of the games used in the data collection experiment – the copyright of the images belong to the respective game creators*

4.1.2 Selection of Participants

The experiment was dimensioned for 60 participants, considering that it would be possible to obtain 20 gameplay samples for each game. Adult volunteers were randomly selected in two locations in city of São Paulo: the campus of the University of São Paulo and a private music school. In both locations students and employees where invited to participate without the provision of any external reward. They were invited by e-mail messages broadcast to student and staff groups and by direct and informal invitations made by the researcher on coffee shops, restaurants and hallways. The volunteers were able to book their participation as desired according to an available hourly schedule, or to play as they arrived in the room if there was no ongoing session.

The only requirements to participate were to be older than eighteen years old, to have basic experience in the use of personal Desktop computers, not to have a history of photosensitive epilepsy or repetitive strain injury, and not to be feeling sleepy or tired at the moment of the session. The limitations imposed were means of safeguarding the health and dignity of the participants, following the applicable ethical guidance. The limitation to a minimum age of eighteen years old was intended not only to make the experiment easier from the ethical perspective (since adults are capable of responding for themselves regarding their participation), but to avoid difficulties in the face tracking.

[7] http://www.lazy8studios.com/free_downloads
[8] http://www.indiedb.com/games/kraven-manor
[9] http://www.indiedb.com/games/kraven-manor/downloads/kraven-manor-demo

Even though the facial muscles are anatomically mature and functional at the moment of birth, there are relevant structural differences between the infant's face and the adult's face, such as proportions and dimensions of skeletal parts, thinner eyebrows, diminutive mandible, underdeveloped chin and absence of proeminent supraorbital ridges (Oster and Ekman, 1978, p.246–247).

The order in which the games were played was randomly generated at the beginning of the experiment and never changed. The games were then assigned to the participants with that order in a circular fashion, as they arrived: the first participant played game A, the second played game B, the third played game C, the fourth played game A, and so on. Since the order of participants was not controlled in any way, this approach permitted to have random assignments of games while capturing a similar number of samples for each game.

The anonymity of the participants was guaranteed by the experiment set-up and insured by a consent form, prepared with the help of the Ethics Committee and signed by both the participant and the researcher in charge of the experiment. The original document was prepared in Portuguese, but there was an English version – presented in appendix A (Free and Clarified Consent) – available for participants that might not speak the language.

4.1.3 Questionnaires

The experiment was planned so each participant would have to play a selected game while her face was captured in video. Next she would have to review the video of her gameplay, which was also captured, as she answered 3 questions of a short questionnaire presented at specific moments of the playback. Finally she would have to answer the 33 questions of the Game Experience Questionnaire (GEQ) (IJsselsteijn *et al.*, 2013) and other 5 questions of an ethnographic questionnaire.

In order to avoid tiring the participants – what could reduce confidence in the provided answers – and still record a large number of responses, the capture of the gameplay was limited to 10 minutes, with another 10 minutes for the review and questionnaires. The review of the gameplay was limited only to the last 5 minutes of the video, with the short questionnaire being displayed at discrete timestamps spaced by 30 seconds (red ticks, marked in the video playback interface used). The review phase presented only the video and audio of the game session played, that is, without showing the player's face, in order to prevent discomfort and to keep the participant's attention in the game experience.

The answers for most of the questions in the questionnaires use a Likert scale varying between "not at all" ("*de jeito nenhum*"), "slightly" ("*levemente*"), "moderately" ("*moderadamente*"), "fairly" ("*bastante*") and "extremely" ("*extremamente*"), valued from 0 to 4 (in parenthesis there are the Portuguese translations employed). The questions were constructed according to the form used by GEQ: they lead the participant to think on how she was feeling during the gameplay by using an affirmative sentence instead of an interrogative one. All questions in all questionnaires were always presented initially with no answer marked or given.

Gameplay Review Questionnaire

The 3 questions of the short questionnaire used in the gameplay review inquiry the levels of frustration, immersion and fun the participant was feeling at a given moment of the gameplay. The items of this short questionnaire will not be used for scoring, but instead will represent a set of 5 discrete states employed for classification of the level of frustration, immersion and fun of the participants.

The word "involvement" was preferred to to word "immersion" because it is considered easier to understand by the general people. The 3 questions, with their Portuguese translations, are the following:

1. I was feeling frustrated (*Eu me sentia frustrado(a)*)

2. I was feeling involved (*Eu me sentia envolvido(a)*)
3. I was having fun (*Eu estava me divertindo*)

That short questionnaire was always presented with a title and subtitle, in order to make clear to the participant that she should consider how she felt at the moment the video is paused:

Title: How were you feeling at that time?

(*Como você estava se sentindo neste momento?*)

Subtitle: Please indicate how you were feeling while playing the game at the time the video is paused, for each of the following items.

(*Por favor indique como você estava se sentindo enquanto jogava no momento em que o vídeo está parado, para cada um dos itens seguintes.*)

Game Experience Questionnaire

GEQ was asked only at the end of each experiment session, with the intention to capture additional information that could be used to validate the review answers and help pointing out outliers. The core module, presented in annex A (The Game Experience Questionnaire) with the used Portuguese translations, was the only one employed because it permits to obtain an overview of the experience in the whole session regarding different components of fun. These components are scored as the average value of its items, as follows:

Competence: Items 2, 10, 15, 17, and 21.

Sensory and Imaginative Immersion: Items 3, 12, 18, 19, 27, and 30.

Flow: Items 5, 13, 25, 28, and 31.

Tension/Annoyance: Items 22, 24, and 29.

Challenge: Items 11, 23, 26, 32, and 33.

Negative Affect: Items 7, 8, 9, and 16.

Positive Affect: Items 1, 4, 6, 14, and 20.

Competence and Challenge are directly related to Flow, and might not be as useful as the others for comparison, although they might help explain responses regarding the self-reported immersion levels. Sensory and Imaginative Immersion are the most directed items, however. Tension/Annoyance is related to the review question regarding frustration, and the Negative and Positive Affects will be helpful in evaluating the detected prototypic emotions.

An important issue regarding the use of GEQ is that there is no Portuguese translation that has been properly validated as the original English version has. GEQ authors mention that, by their experience, translations sometimes result in suboptimal scoring patterns no matter how carefully they are performed. It was not in the scope of this work to validate a translation of GEQ to Portuguese, but great care was taken in performing the translations (found in the annex aforementioned) so the intended meaning in English could be captured as well as possible in Portuguese.

Ethnographic Questionnaire

The ethnographic questionnaire was asked last in the experiment session, with the intention of capturing information regarding the playing habits of the participants and if they had already played the games assigned to them. That information is helpful in settling doubts regarding the validity of the results and also in verifying the representativeness of the collected data. The questions, with their Portuguese translations, are the following:

- How old are you? (*Quantos anos você tem?*)
- What is your sex? (*Qual é o seu sexo?)*)
  - Male (*Masculino*)
  - Female (*Feminino*)
- Do you usually play digital games? (*Você costuma jogar jogos digitais?*)
  - Yes (*Sim*)
  - No (*Não*)
- How many hours per week do you spend playing digital games? (*Quantas horas por semana você costuma jogar jogos digitais?*)
  - 0-2 hours (*0-2 horas*)
  - 2-5 hours (*2-5 horas*)
  - 5-10 hours (*5-10 horas*)
  - 10+ hours (*10+ horas*)
- Have you played [game name] (the game you just played) before? (*Você já tinha jogado [nome do jogo] (o jogo que acabou de jogar) antes?*)
  - Yes (*Sim*)
  - No (*Não*)

4.1.4 Conditions and Equipment

The data collection took place in private rooms booked for a whole week just for its execution. The room used in the music school was artificially lit and well aired (with air conditioning and a ceiling fan) small class room. The window was kept closed to not disturb the participants or cause lighting variations. The room used in the University was an office room, with both artificial and natural light coming from big lateral windows that could not be closed. The window had vertical blinds that were kept closed but still did not block much of the incoming natural light. That was the only room available, so the equipment was positioned sideways in order to not cause reflections on the monitor and the camera. The windows were kept opened to allow air circulation, since the room did not have air conditioning.

Upon arriving at the room, a participant was briefly instructed on the experiment purpose and proceedings and the consent form was given for her to read and sign. Then the participant was asked to sit in front of the computer as comfortably as they would find fit, the camera was adjusted to capture her whole face (depending on the individual's height) in the centre of the image, and the researcher would leave the room leaving the participant alone to play.

In order to conduct the experiment without requiring the presence of the researcher, a guiding software tool was created. This tool helped to minimize the observer influence to the knowledge of being filmed, but also helped to guarantee the anonymity of players, to automatically select the games to play, and to collect the data without human intervention. The tool was also in charge of accounting the 10 minutes of programmed gameplay. When that time expired, the game was forcefully interrupted and the participant was instructed on what would come next. The participants were all previously instructed of this forceful interruption at the beginning of the experiment, in an attempt to minimize the effect in the game experience caused by a sudden interruption. Alternatives to terminate the game in an more elegantly way were sought but could not be achieved in a timely manner.

The software was developed in C++ using the Qt library[10] for the graphical interface and the Open Broadcaster Software Studio (OBS)[11] for the capture of videos from both the gameplay and the player's face. It was executed in full-screen mode, with all keys that would otherwise allow accessing the operating system disabled. For instance, "Ctrl+Tab", which normally allows to alternate tasks, "Alt+F4", which normally allows for closing the current application, and "Win+M", which normally minimizes all running applications, where all disabled via configurations performed in the guiding software or directly in the operating system.

When the selected game was automatically initiated, it was executed over the guiding tool in full-screen mode. Two of the games have configurations that allow exiting from that mode. But the guiding tool was always kept in full-screen mode, so even if a participant would succeed in alternating the tasks she shouldn't be able to access any other resources of the operating system. The guiding software was also executed with elevated privileges under an user session of lower privileges that didn't allow access to the data folder. These measures prevented or, at least, made considerably more difficult for an eventual ill-intentioned participant to have access to the data of other participants or to compromise the equipment.

After the experiment was initiated, it was observed that the videos captured from the gameplay did not always have exact 10 minutes, sometimes presenting a small variation of 1 or 2 seconds in length and always a 1 second black screen at the beginning. This was due to timing variations in the start-up of the OBS integration, which could not be easily improved. The variations didn't prevent the participants to perform the reviews, but caused the guiding software to fail when saving data. Unfortunately the data of a few participants got lost, but the software was quickly corrected by displacing the short questionnaires by a margin of 5 seconds from the actual end of each video. This caused the measurements to no longer have the exact same moments in time for each participant, but prevented the problem from happening again.

The guidance tool, OBS and the games were all executed in the same machine: a desktop computer with an Intel Core i7 2.80 GHz processor and 8 GB of RAM memory, running a Windows 10 64 bits operating system. The performance was not hampered by all that software running, but still the graphical quality of the games was reduced from "Excellent" to "Good". A 3.1 audio system was attached to provide good sound effects and music, and a Logitech HD Webcam C270 was employed for the recordings, using its standard 1280 x 720 (HD) resolution – the camera does not have manual zooming or auto-focus function, so the focal length does not change. The camera was placed over a SyncMaster BX2350 monitor with 23 inches, configured with a 1920 x 1080 (Full HD) resolution, enough to display the games in the good graphical quality.

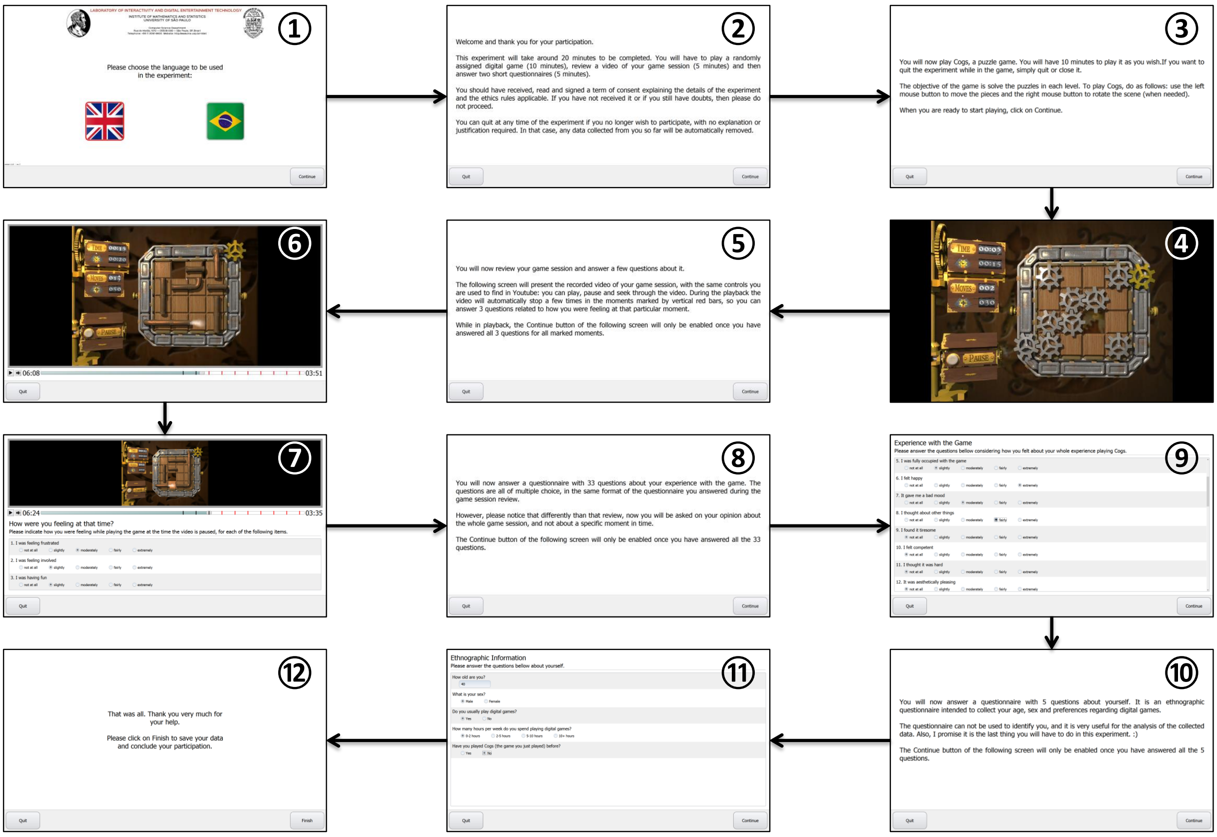
In pre-tests it was observed that the camera frequently moved or vibrated slightly due to the motion of the participants as they interacted with the equipment and particularly due to the camera design: it does not have a firm grip over the monitor. A reposition of the camera to the wall behind the monitor was briefly considered but quickly discarded, because it made harder to capture the participants faces frontally or was obstructed by the monitor itself. Unfortunately, this fragility of the camera set-up caused small problems with the videos collected from two of the participants, whose faces weren't well framed all the time (having the bottom of the face partially cut).

[10] http://qt-project.org/

[11] https://obsproject.com

4.2 Execution

The sequence of a session in the experiment of data collection, as guided by the software tool created, is illustrated in figure 4.2.



LABORATORY OF INTERACTIVITY AND DIGITAL ENTERTAINMENT TECHNOLOGY
INSTITUTE OF MATHEMATICS AND STATISTICS
UNIVERSITY OF SÃO PAULO

Please choose the language to be used in the experiment:

Welcome and thank you for your participation.

This experiment will take around 20 minutes to be completed. You will have to play a randomly assigned digital game (10 minutes), review a video of your game session (5 minutes) and then answer two short questionnaires (5 minutes).

You should have received, read and signed a term of consent explaining the details of the experiment and the ethics rules applicable. If you have not received it or if you still have doubts, then please do not proceed.

You can quit at any time of the experiment if you no longer wish to participate, with no explanation or justification required. In that case, any data collected from you so far will be automatically removed.

You will now play Cogs, a puzzle game. You will have 10 minutes to play it as you wish.If you want to quit the experiment while in the game, simply quit or close it.

The objective of the game is solve the puzzles in each level. To play Cogs, do as follows: use the left mouse button to move the pieces and the right mouse button to rotate the scene (when needed).

When you are ready to start playing, click on Continue.

You will now review your game session and answer a few questions about it.

The following screen will present the recorded video of your game session, with the same controls you are used to find in Youtube: you can play, pause and seek through the video. During the playback the video will automatically stop a few times in the moments marked by vertical red bars, so you can answer 3 questions related to how you were feeling at that particular moment.

While in playback, the Continue button of the following screen will only be enabled once you have answered all 3 questions for all marked moments.

How were you feeling at that time?

Please indicate how you were feeling while playing the game at the time the video is paused, for each of the following items.

1. I was feeling frustrated
○ not at all ○ slightly ● moderately ○ fairly ○ extremely

2. I was feeling involved
○ not at all ● slightly ○ moderately ○ fairly ○ extremely

3. I was having fun
○ not at all ● slightly ○ moderately ○ fairly ○ extremely

You will now answer a questionnaire with 33 questions about your experience with the game. The questions are all of multiple choice, in the same format of the questionnaire you answered during the game session review.

However, please notice that differently than that review, now you will be asked on your opinion about the whole game session, and not about a specific moment in time.

The Continue button of the following screen will only be enabled once you have answered all the 33 questions.

Experience with the Game

Please answer the questions bellow considering how you felt about your whole experience playing Cogs.

5. I was fully occupied with the game
○ not at all ● slightly ○ moderately ○ fairly ○ extremely

6. I felt happy
○ not at all ○ slightly ○ moderately ○ fairly ● extremely

7. It gave me a bad mood
○ not at all ○ slightly ● moderately ○ fairly ○ extremely

8. I thought about other things
○ not at all ○ slightly ○ moderately ● fairly ○ extremely

9. I found it tiresome
● not at all ○ slightly ○ moderately ○ fairly ○ extremely

10. I felt competent
● not at all ○ slightly ○ moderately ○ fairly ○ extremely

11. I thought it was hard
● not at all ○ slightly ○ moderately ○ fairly ○ extremely

12. It was aesthetically pleasing
● not at all ○ slightly ○ moderately ○ fairly ○ extremely

You will now answer a questionnaire with 5 questions about yourself. It is an ethnographic questionnaire intended to collect your age, sex and preferences regarding digital games.

The questionnaire can not be used to identify you, and it is very useful for the analysis of the collected data. Also, I promise it is the last thing you will have to do in this experiment. :)

The Continue button of the following screen will only be enabled once you have answered all the 5 questions.

Ethnographic Information

Please answer the questions bellow about yourself.

How old are you?
40

What is your sex?
● Male ○ Female

Do you usually play digital games?
● Yes ○ No

How many hours per week do you spend playing digital games?
● 0-2 hours ○ 2-5 hours ○ 5-10 hours ○ 10+ hours

Have you played Cogs (the game you just played) before?
○ Yes ● No

That was all. Thank you very much for your help.

Please click on Finish to save your data and conclude your participation.

Figure 4.2: *Screenshots of the guiding software tool during an example session of data capture*

When a participant entered the room, after having had already read and signed the consent form, the computer was already displaying the initial window (1) where she could select the preferred language. Via a secret combination of keys, that only worked in that first window, the camera was tested and manually adjusted to capture the participant's whole face. The researcher would then leave the room.

Upon clicking the button "Continue", a welcome message window (2) explained the experiment once again was presented to the participant. From then on a button named "Quit" was made available, so the participant could quit from the experiment if she wished so at any time. In that case, the software tool conducting the experiment would immediately and permanently eliminate all data recorded so far and return to the initial window. If the participant proceeded, the tool then selected (3) the game to be played in the circular fashion previously explained and automatically started it (4). After the gameplay time is elapsed, the guiding tool would then close the game and present a window explaining the next steps (5).

The reviewing of the gameplay video was performed with an interface very similar to what the participants would find in popular video services such as Youtube (6). The playback of the whole gameplay (i.e. the 10 minutes recorded) was presented from the beginning, and the participant was only interrupted with the short questionnaire at the first ticked position (in red), which occurs only after the first 5 minutes (7). The software tool automatically stopped the playback at those positions and continued when the three answers were provided, but the participant was able to move the progress bar backwards or forwards and change any previously given answer. She was only able to proceed to the next window after filling up all the 3 questions for all the labelled ticks (in red).

When the review is concluded, the participant is informed (8) of the questionnaire with 33 questions on her general experience with the game (GEQ) (9). Finally the participant is informed (10) of the last questionnaire and reminded that she can not be identified by the data provided. The ethnographic data is then captured (11) and the session ends with a thanking message (12).

All data captured is saved to the disk only after the participant clicks the "Finish" button on this last window. The answers to the questionnaires are saved to text files in the Comma-Separated (CSV) format, and the videos – recorded with a frame rate of 30 frames per second – are saved in the MPEG-4 (Moving Picture Experts Group compression format in version 4) compression format.

4.2.1 Overview of the Collected Data

The original plan was to collect data from 60 participants, but only 41 people volunteered. As it has been already mentioned, a problem with the guiding tool caused some data to be lost. So from the 41 participants in the experiment, the data of 6 of then was not saved (subjects 3, 5, 10, 11, 12 and 13) resulting in a number of 35 samples collected.

The number of participants was almost the same for either sex (17 males and 18 females). Their age varied from 18 to 54 years old, with an average of 30 years old (23 for males and 32 for females). There were nearly the same number of sessions for each game, with Cogs being played 12 times, Melter Man being played 12 times, and Kraven Manor being played 11 times. Cogs was more played by male participants and Kraven Manor was more played by female participants. Male participants have reported to play more and for longer periods, and no participant has reported to have played the assigned game before. This information is presented graphically in figure 4.3.

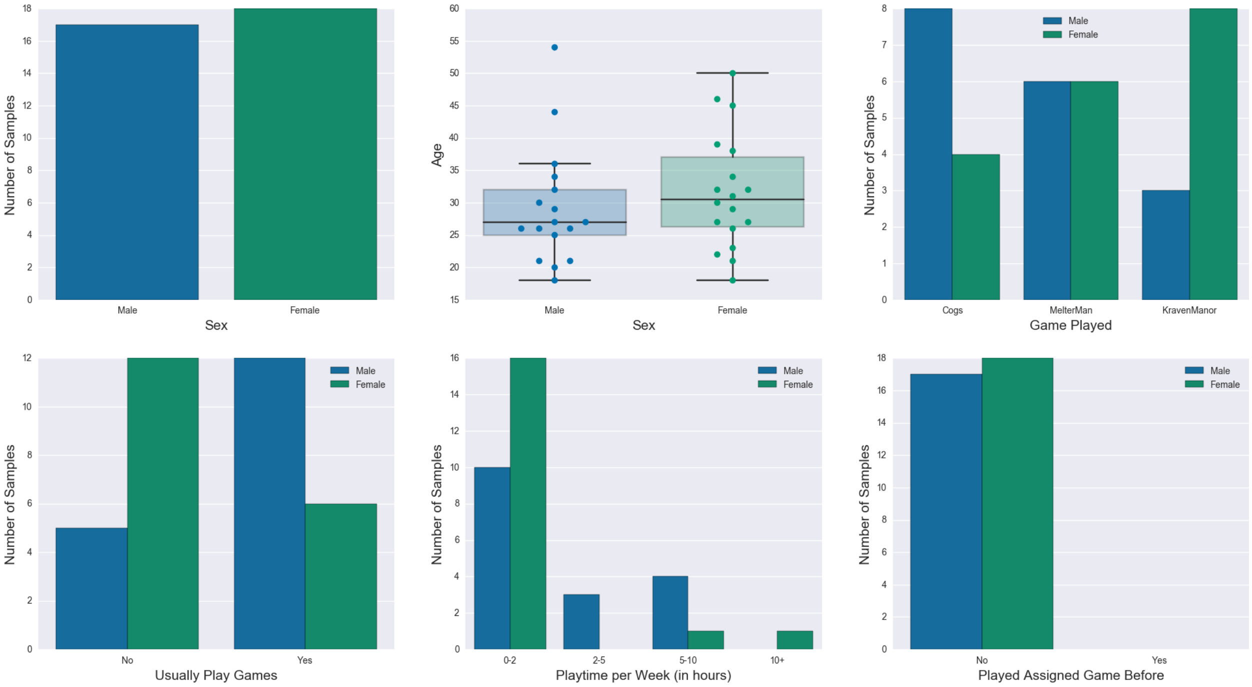


Figure 4.3: *Overview of the data collected in the experiment*

Additional observations, obtained from visual inspection of the face videos, are that almost half of the participants wore glasses (17 with glasses and 18 without glasses) and almost half of the male participants had facial hair (9 with facial hair and 8 without facial hair). This information is relevant for the analysis of results of the face tracker. Also, most of the participants did not know the researcher in charge of the experiment, but about 30% of them were coleagues of the researcher in the university and music school or students of his advisor that freely opted to participate in reply to the e-mail invitation.

All participants were informed on the purpose of the research and that their faces were being recorded during the experiment. Besides authorizing the computational use of their images, all participants were also requested permission of reproduction of anonymous images of their faces in this thesis, in papers and in presentations (via the same consent form reproduced in appendix A). The participants that granted this specific authorization are the subjects of number 1, 2, 4, 6, 7, 8, 14, 16, 20, 21, 22, 23, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 36, 37, 38, 40, 41. The subjects of number 9, 15, 17, 18, 19, 24, 35 and 39 did not grant authorization to reproduce their face images in any media.

Answers to the Questionnaires

The scoring of the answers of each participant in the GEQ questionnaire are presented in figure 4.4. They were calculated as described earlier in section 4.1.3 (Game Experience Questionnaire), and are represented in the same scale as the answers in the questionnaire: 0 ("not at all"), 1 ("slightly"), 2 ("moderately"), 3 ("fairly") and 4 ("extremely"). The neutral (i.e. central) value is 2. An overview of the scores is also presented in figure 4.5, built from box plots displaying the median and the Interquartile Range (IQR) (in the box), the minimum and maximum values (in the whiskers) and eventual outliers (the points above or bellow the whiskers). In both graphs, the data is coloured to separate each game played.

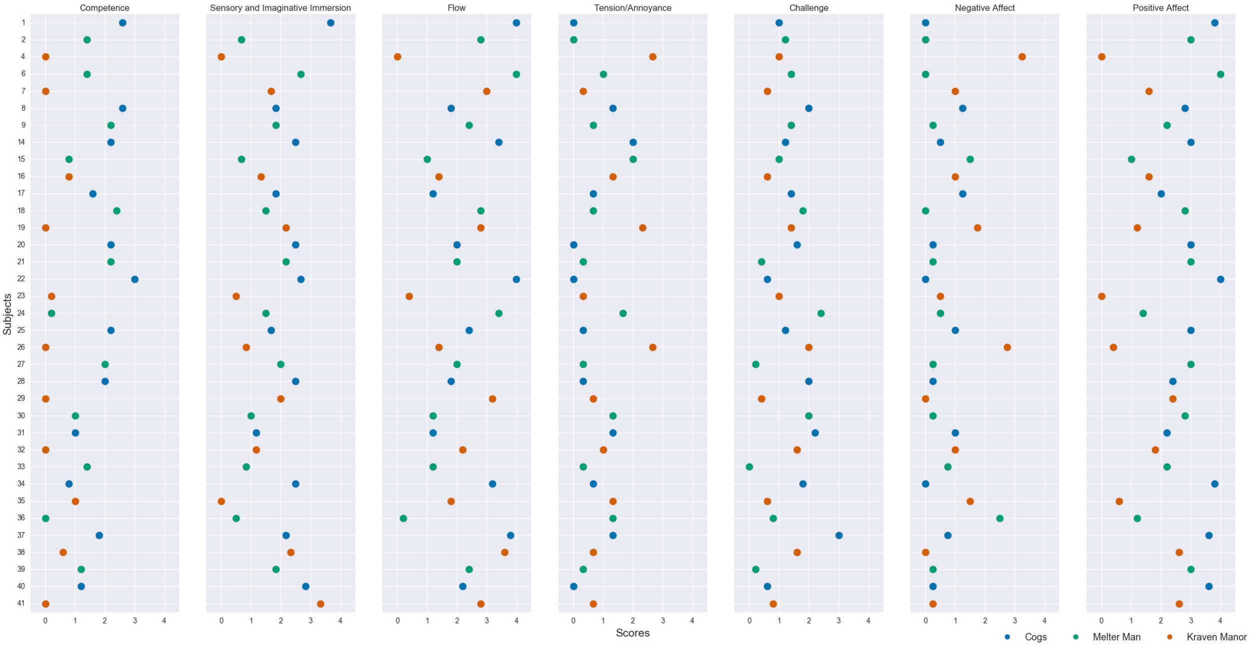


Figure 4.4: *Answers for the Game Experience Questionnaire*

Here it is a brief discussion on the scores obtained in each category foreseen by GEQ:

- **Competence**. The participants had much more difficulties playing Kraven Manor. This was expected since that game is the most complex one, with more options of action and a rich background storyline. For the other games, the median score of Competence was near the neutral value 2 , but the minimum-maximum range varied more towards low scores. It was between 0 and 1 for MelterMan and between 1 and 3 for Cogs. This indicates that Cogs was slightly easier to master than Melter Man, which was also expected since Cogs is the game with fewer options for action.

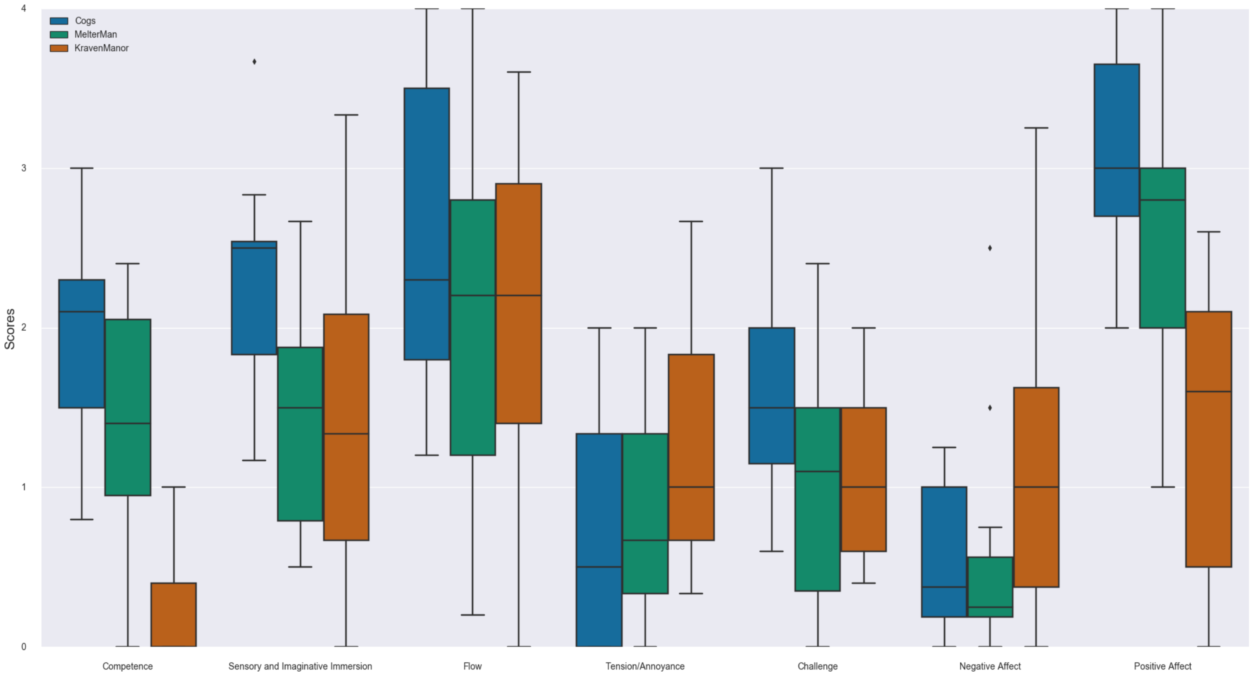


Figure 4.5: *Overview of the answers for the Game Experience Questionnaire*

- **Immersion**. Cogs had a much higher median of Immersion than other games, even though its IQR is still concentrated near the neutral value. The median score of Melter Man and Kraven Manor are very similar, in between 1 and 2. Kraven Manor had the larger minimum-maximum range, from 0 to 3. This larger variation might be explained by the fact that Kravan Manor is the only game with a storyline and a 3D detailed environment.

- **Flow**. All games had mostly the same median scoring, near the neutral value. Cogs had a slightly higher scoring in general, what might be explained by the higher need of attention in the solution of the puzzles (Cogs also obtained the higher scores in Competence and Challenge). Nonetheless, the other games achieved larger minimum-maximum variations, from 0 to 4. This might be explained by preferences. A puzzle tends to be a more neutral and generic type of game, which potentially appeals to many players; whilst the other two games have very specific genres (horror and comedy/action-platform), which may not be easily liked by many players as a puzzle is.

- **Tension/Annoyance**. All games had lower scores of tension/annoyance, with median bellow 1 and minimum-maximum variations from 0 to 2. Cogs and Melter Man have nearly the same scoring in this category while Kraven Manor had the highest scores reaching up to a maximum of 3. Since it is a horror game, higher tension is expected. The scores of Cogs were slightly lower than the scores of Melter Man. A possible reason is that Cogs is a slow paced game in comparison with the other games. It has a timer and a bell that rings as the player takes longer to solve the puzzles, but the game does not stop its gameplay based on that (it is only used to compute the points obtained). Thus, most probably the players didn't feel much the time pressure and have played freely.

- **Challenge**. Cogs was the game considered the most challenging game, with scores varying from 1 to 3. Again this is expected, since it is the only puzzle game. Also, it was the game with highest scores in Competence and Flow. Nonetheless, the median and the IQR in Challenge were concentrated near the neutral value 2. The other games were perceived as less challenging. Particularly Melter Man had the lowest scores, with most of the answers evaluating it from 0 to 1. The scores of Kraven Manor orbitate towards 1, indicating that it was felt as more challenging than Melter Man, yet not enough. The reason for that might be due only to the game interface (it had many options for action), and because the experiment time may have been too short for this particular game.

- **Negative Affect**. The experience of negative affects was generally very low with all the games. Kraven Manor produced the strongest responses in this category, with a median score of 1, an IQR reaching 2 on the third quartile and a maximum of 3 (that is, 50% of the scores are near 1). It is important to notice that this GEQ category is not about the experience of emotions of negative valence (like fear or anger), but to negative attitudes towards the game (i.e. mainly indicating if the participants didn't like the experience: the questions used for this score are "it gave me a bad mood", "I thought about other things", "it felt tiresome" and "I felt bored"). Therefore, this score indicates that the participants tended to dislike Kraven Manor slightly more than the other games. A possible explanation is that the gameplay session in the experiment (10 minutes) might have been too short for this game, not providing enough opportunities for the participant to explore the game and its background story.
- **Positive Affect**. In opposition to the previous category, the experience of positive affect was higher with Cogs and Melter Man, and lower with Kraven Manor (which had all scores essentially bellow 2). Cogs was the game that clearly elicited more positive affects, reaching up to 4 but with a minimum score of 2. Melter Man had larger minimum-maximum variations, from 1 to 4.

The answers provided from each participant during the review of the gameplay are presented in figure 4.6. The questionnaire was asked 10 times, as indicated in the x axis of the graphs. The answers are indicated in the y axis. Most of the participants had variations in their answers, with just a small number of them providing more constant responses (like subjects 1, 4 and 6, for instance). As it would be intuitively expected, frustration seems vary in opposition of fun, while fun varies in accordance with immersion. An opposition between frustration and fun is very noticeable in the graphs of subjects 34, 36 and 39, while an accordance between immersion and fun are very noticeable in the graphs of subjects 1, 4, 6, 7, 34 and 40, among others with one or two divergences only). The term "involvement" was used in place of "immersion" because it was considered to be easier to understand by layman people. It is improbable that involvement would have been confused with fun during the review, since the two questions were always presented together, but it still may not be understood correctly by the subjects.

The review data is the one of real interest to this work, required for the training and evaluation of the classifiers. But the participants may have a positive bias if trying not to displease the researcher conducting the experiment, or get tired and thus not answer all the 10 gameplay review questionnaires with the same care. Since GEQ is a well-validated questionnaire and largely used by the literature, it was also asked in order to collect comparison data of the whole experience which allowed to check the correctness of the review answers.

Figure 4.7 presents the comparisons between the relevant scores in GEQ and the mean (with standard deviation) scores of all the gameplay reviews, for each subject. The review item for frustration is compared with the GEQ score for tension/annoyance, since this is the category directly related to the review item (GEQ's question 29 is indeed the very same question). The review item for immersion is compared with the same GEQ, and the review item of fun is compared to the GEQ score of positive affect (since this is about the positive attitudes towards the experience).

The review responses seem to fit well the respective GEQ scores for most of the subjects. The larger differences in frustration are from subjects 22 and 33. According to the GEQ score, the former has self-reported to not feel frustrated at all (level 0), while the mean score of the review indicates a frustration at level 1. The latter self-reported in GEQ to fell a little frustrated, but this didn't appear in the mean review score. All other subjects had variations, but still inside the deviation in comparison to the GEQ responses.

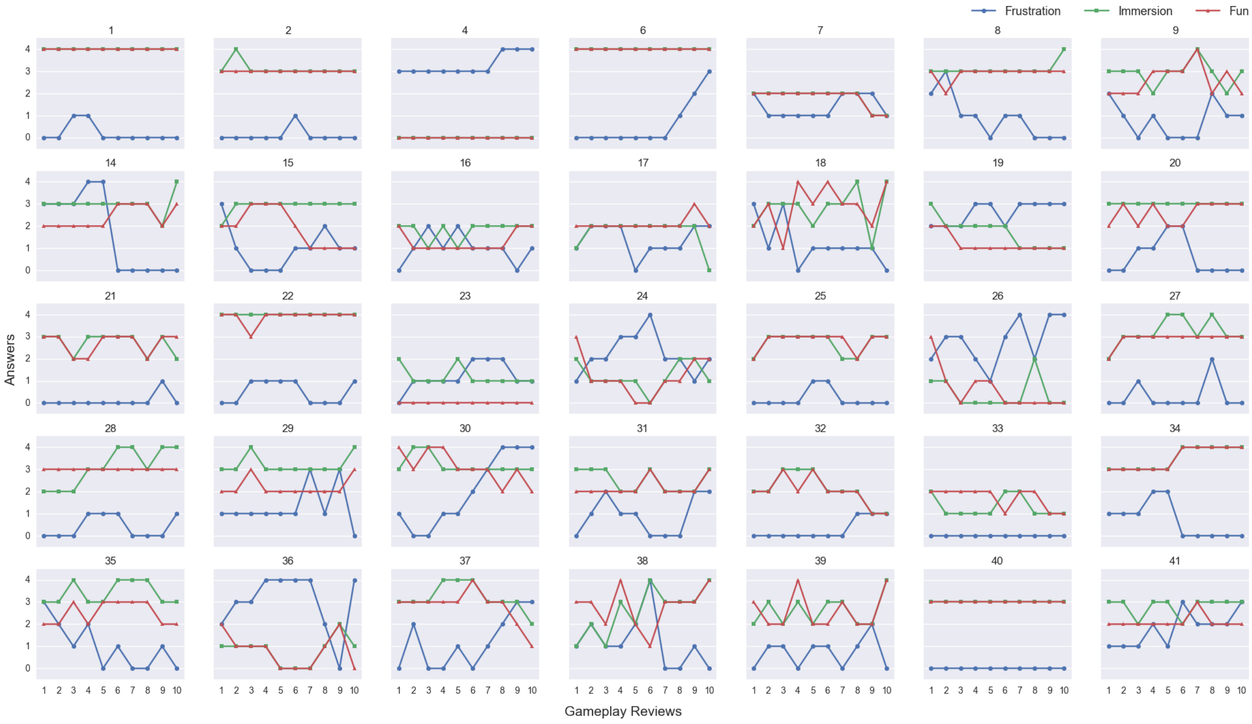


Figure 4.6: *Answers for the gameplay review questionnaire*

The larger differences in immersion are from subjects 2, 15 and specially 35. Subjects 2 and 15 played Melter Man, while subject 35 played Kraven Manor. The GEQ questions regarding immersion use terms like "game story", "aesthetically pleasing", "imaginative" and "rich experience", while the review only asked for "involvement". Melter Man does not have a story, it is the only 2D game (with a graphical quality subjectively worse than the others). The review scores of immersion were always higher the the GEQ score, what might indicate that the immersion considered by the two subjects in the review was more related to the focus of attention in the task than to the sensory and imaginative immersion asked in GEQ. Subject 35 also reported to be much involved in the review but the GEQ score was 0. Again perhaps the participant was considering the task effort only instead of the general feeling of immersion. This indicates that the word "involvement" may not worked as expected with these subjects.

The only large difference in fun is from subject 35 (who also had a huge divergence in the comparison of immersion). GEQ does not ask the questions of immersion or fun directly, so perhaps this was more of a case that the participant didn't like the game and was biased to answer positively during the gameplay review (when the questions were very direct).

Finally it is noticeable how opposed scores consistently vary. Subject 4, for instance, didn't like at all the game she played, as it can be observed by the high frustration (level 3) and totally low immersion and fun (levels 0). A similar situation occurred with subject 26, with a self-reported low level of fun but a relatively high report level of frustration. Subject 23 also self-reported to have not had fun at all with her game, but seems to have experienced some frustration and some immersion. All of these subjects played the game Kraven Manor.

A t-test was also performed for a statistical significance comparison of the GEQ and each review questionnaire (frustration, immersion and fun). The p-values are presented in figure 4.8, where the yellow line indicates the employed level of significance of 0.05 (i.e. 5%). The null hypothesis is that there is no difference between the GEQ and the review scores. As it can be seen, the only statistically significant differences ($p < 0.05$, so the null hypothesis is rejected) are found in the scoring of immersion, for subjects 2, 15, 30 and 35 (subject 8 had $p = 0.05$). This confirms that in those cases the word "involvement" was definitely not understood with the same meaning of

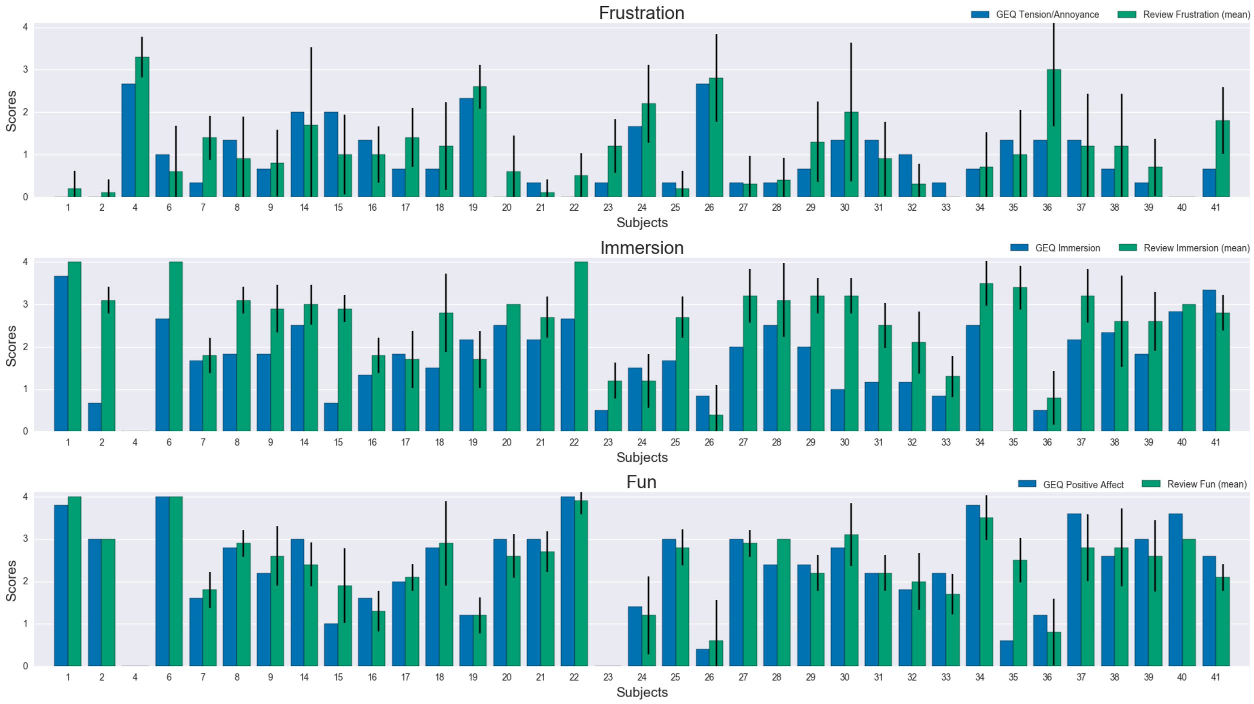


Figure 4.7: *Comparison of scores in GEQ and the gameplay review*

immersion as described in GEQ.

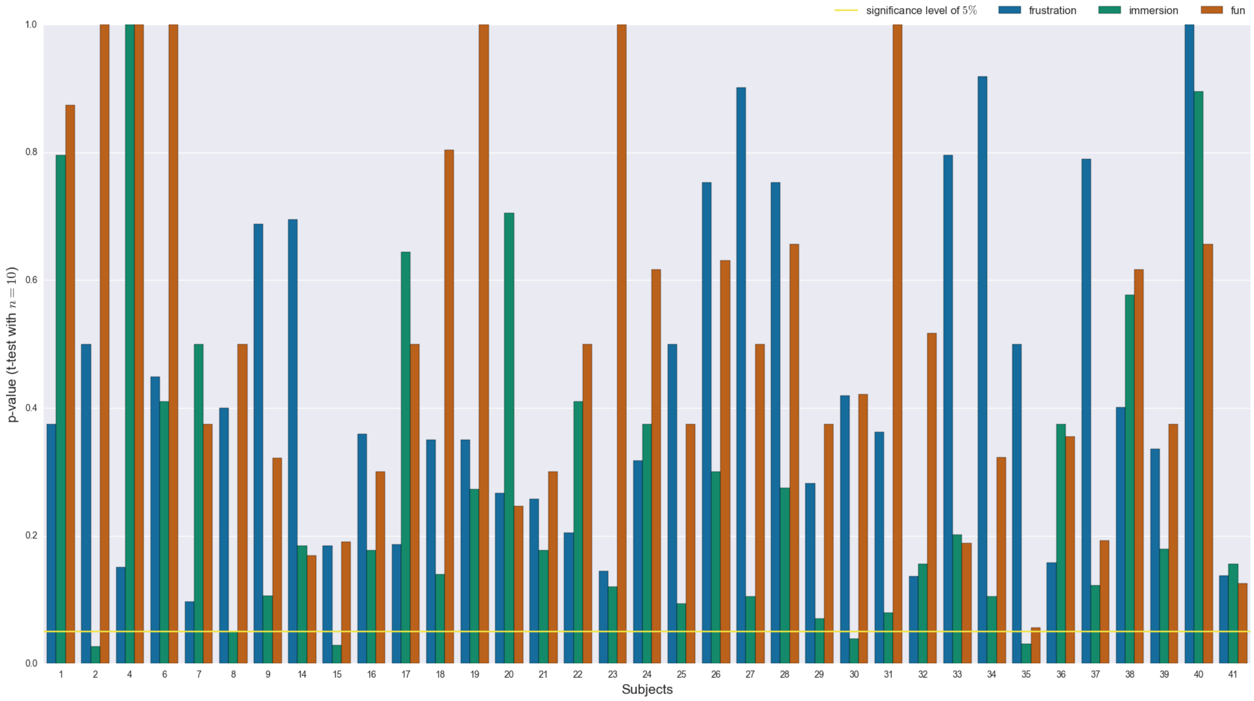


Figure 4.8: *t-test on the differences between the scores from GEQ and the gameplay review*

Chapter 5

Extraction of Features

5.1 Face Detection and Tracking of Landmarks

The main source of information available for this work to attempt the assessment of fun are the videos containing the faces of players. Each video is composed of a set of digital images – the video frames – which are sequentially captured in time in a given frame rate, measure in frames per second.

A digital image is a representations of visual data that is either captured from the real world or created artificially. Mathematically it is a two-dimensional function $f(x, y)$ of discrete values (pixels) for also discrete coordinates x and y. The pixel values are sampled from the light intensities (grey levels) taken from a real scene, for instance. In the spatial domain, a digital image is manipulated through a bidimensional matrix of pixel values in a given scale (commonly a real value in range $[0, 1]$ or an integer value in range $[0, 255]$) and with one or more bands (depending on the colour model, which frequently is RGB: one band for each colour, red, green and blue) (Gonzalez and Woods, 2002, p.1-2).

The detection of faces in digital images is in the core of the problem studied by this research work. Whichever information is needed to help identifying fun, it must be extracted from human faces digitally represented as a set of pixels. Without detecting the face and separating just its region for analysis, the amount of data to process can be very large. For example, a single frame of a video captured with HD resolution using an RGB colour model has $1,280 \times 720 = 921,600$ pixels with 3 band values each, which results in the need to process more than 2 million values per frame.

So, a first and very important step, called Face Detection, is to identify and separate only the region where a face is found. This is usually performed in a grey-scaled version of the original RGB image, because the information of colour is not necessary to describe the facial expressions. As it will be seen, the texture patterns found in the image local edges are more discriminative than colour.

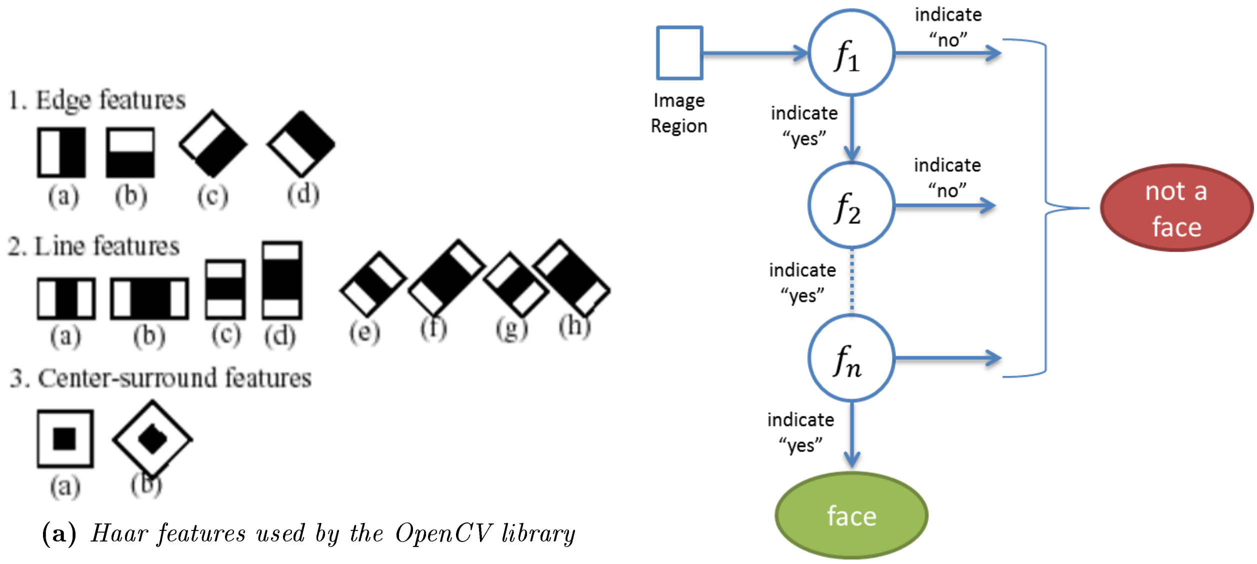
A second important task is locating facial features (eyes, mouth, nose, eyebrows, etc) inside the face region. It is from the changes in those features that the human facial expressions are expressed: a person smiles when happy or bends her eyebrows when disgusted, for example. Hence the location of these features is needed to extract useful data features. Also, the amount of data needed for processing can be further reduced by using only pixel information around those locations. The detection of the location of facial features through coordinates on the image is called Landmark Tracking.

5.1.1 Detection of the Face Region

There have been many proposed solutions for detecting faces in images, using heuristic (histogram analysis), colour-based (back-projection from colour histogram) and template matching (searching for the features from a template image into another) methods, but they are all very sensitive to rotation or present high false-positive rates. The most modern approaches rely on Machine Learning methods, with its most popular algorithm: the Viola and Jones (2001)'s algorithm. This algorithm, also known as the Cascade detector, uses edge features to represent objects in images and it is largely used to detect human faces or individual features such as eyes, mouth and nose.

The Cascade detector works by iteratively searching for an object in an image considering different window sizes. A small set of very simple classifiers, used with the AdaBoost algorithm (Freund and Schapire, 1997), are trained from a large set of positive and negative image samples of the object of interest, from which it is learnt thresholds of mean pixel values calculated from small binary masking features called Haar features (figure 5.1a). Each classifier labels an image window as containing or not containing the object of interest, depending on the mean pixel value calculated being smaller or bigger than the learnt threshold. The classifiers are not able to characterize an object by themselves, but when they are verified in a sequence (or cascade, hence the name of the algorithm) they can quickly rule out a region which clearly does not contain the object (if one classifier f_n indicates a "not-containing" label) and can robustly identify a region that does contain the object (when all classifiers in the cascade indicate a "containing" label). Figure 5.1b illustrates this process for the detection of a human face.

Cascades are very robust to noise and scale, because of the use of multiple small Haar features to characterize an object. They also have a very good performance because the Haar features can be scaled much faster than the image being searched, in order to find objects with different scales. However, a great number of positive samples of the object of interest is required for good detection of objects, particularly when they might be found with different orientations.



(a) *Haar features used by the OpenCV library*

(b) *Illustration of the the Cascade process with weak-classifiers f_1, f_2, ..., f_n*

Figure 5.1: *Illustrations of the Cascade algorithm*

5.1.2 Tracking of Facial Landmarks

Similarly there are many proposed solutions to locate facial features in digital images. The Cascade detector has been used with positive images of the desired features (eyes, mouths, noses, etc).

But the location of the coordinates of specific landmarks is commonly preferred. Using the idea of tracking from one frame into another, there are algorithms capable of estimating the displacements of a previously given set of points. For instance, Mean-Shift (Yizong Cheng, 1995), a very simple algorithm that iteratively shifts a data point to the average of the points in its neighbourhood, has been used to track landmarks manually annotated in a first frame. Other methods use optical flow, the estimation of apparent movement of features based on the gradients of the image signal between two or more frames apart, to relocate randomly or manually assigned landmarks. An example, available in the OpenCV library[1], is the Lucas and Kanade (1981) algorithm.

A problem with these tracking algorithms is that they depend on an initial (usually manual) annotation of the landmarks. To cope with that, other solutions have been proposed that try to fit a given model of the object of interest considering its colour or texture and shape. Hence, the model could be used from start, by randomly assigning the coordinates of the landmarks and then adjusting them to the best positions according to the model. The most famous example of this approach is the Active Appearance Model (AAM) (Cootes *et al.*, 2001).

The AAM works by interactively adjusting (fitting) an image with a statistical model, represented in terms of the principal components of shape and brightness of a set of connected points that characterize the object of interest. During the training phase, image samples annotated with connected landmarks (the shape model) placed over the pixels of the object on the image (the brightness model) are processed with Principal Component Analysis (PCA) in order to find the main modes of displacement and the main variations of grey level that characterize the object. The marks are connected according to relevant features of the object of interest (for instance, the jaw line and the contour of the eyes when using to track landmarks of a face), in a way that they can capture the "hinged" behaviour of the landmarks of the object in real world. A second PCA combines the two models into a single one, so shape and grey-level limit each other's adjustment, so the best fit of an object can be found considering both its appearance and form.

The fitting is performed initially randomly guessing the positions of the landmarks in an image, and then performing an optimization process to reduce the displacement error of the coordinates of each landmark until convergence (figure 5.2 illustrates one step of this process, in which the landmarks are moved from the coordinates in red to the coordinates in green). The process converges when there is no more significant changes in the positions of the landmarks. The initial guessing is usually performed inside a region previously detected with a Cascade, and the tracking from one frame into the next reuses the previously obtained coordinates.

The AAM algorithm is very robust to noise and orientation, as long as the shape model is built with enough landmarks that well-characterize the object of interest. In the case of the human face, the model should connect the landmarks of facial features that move together, like the jaw line, each of the eyebrows, the inner and outer lines of the lips, the contour of the eyes, and the nose bridge. Hence it is very usual to employ about 68 landmarks to model the whole face, even though the face can be tracked with just a few marks (at the eyes and mouth corners, for instance). By itself the AAM algorithm has difficulties to cope with rotation (particularly yaw: the rotation around the top to bottom axis, when the head is moved from side to side), since the model only implements a two-dimensional representation of an object that is three-dimensional in the real world.

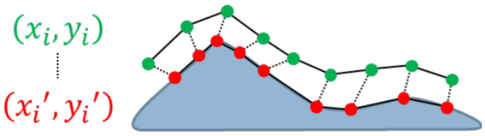


Figure 5.2: *Illustration of one step in the model fitting performed by AAM in landmark coordinates*

[1] http://opencv.org/

5.1.3 Implemented Face Detector and Tracker

When this research work started, there was no open-access library that already implemented face detection and tracking. So an implementation was started based on the Cascade detector provided by the OpenCV library and the development of an AAM implementation with the help of chapter 6 of Baggio *et al.* (2012, p.212-223)'s book. The code was created in C++ to produce a prototype that didn't achieve a good tracking quality[2]. Later on, two open-access libraries with very good face detection and tracking were found, and consequently that custom implementation was halted.

The CSIRO Face Analysis SDK (Cox *et al.*, 2013) was the first one evaluated. It is a C++ library licensed under the GPL2 license that detects and tracks human faces using 68 landmarks. The regularized landmark mean-shift (Saragih *et al.*, 2011) approach employed for the face model fitting considerably improves the tracking under occlusion conditions. In 2014, Dlib (King, 2009) – a machine learning toolkit licensed under the Boost Software License, available for C++ and Python – provided a very good face detection and tracking implementation based on a cascade of estimators (an ensamble of regression trees) that optimize the sum of square error losses to perform the fitting (Kazemi and Sullivan, 2014). This implementation also tracks 68 landmarks and copes really well with missing data from partial occlusions. So, the Dlib implemetation was chosen for the detection and tracking of faces in this project.

The face detection and tracking solution used in this work relies on Python, OpenCV and Dlib, and it was constructed as follows:

1. The image where the face is being searched is scaled down by a factor of 4.
2. Dlib's implementation of the Cascade detector is applied to the scaled image, resulting in the region where a face is found.
3. The coordinates of the face region are scaled back to the image's original resolution.
4. That region is given to Dlib's shape predictor to locate the coordinates of the 68 landmarks.

It was a design choice to re-detected the face on each frame of the videos instead of detecting it only once and keep just track the landmarks from one frame into the next. The reason is because this approach yielded much better results when there was quick head movements. Even by performing a detection in each frame, a good performance could be still achieved because of the simple adjustment describe in steps 1 and 3. By scaling down the original image before doing the detection, the Cascade didn't need to scale much its Haar features and the search was very fast. Also, the scaled-back region coordinates were precise enough for the shape predictor to work with good accuracy in the fitting. The initial detected region was discarded and replaced by the bounding box of the landmarks found by the fitting algorithm.

Figure 5.3 illustrates the results of this solution applied to a clip of 30 frames from the FGNet Talking Face Database (Crowley, 2004), a video database of a person talking to an interviewer which is commonly employed as a test bed for face tracking algorithms. The face region detected by the Cascade is depicted by the red rectangle and the face landmarks found by the landmark tracker are depicted by the yellow points and lines (the points are connected according to Dlib's shape model).

5.2 Immersion Features

Immersion is directly connected to attention, so it is supposedly related to increases in the reception of sensory data. Possible features that can be measured from facial images are variations in the

[2] A video of this prototype is available at https://www.youtube.com/watch?v=_tC7L_lGngc

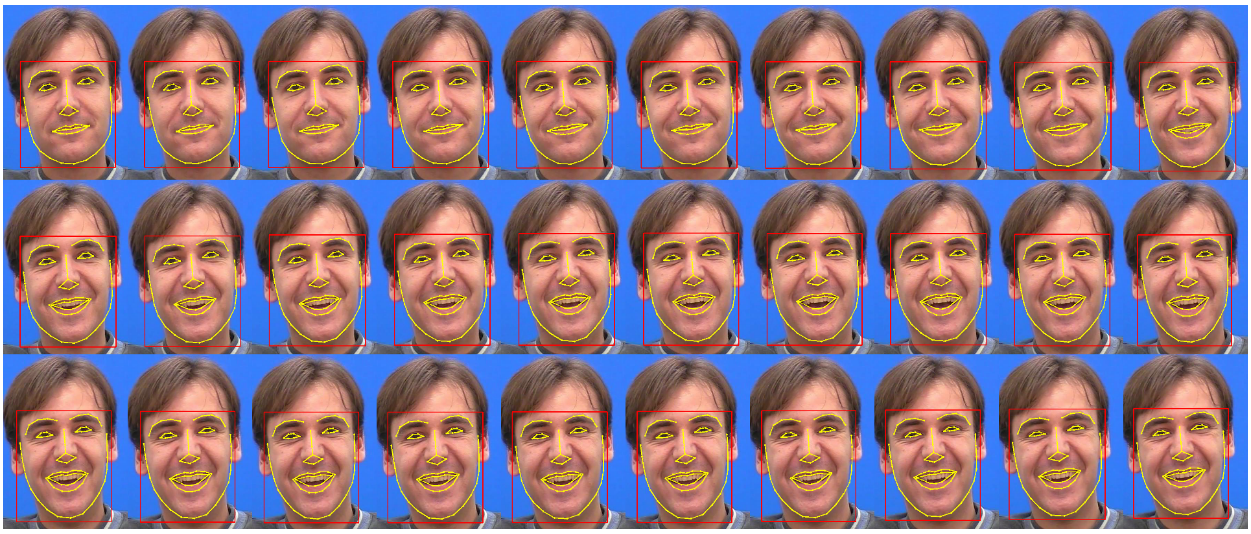


Figure 5.3: *A short clip (frames 20–49 of the FGNet Talking Face Database) with a face being tracked*

distance from the camera, in the gaze direction, in the pupil diameter, and in the blink rate. The distance varies as the player leans the body towards or away the game, the gaze follows the object of interest and the blink rate is supposedly reduced as interest grows.

Among these features, the variations of distance and blink rate seem the best to use. Pupil diameter is very difficult to estimate if there is no enough resolution of the eye region. The data captured used a camera focusing the face of players in a "free" configuration, similar to that would be expected in a normal situation. That is, the face is positioned in the centre of the field of view, and there is not a high resolution of the eye regions. Also, as it has been discussed in the previous chapter, the pupil diameter is more sensitive to light reflexes than it is to cognitive activities (Chen and Epps, 2013, p.112).

Eye gaze is also disregarded because the visual patterns of games vary considerably. While some games require the player to move her gaze around the screen, others concentrate the player's gaze on the screen centre (El-nasr and Yan, 2006, p.6). In that way, it shall probably be very hard to differentiate gaze changes (fixations and saccades) that are merely due to the game visual pattern from the ones that are due to interesting events. As consequence, this work focused on two features for the classification of immersion: the gradient of the estimated distance between the face and the camera, and the estimated blink rate, measured in blinks per minute.

5.2.1 Gradient of the Face Distance

Considering that the camera and the game equipment are fixed, whenever a player leans her body towards or away from the game the distance between the player's face and the camera varies. By knowing the coordinates of a set of landmarks of the face in the image, the coordinates of that same points in a 3D reference model of the face, and some parameters of the camera that captured the image, it is possible to estimate the position of the face in the tridimensional world and from that estimate the distance between the face and the camera.

Pose Estimation

Pose estimation is based on the idea that the coordinates of a known point P in the World Coordinate system, where the real object exists, can be easily transformed into 3D coordinates of that point in the Camera Coordinate system, if the pose (i.e. rotation and translation) of the object with respect to the camera is known (Mallick, 2016). This is done by solving a simple linear equation in which the coordinates of the point are multiplied by the rotation matrix and added to the translation matrix, resulting in the 3D coordinates of the point in the Camera Coordinate system. That remapped 3D point can then be projected to a 2D point p in the Image Coordinate system (which is the plane of the captured image), if some camera parameters are known: the focal length and the distortion matrix. Figure 5.4 illustrates this process, in which o is the origin of the camera, the plane is the image captured, R and t are the rotation and translation matrices of the object, and the segment oc is the focal length of the camera.

In the pose estimation problem, the 2D coordinates on the Image Coordinate system are known and the 3D coordinates of the real object can be assumed within an arbitrary World Coordinate system. But the 3D object rotation and translation are not known. So they are estimated iteratively by the reduction of a projection error: the sum of squared distances between the 3D points projected into the Camera Coordinate system and the 2D points in the Image Coordinate system. This method is called Levenberg-Marquardt Optimization(Mallick, 2016), and it is already implemented in OpenCV via a function called “solvePnP”.

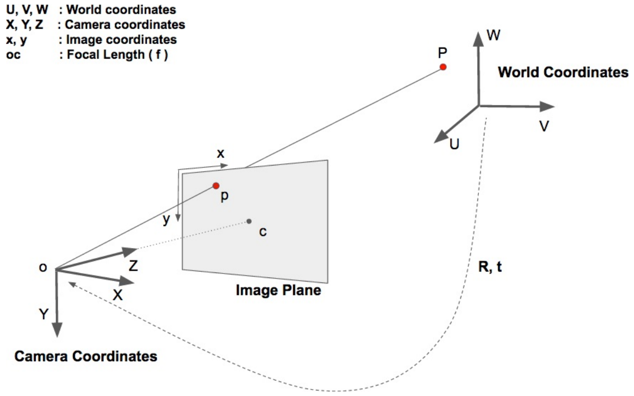


Figure 5.4: *Projection of a point P into the image plane – reproduced with permission from Mallick (2016)*

Although the pose estimation algorithm is implemented in OpenCV and the coordinates of the landmarks are given by the face tracker implemented, the camera parameters are still required. Focal length is the distance vector from the camera’s lens to the camera’s sensor, which is needed for the calculus of the projection into the image plane. The distortion matrix contains the coefficients that represent the natural distortions caused by the lens. Distortion is a problem particularly important with cheap pin-hole cameras: radial distortion makes straight lines to appear curved and tangential distortion makes some areas to appear nearer than expected. These coefficients are fixed for a given camera, so it is possible to pre-process the images to remove distortions before proceeding with the calculations if the coefficients are known.

Camera Calibration

The process of estimating the camera parameters is called calibration. It employs several images of an object with a pattern easy to identify (usually a chess-board or circular grid-like surface), captured from different angles, in different positions, and with the same resolution that will be later used. The number of points (the corners of a checked-board or the circles of a grid-board) and their relative distances (the size of the squares or the circles) are known, so the focal length can be estimated in the same metric unity (usually in milimetres). The distortion coefficients are specified as a percentage of the camera’s field height (Hollows, 2011).

The OpenCV library has the functions “findChessboardCorners”, “findCirclesGrid” and “calibrateCamera” to help with the calibration process. With the known mapping between the 2D and 3D points of the patterns captured, the parameters are iteratively estimate by computing a rectification transformation that makes the camera optical axis parallel. A re-projection error is then calculated to evaluate the quality of the estimated parameters. The error is the average distance (in pixels) between the points found in the original image and the same points found in the image undistorted with the coefficients previously produced. In a well-performed calibration the re-projection error is supposed to be between 0 and 1, but it is desired to have the error as near as possible to 0.

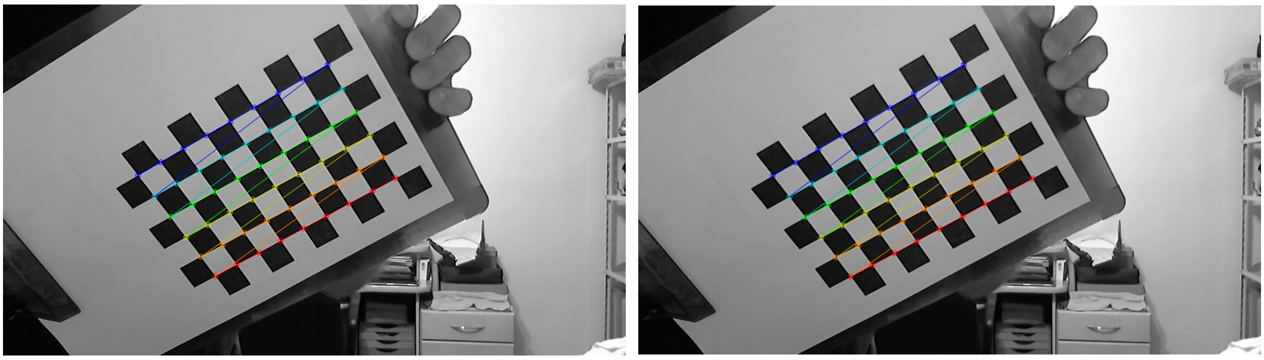
The Logitech C270 camera used in this project was calibrated with the described procedure using a checked-board of 9x7 corners (i.e. 10x8 squares) with squares of 18 milimetres in size (figure 5.5). Many calibration attempts have been executed with different images of the checked-board, and the best setting achieved a re-projection error of 0.366 using 24 images of the checked-board in different positions and angles. The camera matrix, with the focal length given by f_x and f_y (for both axes) and the optical centre given by c_x and c_y, obtained was[3]:

$$A = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1470.178 & 0 & 654.919 \\ 0 & 1476.419 & 364.055 \\ 0 & 0 & 1 \end{bmatrix} \tag{5.1}$$

The distortion vector, with the radial distortion coefficients k_1 and k_2 and the tangential distortion coefficients p_1 and p_2, obtained was:

$$D = [k_1, k_2, p_1, p_2] = [0.004, 0.220, 0.000, 0.002] \tag{5.2}$$

As it can be noticed the distortion coefficients are very low, indicating that the camera has a lens of high quality. In a matter of fact, it is very difficult to perceive with a naked eye the removal of distortions from the original to the undistorted images in figure 5.5. But, despite the camera quality, some distortion exists and hence the coefficients found were used for the estimation of the pose of the players' faces.



(a) *Original image* **(b)** *Undistorted image based on the coefficients found*

Figure 5.5: *Example images of the checked-board pattern used in calibration – the points used for the estimation are drawn in colours for reference*

3D Face Model

In order to estimate the pose of a human face, some of the points of a 3D model of a face are needed. They should be positioned in the World Coordinate system in a way that they capture an average width, height and depth of a human face. Ideally the measurements of each player should be taken for specific estimations, but that is not a practical task. The model used was created by Mallick (2016), in an arbitrary scale. Other more complex models were tried, such as the Candide3

[3] All values are displayed with only 3 decimals, but the real values are calculated with a 10-decimal precision

model (Ahlberg, 2001). But Mallick's simple model was the one that produced the most consistent results.

The model's 3D coordinates are illustrated by figure 5.6 and presented in table 5.1. They rely only on the 6 landmarks that are mostly fixed in the face, that is, that do not move much independently, despite the facial expressions, but still capture the face dimensions (length, width and depth). The coordinates were arbitrarily defined, meaning that an specific unity scale such as milimetres was not considered in its creation, but the model is consistent regarding the relative positioning and distances between the facial points. Also, the origin is defined in the nose landmark, so the distance is estimated from the tip of the nose. This means that small variations are expected when the head is bent forwards or backwards and the nose gets closer or further away from the camera.

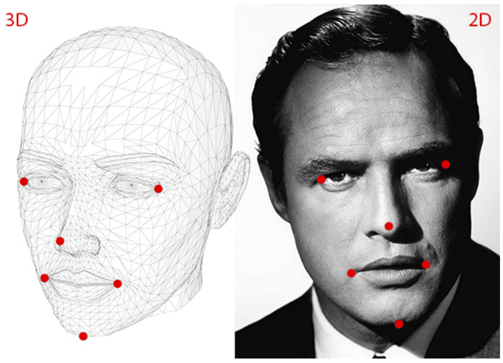


Figure 5.6: *The 3D model used to pose estimation – reproduced with permission from Mallick (2016)*

Table 5.1: *Coordinates of the vertices in the 3D model of the face*

Facial Point	**3D Coordinates**		
	x	***y***	***z***
Tip of the nose	0.0	0.0	0.0
Chin	0.0	−330.0	−65.0
Left corner of the left eye	−225.0	170.0	−135.0
Right corner of the right eye	225.0	170.0	−135.0
Left corner of the mouth	−150.0	−150.0	−125.0
Right corner of the mouth	150.0	−150.0	−125.0

Gradient of the Face Distance

Once the 3D pose of the object is obtained, the z coordinate of its translation regarding the camera is a good estimation of the distance between the object and the camera. The unity of this measurement is the unity of camera's focal length, given in the unity used for camera calibration. Through experimentation it was verified that, with the calibrated camera used, the arbitrary model used is in a scale close to 10 times bigger than if the face points were defined in milimetres. So the value of the z axis of the object posed was divided by 100 (10 times for the model's scale and 10 times for the calibration unity in milimetres) in order to have the estimated distance in centimetres.

The evaluation of the distances calculated was performed by fixating the camera to a movable base, pointing it to a fixed picture of a female model (taken from the page of a magazine and chosen because the face covered almost the whole page) and moving the base closer and further from the picture. A ruler was placed between the picture and the camera. The picture used is slightly smaller than a normal human face, so a difference was expected (due to the distance from the camera that originally took the picture). The intention was to verify that the estimated distance increased and decreased linearly with the physical distance measured by the fixed ruler, and that was indeed the case. A more informal evaluation was also made with live streaming video. It was possible to observe that the estimated distance varies consistently as the face gets closer or further from the camera. Nevertheless, the distance estimated has a fixed error margin of circa 15 centimetres, most probably due to the arbitrary 3D model used.

That error margin in these estimations was disregarded as an important issue because the distance is not really the best feature to characterize immersion. Its variation is more relevant because it better describes the behaviour of leaning the body towards or away from the game. If a person's head is still, the gradient is close to 0. Only when she moves to get closer or away from the camera that value changes. So an accurate estimation of the distance is not a problem for this work's intention as long as the distance variation is consistent.

In order to calculate the gradients, the estimated distance values from each of the video frames were added to a list. The distance was represented in centimetres, with a 10 points decimal precision, so the gradients are in the same unity. Missing values, due to face detection failure in a given frame, where handled by simply copying the last valid estimated distance. The gradients were calculated with an implementation available from the Numpy library in Python[4], using second order accurate central differences for the values in the interior of the list, first differences for the values at the boundaries of the list, and a distance of 1 frame. This way, the list of gradients produced had the exact same size of the list of distances, that is, one value per frame.

5.2.2 Blink Rate

A blink is caused by the voluntary, reflexive or spontaneous movement of the eyelids in response to incentives from the outside world (Roshani, 2011). Therefore, blink detection requires the comparison of two (or more) frames in order to identify the moment when the eye was closed. Basically this can be achieved with two different approaches (Olatsek, 2013): by analysing the frame differences regarding pixel intensity or texture, or by analysing the frame differences regarding movement. The former approach basically compares the mean pixel differences between two frames in order to detect when the eye was opened in one frame and closed in the next. Variance mapping, correlation measures and colour histogram back-projection are some example of methods used. The latter approach focuses on detecting the up and down movement of the eyelids that always accompany a blink. Optical flow or AAM methods are used to track moving local features in the eye region, which are then used to measure the vertical displacement.

Olatsek (2013) has compared these two approaches by using back-projection and optical flow, with the Lucas and Kanade (1981) algorithm. In the first approach, a colour histogram of the whole face is used in back-projection to represent the probabilities of a pixel to belong to skin. The pupil and the sclera are very different than the eyelids, so an image of a closed eye is expected to produce a higher average probability of skin than an image of an opened eye. Bigger variations of skin probabilities between two frames indicate a blink.

The optical flow is used to track randomly selected points in two different regions (of each eye and the whole face, previously detected with Cascades) between two frames. The average diferences in positions of the local features from one frame to the next is called the displacement. The displacement of the eye features is normalized by the displacement of the whole face in order to ignore head movements, and then it is compared to a threshold empirically obtained. If the displacement is bigger than that threshold, a blink is detected.

The author tested both approaches against his own database and the FGNet Talking Face Database

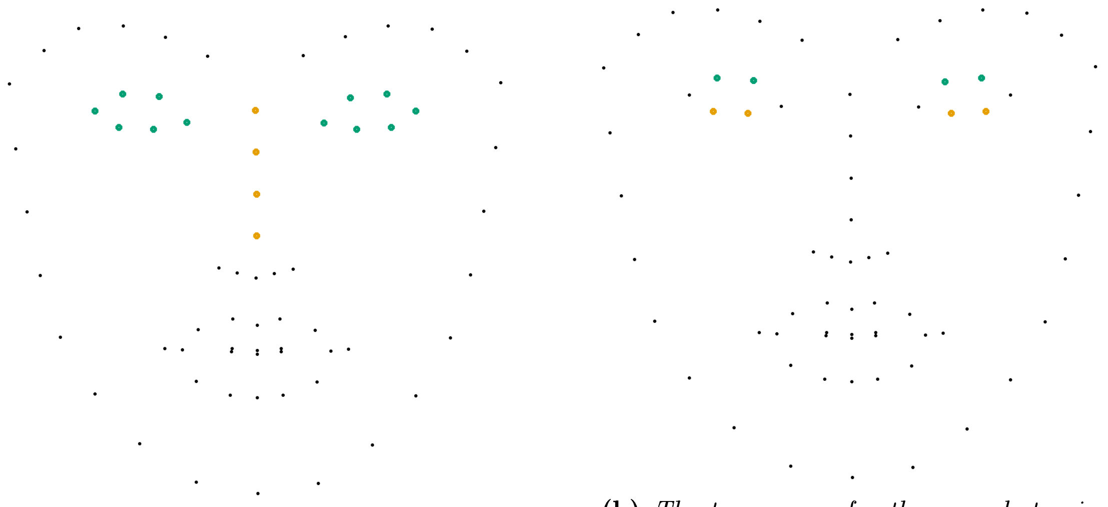
[4] https://docs.scipy.org/doc/numpy/reference/generated/numpy.gradient.html

(Crowley, 2004) (already used in this work to test the face tracker), reporting good accuracies with both solutions (85.25% for the back-projection and 98.36% for the displacement measurement from optical flow), concluding that the displacement measurement was the best method. Following this suggestion, the displacement measurement was used in this work. However, the application of optical flow was unnecessary because the landmarks of the eyelids are already being detected by the face tracker in each video frame.

Blink Detection

The blink detection is performed in two steps. First, the movement of the eyelids is normalized considering the movement of the whole face, in order to distinguish a possible blink from a quick movement of the head. If a blink is possible, then the vertical eyelid movement (that is, the movement of the upper and lower eyelids closing against each other) is verified to definitely indicate a blink.

Two groups of landmarks are used for each step, as illustrated in figure 5.7. The initial verification of movement uses the landmarks of both eyes for one group and the landmarks of the nose bridge for the other (respectively the groups in green and orange in figure 5.7a). The nose bridge represents the head movement well enough, since they moves together with the whole face and almost do not suffer from geometric deformation due to different facial expressions. The second verification uses the upper eyelid features for one group and the lower eyelid features for the other (respectively the groups in green and orange in figure 5.7b). In the two verifications the landmarks of both eyes are used together in each group because blinks of individual eyes are voluntary blinks, intentionally disregarded for blink detection.



(a) *The two groups for the first step in blink detection: eye and nose bridge landmarks*

(b) *The two groups for the second step in blink detection: upper and lower eyelid landmarks*

Figure 5.7: *Groups used for the calculus of eyelid displacement in the blink detection*

In the first step, the displacement within each group is calculated by averaging the the Euclidean distance of each landmark vectors between one frame and the next. In equation 5.3, $p_{i,k}$ refers to the coordinate vector of a landmark i in frame k, and n is the number of features in the group.

$$\text{displacement} = \sum_{i=1}^{n} \frac{\sqrt{(p_{i,k+1} - p_{i,k})^2}}{n} \tag{5.3}$$

The absolute difference between the displacements of the eyes and the nose bridge groups is checked against a threshold calculated as $t_1 = {}^{face.height}/_{150}$ (where the face height is given in pixels). If the absolute difference of displacements is bigger than that threshold, a possible blink is indicated and the next step is verified. Otherwise, a no-blink label is immediately accepted.

The reference paper used the value 165 in the denominator of this threshold value, but the tests performed in this work suggested that the smaller value provided better results. The difference might be due the face height used in the reference work being given by the height of the window detected by the Cascade, while the face height in this work is given by the smallest bounding box of the face landmarks detected (the Cascade is only used to help fitting the face model).

If a possible blink is detected by the first step, then the vertical displacement of only the eye lids is verified. In this step, two groups are formed from the landmarks of the upper and lower eyelids, and their movement from one frame into another are calculated in the same fashion as before (i.e. using equation 5.3). The threshold used for comparison is calculated as $t_2 = {}^{face.height}/300$. The reference work used a denominator value of 110, but it also considered only the vertical axis (the coordinate values y) in the displacement calculus. The Euclidean distance was also considered as a good choice here because it may help making the blink detection robust to rotation of the face on the roll axis (bend the head left or right). Again, the denominator value of 300 was obtained empirically as a good choice considering the differences in implementation and the test results.

The blink rate is obtained by counting the number of detected blinks in the last minute of the video progress. The elapsed time of the video is calculated from the frame number using the frame rate of the recording: 30 frames per second.

Evaluation

The blink detector created was also validated against the FGNet database. This database contains 5000 frames in which the person, engaged in a conversation, blinks naturally (that is, the blinks are either spontaneous or reflex). The database does not provide an annotation of the blinks, so they had to be manually created. A total of 58 blinks were accounted in the video. Figure 5.8 presents the test results. A general accuracy of 84.5% was obtained, with 3 false positives and 6 false negatives. Nevertheless, the real accuracy is better than that, since the detection has not always happened in the exact same frame of the annotation. At least two false positives are very close to other two false negatives, what indicates that those blinks might blink have been correctly detected, but just with a small delay (they are near frames 1280 and 1430).

One false negative (near frame 3975) and one false positive (near frame 4180) were due to two very quick blinks in sequence, which indicates one of the difficulties of this algorithm. Also, as with the work of Olatsek (2013), the algorithm tends to fail when the person lowers the head because the tracking of features becomes much more difficult and great variations between frames are common. This situation does not happen in the FGNet database, so it was not observed in this tests. And it does not happen frequently in the domain of this work because the players' faces are focused on the game screen most of the time.

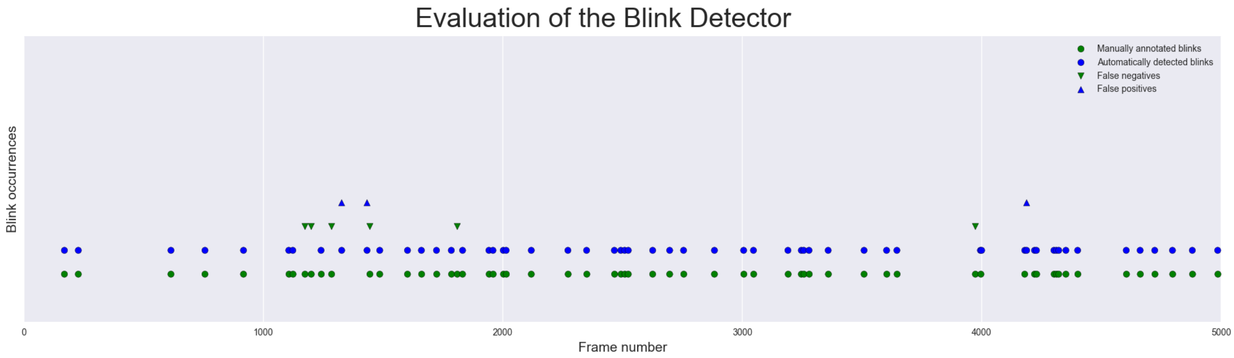


Figure 5.8: *Comparison of the annotated and detected blinks in the FGNet Talking Face Database*

5.3 Emotion Features

Emotions are directly connected to facial expressions, and at least the prototypic emotions are well characterized by variations in facial muscles. The human face is a very important medium of communication of both verbal and non-verbal cues, which change the facial muscles in very distinct ways. These changes might be visually inspected either via geometric measurements between the features or via texture changes in the whole face or in localized areas.

They prototypic emotions have been accurately detected using geometric and texture-based features. Geometric features are measurements taken from the permanent facial features, such as elevation of eyebrows, distance between eyebrows, aperture of eyes, aperture of mouth, etc, commonly normalized by a known fixed value such as the distance between the eyes. Texture-based features are descriptors of the patterns in the face image created by the transient facial features, such as lines, wrinkles and furrows. They are usually extracted with band pass filters in the spatial domain.

While the normalized geometric features are naturally robust to variations in rotation and lighting, they depend on a very precise detection of the facial landmarks. Texture-based features are also robust to variations in lighting and rotation, but are better to capture more subtle variations in facial expressions (Bettadapura, 2012; Fasel and Luettin, 2003). Indeed it has been shown that texture descriptors like Gabor filters yield better classification results than geometric features, even if they are used in combination (Zhang *et al.*, 1998). Gabor filters are a very common way to characterize image textures, being largely used in the emotion classification literature.

5.3.1 Responses to a Bank of Gabor Filters

Gabor Filters

A Gabor filter (Daugman, 1985) is a linear filter used for edge detection. It is created from a sinusoidal carrier attenuated by a Gaussian envelope. The filter has real and imaginary components representing orthogonal directions, which may be used independently or combined into a complex number. Equation 5.4 presents the formal definitions of the real and imaginary components of a two-dimensional Gabor filter (called a kernel in the spatial domain), where the impulse value of each component is defined by the sinusoidal wave multiplied by the Gaussian function. In the equation, λ is the signal wavelength[5], θ is the orientation of the normal to the parallel edges the filter responds to, ψ is the signal's phase offset, σ is the standard deviation of the Gaussian envelope and γ is the spatial aspect ratio of the filter (defining its ellipticity, such as 1 means a symmetrical kernel).

$$
\begin{aligned}
g_{\text{real}}\left(x,y,\lambda,\theta,\psi,\sigma,\gamma\right) &= \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda}+\psi\right)\\
g_{\text{imag}}\left(x,y,\lambda,\theta,\psi,\sigma,\gamma\right) &= \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\sin\left(2\pi\frac{x'}{\lambda}+\psi\right)\\
&\text{where:}\\
x' &= x\cos\theta + y\sin\theta\\
y' &= -x\sin\theta + y\cos\theta
\end{aligned}
\tag{5.4}
$$

Figure 5.9 illustrates a Gabor kernel with wavelength $\lambda = 20$ pixels and orientation $\theta = 30$ degrees

[5] In some definitions it is used the inverse of the wavelength: the spatial frequency $\xi = \frac{1}{\lambda}$

($\sigma = 0.56\lambda$) in two views: a two-dimensional view that shows how the kernel is computationally represented (as a matrix of values, that is, an image) and a three-dimensional view that allows for an easier observation of the sinusoidal carrier attenuated by the Gaussian bell towards the outer limits.

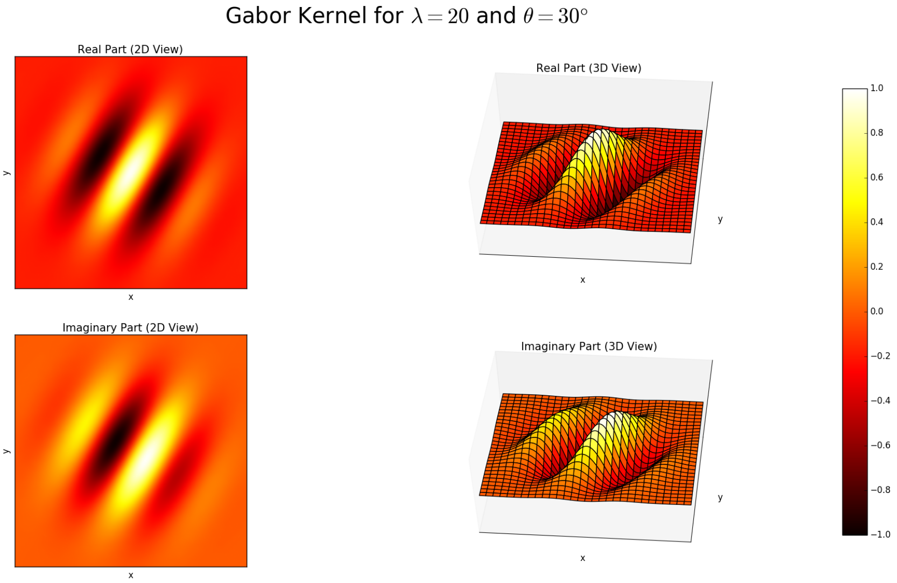


Figure 5.9: *Visualization of the real and imaginary components of an example Gabor kernel – angles are in degrees and values are normalized to* $[-1, 1]$*for easy viewing*

Filtering an image with a Gabor kernel is commonly performed by convolution in the spatial domain. The filter responses are produced by the magnitude of the real and imaginary components (response $= \sqrt{r^2 + i^2}$), and are proportional to how well the local features of the image match the scale (i.e. the wavelength, given in pixels) and the rotation (i.e. the orientation, given in radians) of the kernel used (Lyons *et al.*, 1998, p.201;Henriksen, 2007, p.86). The size of the kernel is usually calculated from the values of λ and σ in order to limite the cutoff of the filter (i.e. get values very close to 0 at the outer limits of the kernel). The aspect ratio γ is usually 1, in order for the filter to respond equally in both axes, and the phase offset ψ is usually set to $\pi/2$ by convention. Figure 5.10 illustrates the responses of different filters on a test image.

An important caveat of this method is that it is very computationally expensive. Filtering an image with a Gabor filter requires two convolutions, one for each of the real and imaginary parts, plus the calculus of the magnitudes. Nonetheless, modern implementations rely use the GPU to paralelize these computations and achieve good performance.

Bank of Gabor Filters

The filters used to process an image can be designed specifically for a problem, with fine tuning of parameters and human intervention, or created *ad hoc* from a set of kernels with different parameter values. The set of kernels are not necessarily optimal, but can represent many the variations of the image samples (Tsai, 1993, p.38). Using optimal parameters is not always a good approach, since it may lead classifiers to over fitting. The surface features of images are composed of several frequencies and orientations, so the mapping is never one to one(Henriksen, 2007, p.17).

Thus it is more common in the literature to employ a bank of filters with different scales and rotations. A Gabor bank is considered a very good general texture descriptor and the image processing with such a bank is said to be very similar to how the visual cortex of mammals work, where different cells are sensitive to specific edges of a given scale and rotation (Fasel and Luettin,

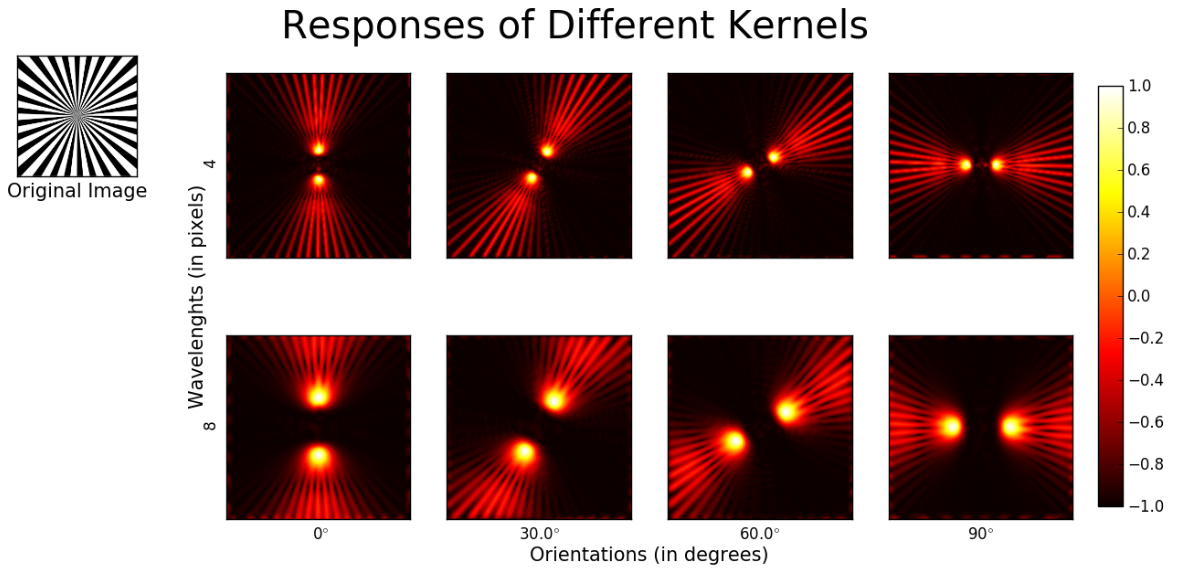


Figure 5.10: *Visualization of the filter results of different Gabor kernels – again angles are in degrees and values are normalized for easy viewing*

2003; Schindler *et al.*, 2008). Nevertheless, this approach has the important caveat of increasing the computational cost of the solution. With 2 convolutions per filter plus 32 filters, the filtering process uses a total of 64 convolutions per frame image. But it is important to remember that the amount of processing can be considerably reduced by filtering only the image region where the face was located.

Typical Gabor banks employ 32 kernels, with 8 orientations times 4 wavelengths (Henriksen, 2007, p.86). The orientations vary from 0° to 180° because they can capture enough variations in rotation. No value above 180 is used because the responses are the same, just with the phase rotated by 180°. The wavelengths (or frequencies, depending on the preferred choice of implementation) are more of a matter of debate. The most important rule of thumb is not to use wavelength values smaller than 3 or bigger than half the image size (width or height, whichever is smaller), because they produce useless results due to the fact that the image and the kernels are functions of discrete values. The literature commonly employs small arbitrarily values chosen with basis on the experience of the designer.

Regarding the choice for the other parameters, it is common to simply mimic the human visual cortex. The human visual cortex supposedly responds to signals withing a bandwidth close to 1. Since the Gabor filter is a band pass filter, the relationship $^{\sigma}/_{\lambda}$ defines the bandwidth of the filter (Henriksen, 2007, p.88), as described by equation 5.5. Considering a fixed bandwidth of 1, this ratio leads to a value of $\sigma \simeq 0.56\lambda$, which is a common choice in the literature(Grigorescu *et al.*, 2003). An aspect ratio of $\gamma = 1$, gives a symmetrical Gaussian envelope (i.e. bell shaped), which suffices because the variations in rotation are already captured by the different values of θ employed. And as it has been said, an offset of $\psi = {}^{\pi}/_{2}$ is used by convention.

$$b = \log_2 \frac{\frac{\sigma}{\lambda}\pi + \sqrt{\frac{\ln 2}{2}}}{\frac{\sigma}{\lambda}\pi - \sqrt{\frac{\ln 2}{2}}}, \quad \frac{\sigma}{\lambda} = \frac{1}{\pi}\sqrt{\frac{\ln 2}{2}} \cdot \frac{2^b + 1}{2^b - 1} \tag{5.5}$$

Gabor Bank Used

The Gabor bank used for the detection of prototypic emotions was created from 32 kernels built from the combination of the wavelengths and orientations in table 5.2 (angles are in radians, as expected by the implementation). Their real components are illustrated in figure 5.11, with images scaled to the same size and normalized to range $[-1, 1]$ for easy viewing.

Table 5.2: *Parameter values that will be used to build the Gabor filters in the bank.*

Parameter	**List of Values)**							
θ	0	$\frac{\pi}{8}$	$2\frac{\pi}{8}$	$3\frac{\pi}{8}$	$4\frac{\pi}{8}$	$5\frac{\pi}{8}$	$6\frac{\pi}{8}$	$7\frac{\pi}{8}$
λ	4	7	10	13				

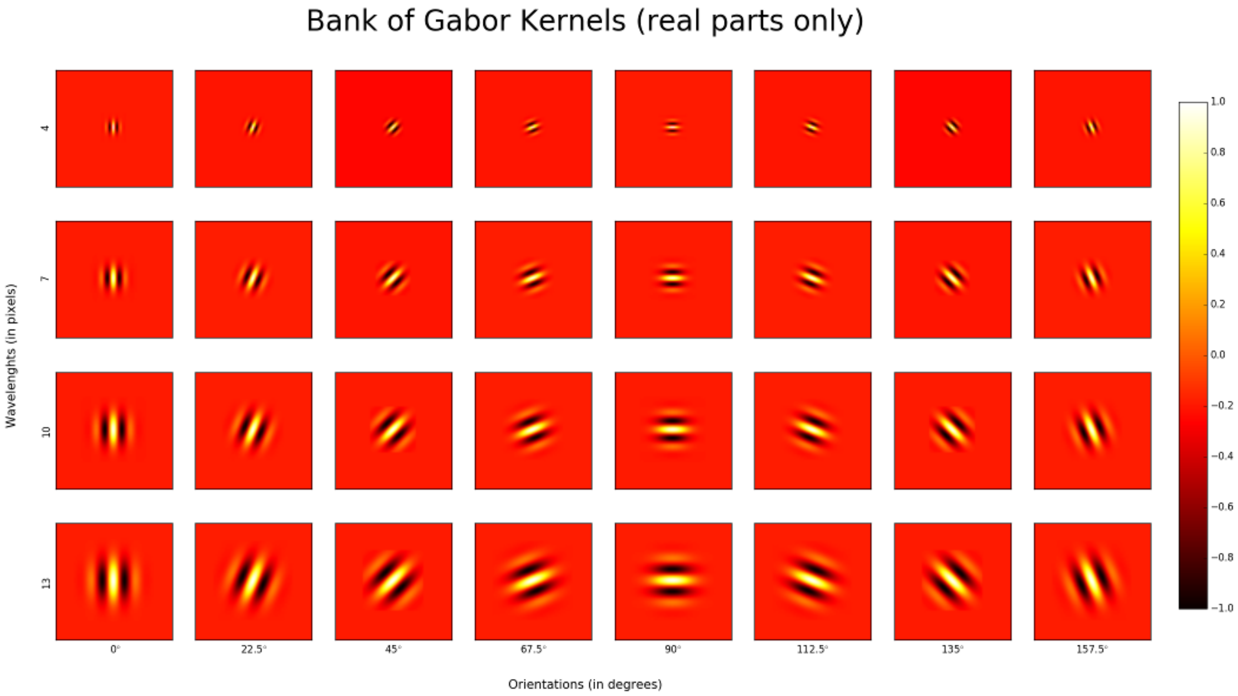


Figure 5.11: *Gabor bank used for feature extraction*

The angles were chosen as described in the literature, in order to capture a good range of rotations in all directions. The wavelengths were chosen based on the observation that even though the image resolution is high (HD), the facial landmarks have a much lower resolution.

The faces detected in the captured data have an average resolution of 337 x 333 pixels (width x height), including the whole face width and the height from the top of the eyebrows to the inferior border of the mandible. A measurement in nature very similar to the detected height is the Morphological Facial Height[6](Weinberg and Marazita, 2010). That measurement has an average of 122.59 milimetres in 30-year-old humans of both sexes. The mouth, with an average height of 16.23 milimetres in the same group, represents only 13.23% of the face height (Weinberg and Marazita, 2010).

Therefore, the expected height in pixels of that particular facial feature in the captured images is circa 44 pixels, and therefore the maximum upper limit for the wavelengths should be half of

[6] The Morphological Facial Height does not start exactly from top of the eyebrows, but from the middle point between the eyes

that (22 pixels). But the transient facial features, such as lines, wrinkles and furrows that produce textural changes in the skin due to the facial expressions, are even much smaller than that. Hence it was opted to use a smaller value near to a quarter of that (i.e. 11 pixels), but with two odd and two even wavelengths. More wavelengths could be used, but the computational cost would increase with the convolutions. Also, Gabor vectors at neighbouring pixels are highly correlated and redundant so it suffices to use the responses on the pixels marked by the landmarks found by the face tracker (Lyons *et al.*, 1998, p.201). Bigger wavelengths would produce larger pixel areas with the same response, and that is another reason to not employ wavelength values larger than 11 pixels.

Despite all this rationale, the values are still arbitrarily chosen and do not necessarily produze optimal responses for the images used. Nonetheless, it is expected that their combined use is enough to represent the variations in orientation and scale of the changes in the facial expressions. An illustration of the responses produced for a test facial image (cropped from the region detected in an arbitrary frame, and without any scaling) with the bank created from this choice of parameters is presented in figure 5.12. Again the images have been normalized for easy viewing. The choice of parameters seems to capture many distinct variations. The responses of small wavelengths capture very subtle nuances of the face, like the lip fissure and the nostrils, but are sensitive to glasses in some of the orientations (near 0 and 90 degrees). The responses of bigger wavelengths capture larger nuances, like the cheeks, the lips and the eyebrows, and are less sensitive to glasses. Responses of wavelengths bigger than 13 become very similar blurrier versions of that one, with higher activations for bigger regions of the face.

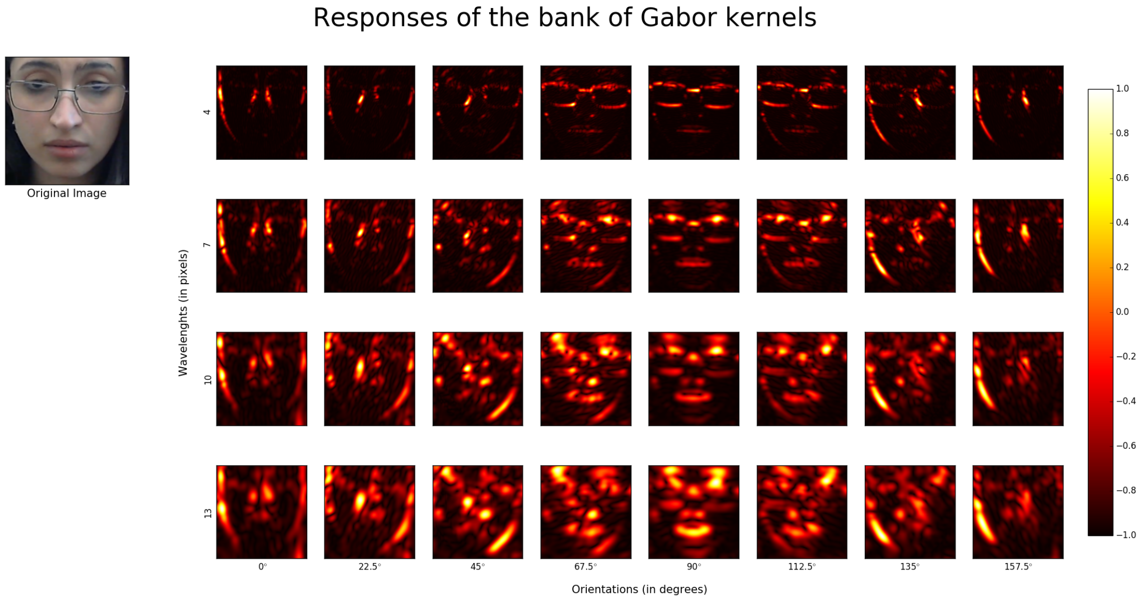


Figure 5.12: *Responses for the cropped face region of subject 23 extracted from an arbitrary frame*

Chapter 6

Towards the Assessment of Fun from Facial Images

With the features extracted from the videos collected, the next step is to build Machine Learning (ML) algorithms that should be able to predict frustration, immersion and fun from that data. This input data is not composed of only still images but of sequences of video frames. This adds complexity to the problem because the data should be processed as a sequence: information from one frame alone might not be enough to properly identify a state because the history of changes is relevant.

For example, the pleasure that results from the achievement of difficult but achievable goals (the fun aspect of challenge, related to immersion and flow) most commonly follows a sequence of frustrating events that happened when the player's first attempts have failed. Similarly, the pleasure that results from the experience of emotions (the emotional aspect of fun) might have a structured nature: in horror games, for instance, a sequence of fearful moments may lead to happiness related to relief from a tense situation.

This chapter describes how the construction of the classifiers was planned and executed, considering these issues. It starts by briefly describing the ML models used and then describes how the specific detectors have been implemented.

6.1 Machine Learning: Concepts and Algorithms

The prediction of a player's state (regarding emotions, immersion and fun) requires the construction of algorithms that should be able to indicate the state based on the features extracted from the facial images. Machine Learning (ML) techniques help with this task by building models from a particular set of samples in order to obtain generic conclusions from them. There are basically two approaches followed by ML models (Theodoridis and Koutroumbas, 2008, p.7): supervised and unsupervised learning.

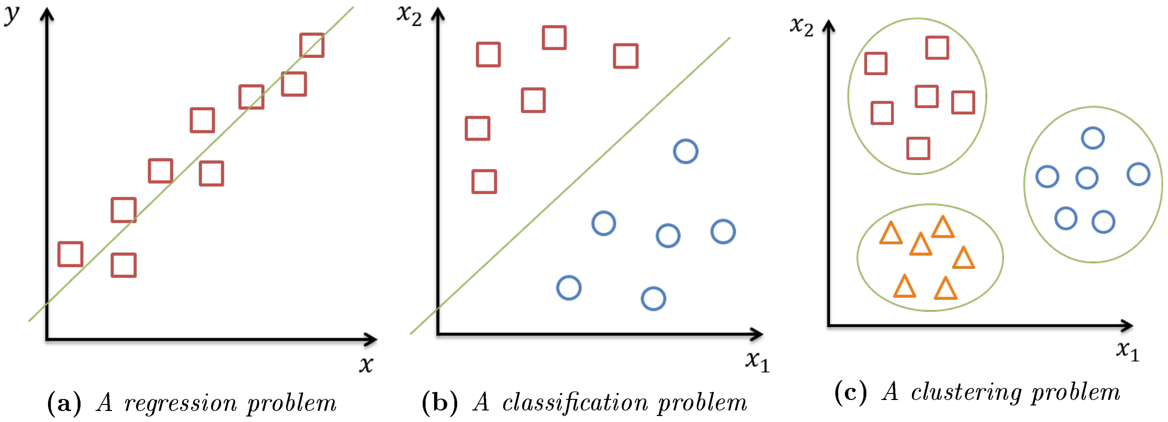
In supervised learning the source data contains samples with both the input features and their respective targets (i.e. the states or responses being modelled in the problem domain), so the model can learn as if a teacher had provided the correct answers. Therefore, the source data is called training data. A well trained supervised model should be able to correctly predict the targets of new but yet unknown samples presented. If the targets in the training data are expressed in a continuous space, the model is used for regression (the prediction of values in stock market, for instance); if the targets are expressed in terms of discrete classes, the model is used for classification (the classification of fruits as bad or good, for instance). In classification, the number of possible discrete targets (i.e. the classes) further defines the problem as binary if there are only 2 possible classes, and multi-class if there are more than 2 classes. Examples of supervised models are Linear

Regression for regression problems and Perceptrons for binary classification problems.

In unsupervised learning the source data only contains the input features, with no targets, hence there is no teacher to present the correct answers. According to a quality measure the model learns how to represent the data in order to be able to find patterns that help understanding it. The type of association represented by the quality measure further defines the problem as clustering, if it discovers the inherent groupings in the data, and association, if it discovers the tendencies of change or the rules that relate different samples.

The data used in supervised learning is a set of labelled samples in the form (x, y), where x is a feature vector with n features $[x_1, x_2, x_3, \ldots, x_n]$ and y is the target label of the sample x. After training, the model can be used to predict the target label of a unknown sample. In unsupervised learning, the used data is unlabelled, so it is just the set of samples in the format (x).

Figure 6.1 illustrates two supervised and one unsupervised learning generic models applied to simple abstract problems. In 6.1a the regression of a set of samples with a single feature aims to find a function (the thin green line) that describes the linear distribution of the data by mapping the feature x to the target y. In 6.1b the classification of a set of samples with two features aims to find the line (the thin green line) that separates the two classes of the data (in this case, $y = 1$ is a red square while $y = -1$ is a blue circle). And in 6.1c the clustering of a set of samples with two features aims to find the groupings (the thin green ellipses) that relate the samples. The data used by the problem has no target labels (that is, no y), so the groupings are learned from the patterns in the data according to a given quality measure (in this case, the Euclidean distance between the samples in the features space).



(a) *A regression problem* **(b)** *A classification problem* **(c)** *A clustering problem*

Figure 6.1: *Illustration of different problems handled by supervised and unsupervised learning methods*

The details of each approach are beyond the scope of this work. So the next paragraphs are focused on supervised learning for classification because the problems being analysed in this work are of that type. Then, the two used models are explained.

6.1.1 Classification with a Linear Discriminant

The basis for all classification models comes from the definition of the separation line previously illustrated in figure 6.1b, also called decision boundary. In that example the boundary is a line because the features space is two-dimensional, but it would be a plane if the features space were three-dimensional (i.e. if the samples had 3 features) and so on. Therefore, the decision boundary is a hyperplane: it is a subspace with one dimension less than its ambient space.

The boundary in the that figure is also a line because the illustrated problem is linearly separable. Real world problems might not be linearly separable, cases in which a curve is necessary to describe the decision boundary. The concept of hyperplane is still applicable though, as it will be described further ahead. The initial focus is given on linear binary problems because they make the formalization easier to understand.

The geometrical formulation of a hyperplane is illustrated by figure 6.2 and defined by equation 6.1, in which w is the normal vector to the hyperplane, $w \cdot x$ is the inner product between vectors w and x, $d = \frac{|b|}{\|w\|}$ is the distance from the hyperplane to the origin, with $b \in \mathbb{R}$, and $z = \frac{|f(x)|}{\|w\|}$ is the distance from the sample x to the hyperplane (Lorena and de Carvalho, 2007, p.53;Theodoridis and Koutroumbas, 2008, p.91).

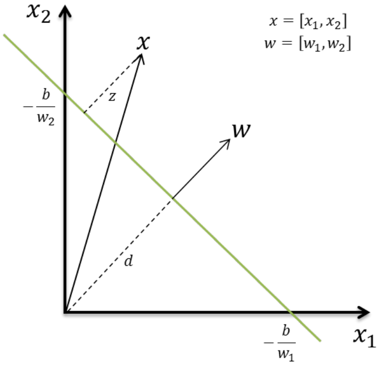


Figure 6.2: *Geometry of the decision boundary*

$$f(x) = w \cdot x + b = 0 \tag{6.1}$$

In one side of the hyperplane, the function $f(x)$ assumes positive values, and on the other negative values. So the process of training a linear binary classifier with the labelled data (x, y) involves finding the values of w and b that characterizes the hyperplane separating all the training samples. Then, with that hyperplane formalized with the correct parameters, the classification of a new sample is easily obtained by doing:

$$g(x) = sgn(f(x)) = \begin{cases} +1 & \text{if} \quad w \cdot x + b > 0 \\ -1 & \text{if} \quad w \cdot x + b < 0 \end{cases} \tag{6.2}$$

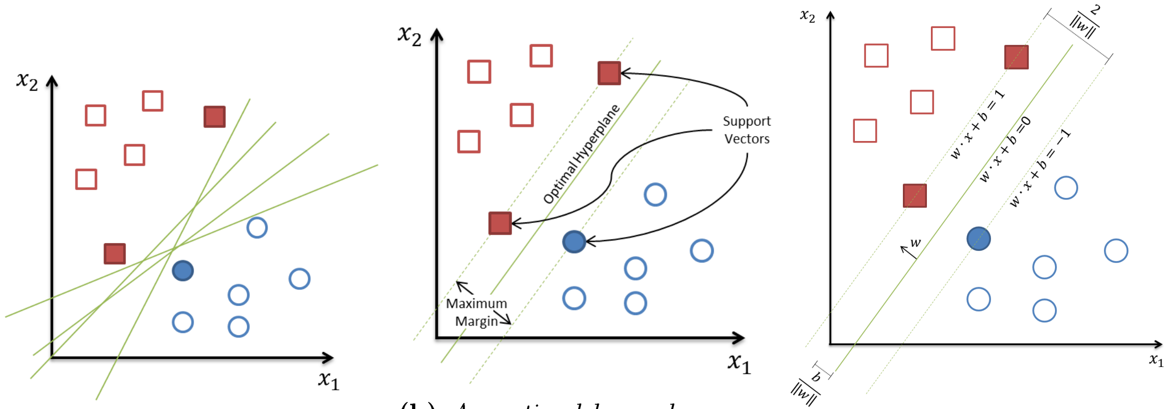
Different methods employ different approaches to find the separation hyperplane. Also, depending on the model the values in w are called weights and the value b is called bias or threshold (sometimes referred to as w_0 instead of b). The classifiers employed in this work are described in the following.

6.1.2 Support Vector Machines

A very popular supervised model used for classification (although it can be adapted for regression) is called Support Vector Machines (SVM) (Cortes and Vapnik, 1995). The principle of a linear binary SVM, the basis of all other SVM variations, is that while there are infinite hyperplanes separating the sample vectors of two classes (illustrated in figure 6.3a), there is an optimal hyperplane that separates the classes by a maximum margin (illustrated in figure 6.3b). This margin is the largest distance between the optimal hyperplane and the nearest samples of each class, which are called support vectors (hence the name of the model).

A SVM model learned from labelled data is represented by the separating hyperplane, the support vectors of each class, and the value of b, as illustrated in figure figure 6.3c. The classification is made by evaluating equation 6.1 for a given new sample and checking the signal of the result. For example, in figure 6.3c, all the red-squared samples have $w \cdot x + b > 0$, while all the blue-circled samples have $w \cdot x + b < 0$.

The maximization of the margin with relation to $w \cdot x + b = 0$, which the training process aims to



(a) *Some of the infinite hyperplanes separating the two classes*

(b) *An optimal hyperplane separates the two classes by a maximum margin, given a set of support vectors*

(c) *The linear binary SVM model, described in terms of the hyperplane w and the set of support vectors of each class*

Figure 6.3: *The linear binary SVM*

find, can be obtained by the minimization of $\|w\|$ with the following optimization problem:

$$\begin{aligned} \underset{w,b}{\text{Minimize}}\ \frac{1}{2}\|w\|^2 \\ \text{subject to } y_i\left(w \cdot x_i + b\right) - 1 \geq 0, \forall i = 1, 2, \ldots, n \end{aligned} \tag{6.3}$$

The restriction is intended to ensure that there are no training samples between the separation margins, resulting in a SVM of rigid margin. However modern implementations, based on the work of Cortes and Vapnik (1995), are soft margin. They use what it is called slack variables (ξ_i) to allow for a few mistakes in the classification, with a penalty imposed by a regularization parameter (C). This allows the classifier to tolerate a few occurrences of noise or outliers. In the soft margin SVM, the minimization problem becomes:

$$\begin{aligned} \underset{w,b}{\text{Minimize}}\ \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \\ \text{subject to } y_i\left(w \cdot x_i + b\right) \geq 1 - \xi_i, \forall i = 1, 2, \ldots, n \end{aligned} \tag{6.4}$$

A non-zero value of ξ_i allows for a sample vector x_i not to meet the margin requirement at a cost proportional to the value of ξ_i (figure 6.4). In this problem there is a treading off on how large the margin can get versus how many points can be placed inside the separation margins. The margin can be less than 1 for a point x_i by setting $\xi_i > 0$, but a penalty of $C\xi_i$ is paid for doing that. The value of C is the regulation parameter that controls the level of over-fitting: small values of C allow for a larger margin with many points in between the separation margins; as the value of C becomes larger, the reduction of the geometric margin becomes less attractive.

Non-linear SVMs

In the case of non-linear problems, the linear SVM is not enough even with the soft margin implementation. The solution is to have the sample data mapped to a higher dimensional space where the separation is presumably easier (figure 6.5 (as long as the transformation performed is non-linear).

If the mapping function is $x \rightarrow \Phi\left(x\right)$, the hyperplane equation becomes $f\left(x\right) = w \cdot \Phi\left(x\right) + b = 0$,

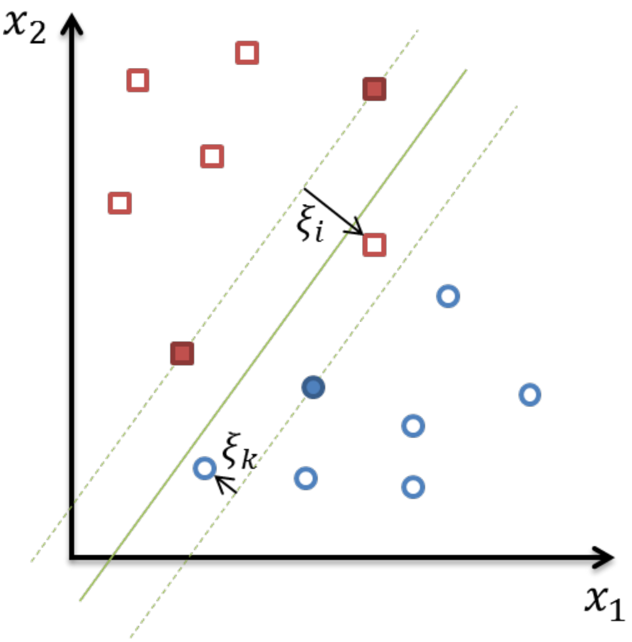


Figure 6.4: *An SVM of soft margin, with sample vectors between the separation margins*

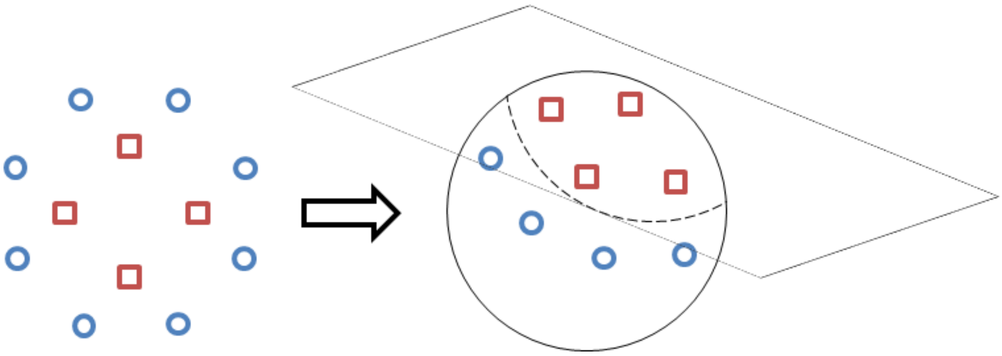


Figure 6.5: *Mapping data to a higher dimension allows constructing a hyperplane to separate classes that are non-linearly separable in the original dimension*

with all rest the same. But the dimension of $\Phi(x)$ can be very large, so in order to keep computation feasible this is achieved by using the so called kernel trick. It has been shown that for same variables α_i (Lorena and de Carvalho, 2007):

$$w = \sum_{i=1}^{m} \alpha_i \Phi(x) \tag{6.5}$$

So, instead of optimizing w it is possible to optimize α and the separating hyperplane function becomes:

$$f(x) = \sum_{i=1}^{m} \alpha_i \Phi(x_i) \cdot \Phi(x_j) + b \tag{6.6}$$

The inner product $\Phi(x_i) \cdot \Phi(x_j) \quad \forall i, j = 1, 2, 3, \ldots, m$ with $i \neq j$ in this new space is called a kernel function. It is essentially a similarity function $K(x_i, x_j)$ computed over each pairs of data points in the mapped space. Common kernels used in the literature are:

- Polynomial: $(\gamma (x_i \cdot x_j) + k)^d$, with parameters γ, k and d. The variable d indicates the degree of the polynomial, so when $d = 1$ it falls back to a linear kernel.

- Gaussian: $\exp\left(-\sigma\|x_i - x_j\|^2\right)$, with parameter σ.
- Radial Basis Function (RBF): $\exp\left(-\gamma\|x_i - x_j\|^2\right)$, with parameter $\gamma = \frac{1}{2\sigma^2}$).

The RBF kernel is particularly popular with SVMs because it commonly outperforms the other kernels with much less data needed for training. Figure 6.6 illustrates the linear and non-linear separations obtained with some of these kernels using the SVM implementation of the Python library Scikit-Learn (called there SVC to stand for Support Vector Classifier)[1].

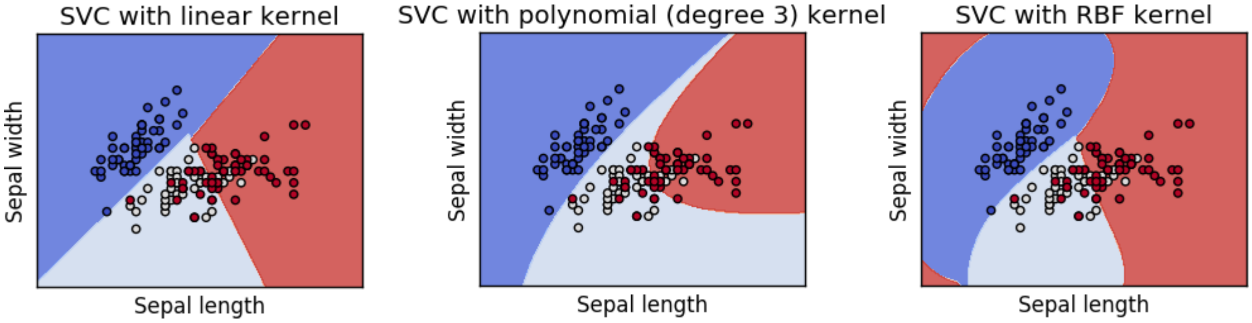


Figure 6.6: *Classifications obtained with different kernels on the 3 classes of the Iris Dataset – image reproduced from the Scikit-Learn documentation*

Multi-class Classification

SVMs are binary classifiers, that is, they can only make class predictions among two different classes according to the side of the hyperplane where a sample is defined. However, multi-class problems can use SVMs by applying one of two strategies: One-versus-Rest (OvR) or One-versus-One (OvO).

The OvR strategy consists of training one SVM classifier per class with the data labelled as belonging and not-belonging to that class. That is, the samples of the class for which the classifier is being produced are labelled with 1, while all other samples (of all other classes) are labelled with -1. During prediction time, the class with the higher score (distance from the hyperplane) is chosen. On the other hand, the OvO strategy consists of creating one SVM classifier per pair of classes, using only the subset of data with the samples of the classes in each pair. During prediction time scores are obtained for each binary combination and the most voted class is chosen.

OvR is commonly preferred due to its computational efficiency: for a problem with n classes OvR requires just n classifiers, while OvO requires $\frac{n\times(n-1)}{2}$ classifiers. However, OvO is useful when the training does not scale well to the volume of data, since in this strategy each individual classifier is trained with only a small subset of the original data.

Estimation of Probabilities

When working with multi-class problems, as it is the case of the classification of the 6 prototypic emotions, it is usually more convenient to have class probabilities instead of a single class prediction. The SVM model alone can only produce single class predictions, but probability estimations can be produced with a process called Platt Scaling (Platt, 1999), that transforms the outputs of the SVM model into a probability distribution over the classes in the problem domain.

The idea behind this process is to train another model with new data created with the same labels but with a single feature (i.e. one dimension): the responses produced by the original SVM model.

[1] http://scikit-learn.org/

Hence, this new model outputs probabilities that serve as confidence levels on the class prediction. In essence, a logistic regression is performed, where the dependent variable is the true class label and the predictor is the SVM score value (also called confidence).

To avoid over-fitting, the regression is calibrated with cross-validation using the k-fold method – in which the data available for training is randomly partitioned into k equal sized blocks of data, so the model is trained against k-1 partitions and then tested and scored against the remaining partition. Therefore, the use of Platt Scaling is an expensive operation for large datasets due to the need of k iterations of training and scoring from the cross-validation. Nonetheless, the probability predictions are commonly very good.

6.1.3 Structured Perceptron

Some classification problems have the additional difficulty that both the input x and the output y are structures: there is an intrinsic organization in the samples of the data that makes it either a sequence, a tree or a graph (Daumé III, 2006, p.16;Zhao, 2014, p.1). That is the case of many natural language processing tasks, for instance, since words and sentences occur in a given order. And, as it has been said at the beginning of this chapter, it is also the case with fun in a game session. Therefore, specific models and algorithms have to be used for this type of problem to consider the sequence of events. The Structured Perceptron (Collins, 2002) is an extension of the standard Perceptron (Rosenblatt, 1957) that has become very popular to handle structured problems due to its simplicity and efficiency.

Perceptron

The Perceptron is an earlier model and one of the most simple linear binary classifiers. It is inspired in the biological neuron, with the dendrites receiving the input signals, the nucleus and cell body applying a threshold to the weighted sum of the input values, and finally with the axon terminals emitting an activation signal. The model is illustrated in figure 6.7, where $x_1, x_2, \ldots, x_n$ are the inputs (the values from the feature vector of a sample) and $w_1, w_2, \ldots, w_n$ are the weights applied to each input value. The weighted inputs are summed together with a bias, from which an activation function F (also called transfer function) emits the Perceptron's output signal (o) in range $[0, 1]$.

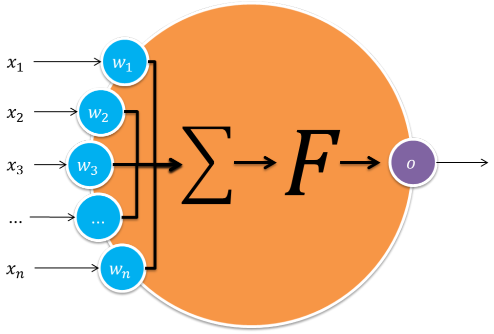


Figure 6.7: *The Perceptron model*

The weighted sum from the input values plus a bias is the same as equation 6.1, that is, the characterization of a hyperplane separating two classes through the inner product between the normal vector w and the features vector x. The bias b only shifts the decision boundary away from the origin, and does not depend on any input value. The activation function F controls how the binary output (i.e. the classification) is produced by checking the side of the hyperplane a sample is located. With a simple step function (the Heaviside Step function) the output is 0 or 1 depending on the signal of the weighted sum plus the bias (called *net*):

$$o = F\left(net\right) = F\left(\sum_{i=1}^{n} w_i x_i + b\right) = \begin{cases} 1 & \text{if} \quad net > 0 \\ 0 & \text{if} \quad net \leq 0 \end{cases} \tag{6.7}$$

It is a common practice to add the number 1 to the feature vector, making it $x = [x_1, x_2, \cdots, x_n, 1]$, and drop the bias b (Zhao, 2014, p.6). This makes the verification simpler, as indicated in equation 6.8, and also allows for the bias to be learnt via a weight w_{n+1}.

$$o = \begin{cases} 1 & \text{if} \quad F\left(\sum_{i=1}^{n} w_i x_i\right) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.8}$$

The training of the Perceptron model is made by iteratively adjusting the weights from a loss-function. That is, the weights are gradually corrected from the labelled samples presented to the model, and the hyperplane is moved towards the training samples. The learning algorithm works as follows (Theodoridis and Koutroumbas, 2008, p.93-94):

1. The weights (including the bias) are randomly initialized with real values, equal to 0 or very small.

2. The feature vector x of a training sample is given as the inputs. The Perceptron calculates the *net* by applying the weighted sum and passing the value through the activation function.

3. The output signal o is compared to the expected value y (the known label for that given training sample) via a loss function $E = o - y$.

4. If $E = 0$, then the algorithm goes back to step 2 to try with another sample; if $E \neq 0$, then the weights are adjusted with the error, proportionally to a learning rate η. This is done by calculating $w_i = w_i + \eta x_i E$.

5. The algorithm goes back to step 2 if an adjustment has been performed or terminates if it converged (i.e. no adjustment was performed for any of the input samples).

The learning rate η is defined in interval $]0, 1]$ and controls overfitting by indicating how much of the error of each sample is applied to the weights per training step. The adjustment of the weights simply moves the hyperplane closer to the sample points as they are seen. It has been proved that if the data is linearly separable, the algorithm converges with finite steps and produce a weight vector w that separates the two classes (Zhao, 2014, p.7). However, the separation hyperplane found might not be optimal. Other variations use the same theoretical basis of the SVM classifier to achieve optimal separation. For example, averaged Perceptron assigns more weights for the examples learnt at the beginning of the training, allowing the Perceptron to achieve some kind of large margin effect (Zhao, 2014, p.8).

The Perceptron model easily generalizes to multiclass problems, in which each feature vector x has a corresponding output y that belongs to a finite set $\mathcal{Y}$ with $m > 2$ classes. A feature function $\Phi\left(x, y\right)$ is used to map the pair (x, y) to a vector of real numbers that represent their relation (for example, with simple counting of occurrences or with statistical measures such as variance). As such the value of $\Phi\left(x, y\right)$ is also a vector in Euclidean space but which depends on the output y. With the basic Perceptron algorithm this new input vector is multiplied with a weight vector w, and the intended response is obtained by choosing the value of y that maximizes the responses for $\Phi\left(x, y\right)$:

$$\hat{y} = \underset{y \in \mathcal{Y}}{\arg\max}\, \Phi\left(x, y\right) \cdot w \tag{6.9}$$

This is equivalent of having m independent error minimizations that are designed to output 1 for vectors belonging to the correspondent class and 0 for all the others (Theodoridis and Koutroumbas, 2008, p.104;Zhao, 2014, p.8), and hence very similar to what is performed with the SVM classifier using the One-versus-All strategy.

The Perceptron model is said to have two layers (the input and the output layers) with a single neuron. The addition of more layers – with neurons interconnected in an input, an output and at least one hidden layers, with weights also in the connections between the neurons – allows the classification of non-linearly separable data. In this case the model is called Multilayer Perceptron (MLP). However, the activation function must be non-linear in some of the neurons. If a multilayer Perceptron has a linear activation function in all neurons, it can be demonstrated with linear algebra that the number of layers can be reduced to the standard two-layer model (Theodoridis and Koutroumbas, 2008, p.103), hence making the model only capable of dealing with linearly separable data. Because of that MLPs usually employ sigmoid functions like the logistic function, in order to still produce outputs in range $[0, 1]$.

The training of a MLP also employs a different algorithm called Back-propagation. It is a generalization of the least mean squares error of the linear Perceptron that similarly uses the prediction error to adjust weights. The inputs are feed-forwarded through the network of neurons until the processing reaches the output layer. Then the error is propagated backwards in the network in order to adjust the connection weights, using the gradient descent method. That is, the weights are moved towards the negative of the partial derivative of the error with respect to each weight in a neuron in order to try to find the local minimum:

$$w_i = w_i - \eta \frac{\partial E}{\partial w_i} \tag{6.10}$$

Generalization to Structured Problems

The structured prediction is similar to a multiclass classification task, with the difference that the label set $\mathcal{Y}$ now represents a set of structured responses that can be generated from the structured input x (Zhao, 2014, p.2). The label set is denoted as a function $\mathcal{Y}(x)$ of the given input, and the prediction with the Perceptron is rewritten as:

$$\hat{y} = \underset{y \in \mathcal{Y}(x)}{\arg\max}\, \Phi(x, y) \cdot w \tag{6.11}$$

However, this is a much more difficult problem because the number of combination of labels in $\mathcal{Y}(x)$ is exponentially large. Thus finding the value of y that maximizes the responses for all possible combinations of sequences of labels in $\mathcal{Y}(x)$ is not tractable in the general case. But for particular problems, if Φ decomposes over the vector representation of $\mathcal{Y}(x)$ such that no feature depends on elements of y that are more than k positions away, the Viterbi algorithm can be used to solve the $\arg\max$ problem in time $O\left(M^k\right)$, where M is the number of possible combinations of labels (Daumé III, 2006, p.18).

The Viterbi algorithm (Viterbi, 1967) uses dynamic programming to find the sequence of outputs y that maximizes the values $\Phi(x, y)$ for the presented structure in x. Suppose that the trained Perceptron is presented with a structure x composed of n sequential samples $x_1, x_2, x_3, \cdots, x_n$. The Viterbi algorithm maintains a table in which the nodes in the a time t of the sequence (what is called a trellis, and it is illustrated in figure 6.8) have the maximum value of the functions Φ for the previous (i.e. time $t - 1$) nodes that lead to the node being calculated. These values are calculated forward in the trellis and a back pointer (in red, in the figure) is kept to allow tracing back the path that maximized for the presented input x. In this case, it is assumed that the observation of a value only depends on the immediately previous value (what is called the 1st Markov Assumption), so $k = 1$.

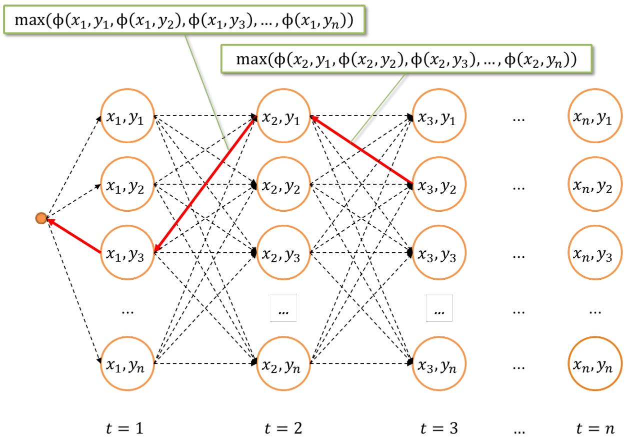


Figure 6.8: *The trellis of interconnected nodes that represent the dynamic programming of the Viterbi algorithm*

6.1.4 Test and Evaluation

The evaluation of a supervised classifier requires the execution of tests with unseen data, so the real accuracy of the classifier can be estimated from the comparison between the predicted and the known class labels (Russell and Norvig, 2010, p.708). Ideally these tests should be performed against a completely different dataset, also containing enough samples representative of the problem domain. But this is not always easy to do because the amount of available data might be insufficient for training and testing the model. A simple approach is to randomly split the dataset into two partitions, so the classifier would be trained with one partition (the training data) and tested with the other partition (the test data). This method, called Holdout cross-validation, has the disadvantages that it fails in using all the data available for training and the results are influenced by the partition employed. For instance, if by chance all the significant samples of a given class are let in the test partition the classifier produced will have a very poor accuracy.

Hence another, more popular, approach is called K-Fold cross-validation. The idea is that each sample will be used twice, both for training and testing. The data is split into k equal partitions, and then k rounds of training and testing are performed, in which $\frac{1}{k}$ of the data is hold for testing and the rest of the data is used for training the classifier. The average score of all tests is expected to yield a better estimate of the overall accuracy of the classifier. Common parameter choices are $k = 5$ or $k = 10$, considered to be statistically accurate while limiting the computational time to a feasible amount. In the extreme of $k = n$ (where n is the number of samples), the approach is called Leve-One-Out cross-validation, meaning that there will be n rounds of training and testing in which the classifier is trained with almost all the samples $(n - 1)$ and verified against a single sample.

Scoring Metrics

The verification of the quality of a classifier tested using any of the methods described above usually employs different empirical measurements (Sokolova *et al.*, 2006, p.1015). The most simple and commonly used one is the accuracy score. This metric simply compares the predicted against the expected class labels, providing an estimation on the number of correct predictions from the classifier in range $[0, 1]$ (with 1 indicating the best value). The accuracy score of a binary classifier is calculated as defined in equation 6.12, in which *tp* is the number of true positives (the test samples correctly predicted as belonging to a class), *tn* is the number of true negatives (the test samples correctly predicted as not belonging to a class), fp is the number of false positives (the test samples incorrectly predicted as belonging to a class) and fn is the number of false negatives (the test samples incorrectly predicted as not belonging to a class).

$$\text{accuracy} = \frac{tp + tn}{tp + fp + fn + tn} \tag{6.12}$$

The problem with this metric is that it assumes equal cost for both kinds of errors (i.e. false positives and false negatives), what might not be desired depending on the problem domain. For instance, in a medical domain in which a classifier is built to predict cancer recurrence, a high accuracy of 90% does not indicate a quality classifier if the 10% of errors include all false negatives (which could lead to disregard of health risks for patients).

Also, when the number of samples for each class are not balanced, the accuracy score suffers from the accuracy paradox. A famous example[2] is the prediction of insurance fraud from a training set of 10,000 samples in which there are only 150 positive cases of fraud. A classifier that would predict $tn = 9,700$, $tp = 100$, $fp = 150$ and $fn = 50$ on that test set would have an accuracy score of 98%. However, due to the large number of non-fraud samples, this accuracy could be *easily improved* to 98.5% by always predicting negative fraud (despite the obvious effect that the classifier would then became completely useless for fraud detection). That is why other metrics – namely the precision and recall scores, originated from the domains of medical trials information retrieval (Sokolova *et al.*, 2006, 1016;Russell and Norvig, 2010, p.869), are employed in conjunction or in place of the accuracy score.

The precision score is calculated as defined in equation 6.13. It measures the number of true positive predictions of the total of positive predictions (i.e. including both true and false positive predictions), in a way that a low precision indicates a high number of false positives. Hence, it can be intuitively understood as the ability of a classifier not to label as positive a sample that is negative.

$$\text{precision} = \frac{tp}{tp + fp} \tag{6.13}$$

On the other hand, the recall score is calculated as defined in equation 6.14. It measures the number of true positive predictions of the total of expected correct answers (i.e. both the true positive and the false negative predictions), in a way that a low recall indicates a high number of false negatives. Hence it can be intuitively understood as the ability of a classifier to find all the positive samples that exist in the data.

$$\text{recall} = \frac{tp}{tp + fn} \tag{6.14}$$

While precision measures the exactness of the classifier (the responses that are actually relevant), recall measures its completeness (the relevant responses that could actually be found). Ideally a classifier with high values of precision and recall is desired, but this is unlikely to be achieved (Russell and Norvig, 2010, p.869). When a classifier is trained the parameters employed help in controlling the amount of fitting performed to the training data. A good classifier will not be completely tuned to that data but instead it will generalize from it in order to achieve good results on unseen data (avoiding the so called overfitting). So there is always a tread-off between exactness and completeness that must be accepted.

These measures consider a binary classifier. They can be calculated per class in a multiclass classifier, although it is more usual to produce an unique score using different average strategies. The micro strategy calculates the score globally by counting the true positives, false negatives and false

[2] Famous enough to be described on Wikipedia: https://en.wikipedia.org/wiki/Accuracy_paradox

positives without regarding the different classes. The macro strategy calculates the metrics for each class – using the OvR (One-versus-Rest) approach – and then takes the unweighed mean of the results to produce the final score. And the weighed strategy calculates the metrics for each class just like in the macro strategy, but uses a weighed mean to account for label imbalance (differences in the number of samples for each class that can occur because of limitations of a dataset used or because of the partition performed by the cross-validation). This work uses the weighted strategy, in order to avoid biases towards the most populated classes.

6.2 Detecting Emotions

As seen in chapter 3, two models have been used with greater success for the classification of prototypical emotions from facial images. Hidden Markov Models (HMM) have been traditionally used for the detection of prototypical emotions, but Support Vector Machines (SVM) are becoming more popular because of their higher accuracy and much lower number of false positives (Bettadapura, 2012, p.21). The HMM model has the advantage of dealing with the temporal dynamics of the facial expressions, which involves the progress changes from onset, apex and offset of the emotions. However, their training requires datasets labelled with structured data and with a large number of samples, which is not easy to find in publicly available datasets (Bettadapura, 2012, p.21). For this same reason, this work also used a SVM for the classification of emotions.

Since the detection of emotions will not consider their temporal dynamics (that is, the classification of emotions will be performed in each video frame independently from the previous frames), a reduced accuracy in emotion detection is expected in frames in between the occurrences of the apex of an emotion. Nonetheless, this is not considered a problem for their use as features in the detection of frustration and fun. Emotions will be estimated in terms of probabilities for each prototypic class and used as features in a structured classification (with the Structured Perceptron model). Thus, the temporal dynamics is still considered, although indirectly.

In order to create and test the emotion classifier, facial images labelled with the prototypic emotions were needed. Then, the SVM model parameters (i.e. the regularization parameter C and the parameters of the kernel used) had to be chosen. Finally the model was evaluated with cross-validation and tested with a different dataset. These steps are described in the following sections.

6.2.1 Emotion-Labelled Datasets

Instead of creating a dataset of labelled images to use for feature extraction and training of a classifier, two publicly available datasets were used: the Extended Cohn-Kanade Dataset (CK+) (Kanade *et al.*, 2000; Lucey *et al.*, 2010) and the 10k US Adult Faces Database (Bainbridge *et al.*, 2013).

CK+ is a very famous facial image dataset, employed by many works in the literature for the detection of prototypical emotions. In a matter of fact, the authors have created it with that purpose in mind. The dataset contains 593 image sequences of frontal faces with both posed and natural facial expressions collected from 123 subjects, and annotated coordinates of 68 facial landmarks detected with the AAM algorithm.

The image sequences in the CK+ dataset vary in duration (10 to 60 frames), starting from a neutral expression and ending with the peek of a prototypic emotion, labelled by experts in the Facial Action Coding System (FACS) (Ekman and Friesen, 1978) – a system used to describe the facial expressions in terms of the contraction and relaxation of facial muscles. Besides the 6 prototypic emotions (anger, disgust, fear, happiness, sadness, surprise) CK+ has also image sequences for contempt, but those samples were not considered. A sample image from the CK+ dataset is reproduced in figure 6.9, with due authorization of the subject.

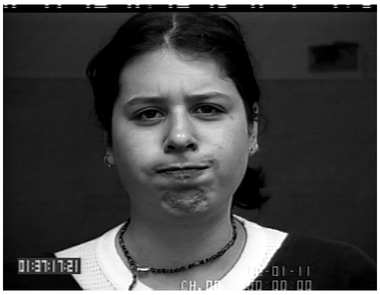


Figure 6.9: *Image from the CK+ dataset with subject S111 expressing anger (©Jeffrey Cohn)*

The 10k dataset was created by psychologists that studied memorability – the degree to which face images are remembered or forgotten. To execute the study the authors collected 10,168 individual faces from the Internet using Google Images. To search for images they used randomly generated names based on the 1990 US Census name distribution. The pictures obtained were then evaluated by five observers in order to remove celebrities, children, low-quality images and faces occluded by objects or with unusual makeup.

During the study the collected images were used in a face memory game in which volunteers should have to verify a target image for repetition in a timed sequence filled with images from other faces. The volunteers also reported on the evaluation of 20 roles of personality, social, and memory-related traits related to the faces in the target images, among which there were the 6 prototypic emotions (i.e. a participant had to judge which emotion the face on the target image was expressing). Besides the prototypic emotions the neutral face was also an option available. From the 10,168 images in the dataset only 2,222 were randomly assigned to be target images, and these are the ones available for training a SVM classifier together with the images from the CK+ dataset.

No image from the 10k dataset can be reproduced in articles or presentations, but the authors have provided a smaller version with other 49 face images that can be used for that purpose. They have been handled with the same procedures, but, differently than the images in the 10k dataset, these 49 images have been manually collected from Creative Commons image resources. Hence, their distribution do not necessarily matches that of the US population. A sample image from this smaller dataset is reproduced in figure 6.10, with due authorization of the subject.

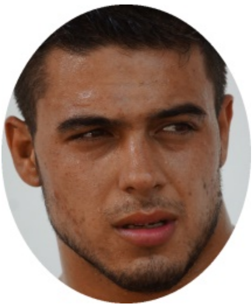


Figure 6.10: *Image from the publication-friendly smaller version of the 10k dataset, with subject 6748122431_3286f0526a_b expressing disgust (©Wilma Bainbridge)*

6.2.2 Building the Multi-class SVM

The prototypic emotions were classified using a multi-class SVM available in the Scikit-Learn library, with the Platt Scaling procedure used to estimate probabilities. The features employed to characterize the training data were the responses of the Gabor bank for each of the 68 facial landmarks detected in a face image from the dataset composed of images from CK+ and 10k. Since the used bank has 32 kernels, each training sample is a vector of 2176 features. The literature only employs dimensionality reduction methods such as PCA when the responses for the whole face are used (case in which the dimension is much larger), or when performance is very important. But since the critical aspect of performance is the convolution required to obtain the Gabor responses, it is very rare in the literature the application of PCA when only the responses at landmark coordinates are used.

The CK+ dataset has sequences of emotions varying from neutral to apex, which could be used for a structured prediction. However, the number of samples is too small. It also has the downside of employing many posed expressions performed by actors. So the images from the 10k dataset, which has more naturally posed samples, were added to compose the final dataset used to train the SVM classifier built.

The SVM model was constructed as follows. From the CK+ dataset, the first and the last images in each sequence were respectively employed as the neutral and the prototypic emotion samples. From the 10k dataset, the 2,222 target images were used as provided. The emotion labels used in these datasets are not the same, so they had to be normalized. Table 6.1 presents the labels used by the datasets. CK+ seems to index the labels in alphabetical order of the emotion names, while 10k uses the same order commonly employed in the literature (that is: happiness, sadness, anger, fear, surprise and disgust). Following the literature convention, the order employed by 10k was kept and the labels from CK+ were converted accordingly.

Table 6.1: *Emotion labels used by the CK+ and the 10k datasets*

Emotion Label	**Dataset**	
	CK+	**10k**
neutral	0	0
anger	1	3
contempt	2	not used
disgust	3	6
fear	4	4
happiness	5	1
sadness	6	2
surprise	7	5

An overview of the samples used is presented in figure 6.11. Both CK+ and 10k datasets have much more samples of neutral and happiness expressions than of the other emotions, what might cause the classifier to bias the predictions towards these classes. Instead of excluding randomly selected samples to make the distribution even, a different approach was preferred to not reduce the number of samples.

If the number of samples for each class is very similar, a problem is called balanced; otherwise, it is called unbalanced. In balanced problems, a fixed value of C (the regulation parameter) is used because the same level of over-fitting is desired for all binary classifiers. In unbalanced problems, the effect of the bias caused by the difference in the number of samples can be reduced by making the amount of over-fitting on the binary classifiers inversely proportional to the frequency of samples for each class. This is achieved by weighting the value of C by a factor $0 < w_i < 1$, calculated as shown in equation 6.15, where $i \in 0, 1, \ldots, m-1$, are the weights for each of the m classes, n is the total number of samples, and $count(y_i)$ is the frequency of each class label y_i.

$$w_i = \frac{n}{m \times count(y_i)} \tag{6.15}$$

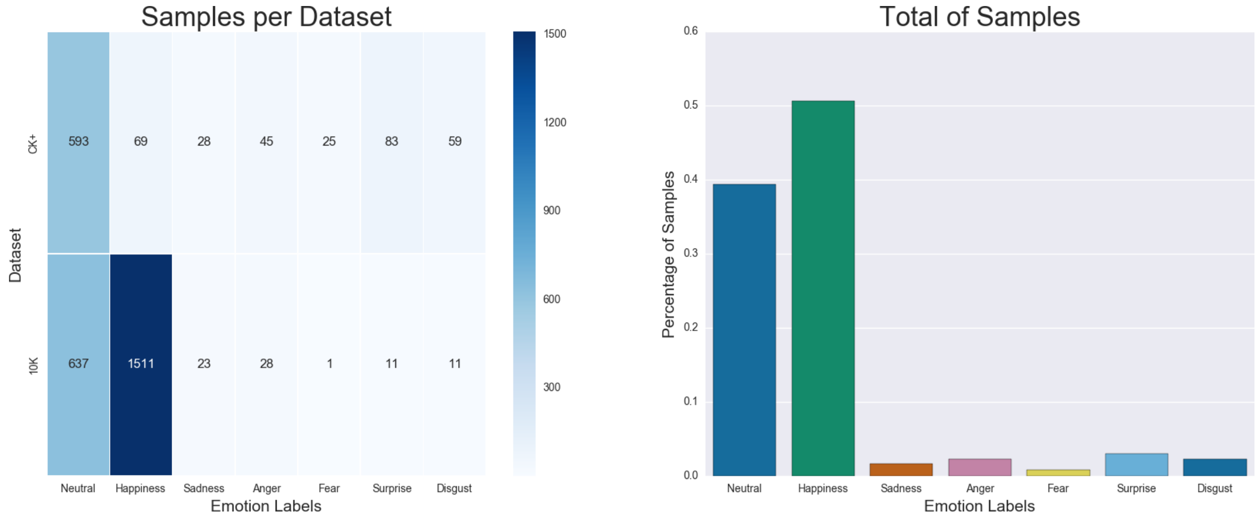


Figure 6.11: *Overview of the emotion-labelled datasets used*

The Scikit-Learn library already implements this balancing method, which was just configured to be used with the employed value of C.

The training images from the datasets were processed to detect and extract the coordinates of the landmarks in the face. The CK+ dataset already provides their own landmark annotations, but those were disregarded in favour of the landmarks detected by the implementation used in this work. Then, the face region was cropped and filtered with each kernel in the Gabor bank used. The magnitude of the real and imaginary parts was calculated to produce a response image with the same size of the face cropped. Finally the Gabor responses at the coordinates of landmarks were collected, normalized to range [-1,1], and used to build the feature vector.

6.2.3 Selection of Parameters

Using the feature vectors of the images from the datasets, an exhaustive grid search was performed in order to find the best parameters for the SVM classifier. This search process simply performs cross-validation on the available data with different classifiers trained from the combination of a list of values provided for each parameter. The combination of parameters that yields the best accuracy are considered the best choices to use.

The linear and RBF kernels were included in the evaluation, and a discrete choice of real values was used for both C and γ: $[0.001, 0.01, 0.1, 1, 1.0, 10.0, 1000.0]$. The parameter C is required by both kernels, while the parameter γ is only required by the RBF kernel. The cross-validation used $k = 5$ folds, and the score was measured with the precision and recall metrics.

The complete list of results obtained is reproduced in B (Grid Search for SVM Parameters). Both the precision and the recall scores indicated that RBF is the best kernel to use, with a value of $\gamma = 0.001$. Regarding the values of C, the precision score indicated $C = 10.0$ as the best choice, while the recall score indicated $C = 1000.0$ as the best choice. The difference of the recall score for values of C in range $[10.0, 1000.0]$ was of only 0.02, hence the value $C = 10.0$ was selected because it yields the best precision score.

With this choice of parameters (RBF kernel, $\gamma = 0.001$ and $C = 10.0$), the Grid Search indicated a precision of 0.78% and a recall of 0.55% on the training data. This means that the produced classifier prioritizes exactness in detriment of completeness. These were judged as good values because the data in which the classifier will be applied to (the videos collected from the participants in the experiment) have a very large number of samples, so it will be more useful for the detection of fun to have more precise classifications of emotion even if they dot not occur all the time.

6.2.4 Evaluation of the Classifier

The multi-class SVM created was evaluated in three steps. The first step consisted of the Grid Search just described. As it helped to decide for the best parameters, it also yielded the precision and recall estimates of the classifier produced.

The second step consisted of a cross-validation performed with 5 folds (i.e. $k = 5$) of the training data (composed of the images from the CK+ and the 10k datasets), in order to obtain the accuracy score. The accuracy obtained was of 0.85, with a 95% confidence interval of ± 0.10 among the 5 tests.

The third step consisted of an independent validation using the trained classifier to predict the emotions on the images from the small dataset of 49 images available from the same authors of the 10k dataset. This smaller dataset was labelled with the same process used for the 10k images, but it contains different, unseen images. The classifier correctly predicted 35 of the 49 samples, resulting in an accuracy of 0.71.

6.3 Detecting Fun

Differently than with the prototypic emotions, there is no publicly available dataset of facial images labelled with levels of fun or with any of its component aspects – at least to the best of the knowledge of this work's author. That is why the levels of frustration, immersion and fun were asked to the subjects during the review of the gameplay: to collect the needed data to train the respective classifiers.

The hypothesis verified in this work is that fun can be inferred from facial images using measures related to challenge, immersion and emotions, in which challenge is indirectly represented in both the levels of frustration and immersion. The level of each affect is measured in the same 5-valued discrete scale used for inquiring the subjects (i.e. the Likert item), a choice intended to make it easier to map the predictions obtained to the responses given by the participants. This means that the level of each affect is represented as a discrete integer value in range $[0, 4]$, in which 0 means not at all frustrated/immersed/having fun, 2 is an intermediate level and 4 means completely frustrated/immersed/having fun.

6.3.1 The Detection of Frustration, Immersion and Fun

As it has been discussed in chapter 2 (What is Fun?), frustration is related to the struggle for achieving difficult but achievable goals, and thus it is something that might be experienced *before* a player achieves flow. It is not a prototypic emotion but a more complex one, as it can relate to sadness or fear, and specially anger when failures are frequent to the point of the person feeling powerless (Klein *et al.*, 2002, p.121;Canossa *et al.*, 2011, p.61). Immersion, on the other hand, is about the absorption of attention that emerges not only from the execution of challenging tasks, but also from the stimulus of the senses and the elicitation of empathy and imagination. Therefore, frustration and immersion were assessed separated in order to try to characterize fun from the aspects of challenge and curiosity/fantasy respectively.

It seems quite clear that an structured approach is required to measure frustration, because a player in flow won't be experiencing frustration at that very moment but instead should have experienced it before, as minor frustrations that might have helped reaching that state. A similar reasoning can be made regarding immersion, since its evolution in time is an important factor to reach flow and hence produce fun. Therefore, structured classifiers were used for their detection.

So the classifiers for frustration and immersion were produced using the Structured Perceptron model implemented in the freely available library seqlearn[3]. This library uses the multiclass approach and the Viterbi algorithm to find the sequence of class labels that maximizes the prediction from the structured input (sequence of features observed in each frame of the videos). As with each of this dependent affect, fun also seems to be produced from a sequence of events, even though it is intuitively understood and described by people as a fleeting and intense affect. Therefore its classifier was also created with the Structured Perceptron model.

Since frustration is a more complex emotion, the features (x) used for training its classifier were the probabilities of the prototypic emotions detected. The classification of immersion, on the other hand, used the gradient of the face distance and the blink rate as features, due to the relation to attention as observed in the literature. The training is supervised, so the answers provided by the subjects for each affect during the gameplay review were used as the class labels (y) for the respective classifier. As it was discussed in chapter 4 (Data Collection), the review scores are statistically significant in comparison with the scores produced from the Game Experience Questionnaire (GEQ), with the exception of immersion for a few subjects. So the review scores were trusted to be correct.

The ideal solution would be to have fun classified from these component affects. However, in order to avoid biases produced from the intrinsic difficulties of each intermediate classifier (frustration and immersion), the fun classifier was trained directly from the review responses self-reported by the subjects for frustration and immersion. This way, an independent analysis on the feasibility of using these affects to detect fun can be performed.

For comparison of prediction gain, the immersion classifier was trained with different combinations of the features: distance gradient and blink rate, distance gradient only, and blink rate only. The fun classifier was also trained with this strategy, using directly the base features combined (prototypical emotions, distance gradient and blink rate) and the levels of frustration and immersion (also combined). In all cases the class labels are the self-reported levels, trusted to be correct.

6.3.2 Data Selection and Preparation

Not all the available data from the video frames was used to train the classifiers. First of all, only the data from subjects 1, 2, 4, 6, 7, 14, 15, 17, 18, 20, 21, 22, 23, 25, 26, 27, 30, 32, 33, 34, 37, 38, 39, 40 and 41 was considered. The other subjects had very poor face detection and landmark tracking, which would certainly encumber the results obtained. Hence they were ruled as unusable – this analysis will be detailed in the first sections of the chapter 7 (Results and Discussion).

Secondly, the frames from the used subjects that still had face detection failures were discarded, since no data extraction would be possible if the face is not detected and tracked.

Lastly, as it was described in chapter 4, the gameplay review was limited to the last 5 minutes of the video, in order to avoid tiring the subjects during the experiment. So the data from the first half of the videos had no labels and was also discarded.

The gameplay review was performed in discrete intervals of 30 seconds, so the missing responses in the frames in between those intervals had to be filled. A linear interpolation based on the existing responses was used, judged as appropriated since that interval is short in terms of the possibilities of action in the games used. The interpolation outputs real values, which were rounded to be transformed into discrete values in the same range $[0, 4]$, as the responses obtained in the review. The real values were rounded down or up to the nearest smaller or bigger integer (for instance, 1.4 rounds to 1 and 1.6 rounds to 2); values exactly halfway between two integers were rounded to the nearest even integer (for instance, 1.5 and 2.5 are both rounded to 2).

[3] https://github.com/larsmans/seqlearn

6.3.3 Validation Procedure

The data used for training the classifiers is sequential and thus can not be easily split. Therefore, instead of using the K-Fold cross-validation method, the Leave-One-Out method was employed considering each subject apart. That is, for each subject s_i for $i \in 0, 1, \cdots, n$, a classifier c_i was trained with the data from all other subjects s_j for $j \in 0, 1, \cdots, n$ and $j \neq i$. The classifier trained was then tested against subject s_i (whose data had not been seen), in order to obtain individual scores of precision and recall. The values were analysed with a box plot, to use the median and Interquartile Range (IQR) to characterize the general quality of the classifier produced.

Chapter 7

Results and Discussion

7.1 Feature Extraction

7.1.1 Face Detection

A general measure of the quality of the face detection is given by the number of frames in which happened a detection failure – the impossibility to find a face, either because the Cascade couldn't find it or because the landmarks model could not be fit on the region located. Table 7.1 presents the failure rates (i.e. the percentage of frames with detection failures) accounted from the collected videos of each subject.

Table 7.1: *Detection failure rates of the collected videos*

Subject	Failures	Subject	Failures	Subject	Failures
1	0.33%	2	0.24%	4	0.54%
6	0.23%	7	0.24%	8	1.01%
9	0.69%	14	0.30%	15	0.23%
16	15.54%	17	0.26%	18	0.33%
19	7.69%	20	0.49%	21	0.71%
22	4.12%	23	0.24%	24	5.63%
25	0.58%	26	0.24%	27	0.28%
28	14.15%	29	0.94%	30	0.24%
31	0.71%	32	1.50%	33	3.73%
34	0.24%	35	0.24%	36	0.24%
37	0.58%	38	0.33%	39	0.29%
40	0.50%	41	0.23%		

The highlighted cells indicate the subjects with failure rate exceeding 1.70%. This value is the statistical upper “fence‘”, calculated from the quartiles of the ordered data as $Q3 + 1.5 \times IQR$, where $Q1$ is the first quartile (equals to 0.244%), $Q3$ is the third quartile (equals to 0.827%), and IQR (the Interquartile Range) is $Q3 - Q1$ (equals to 0.583%). Figure 7.1 represents this information graphically: the horizontal red line is the upper fence, the horizontal blue line is the median (i.e. the second quartile), and the horizontal lighter blue area is the IQR. With that figure it is possible to verify that subjects 16, 19, 22, 24, 28 and 33 are outliers with the worse facial detection quality. The failure rates in the videos of other subjects are mostly inside the IQR with the noticeable exception of subject 32, whose failure rate almost reached the upper fence.

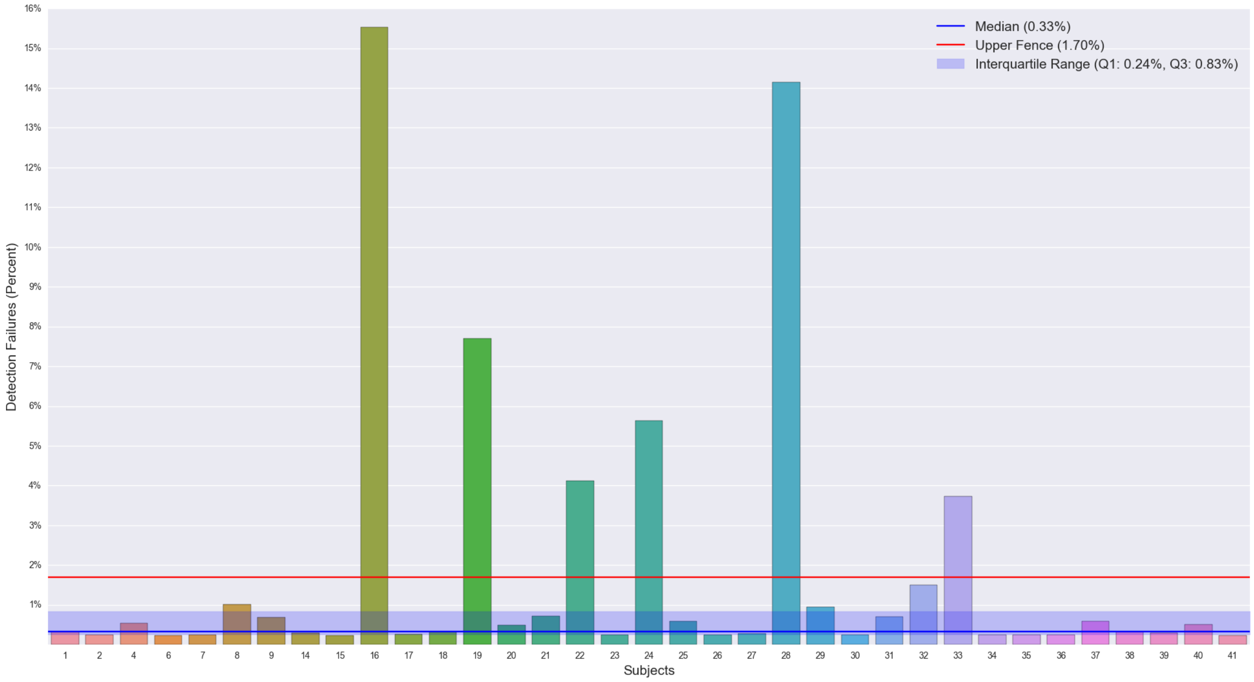


Figure 7.1: *Comparison of the detection failures in the videos of each subject*

A qualitative measure of the detection quality is how "spread" the failures occur in each video. Some failures are expected due to occlusion from the hands, when an individual scratches her nose or adjusts her glasses, for instance. But many spread failures are an indication of general bad quality in face detection. Figure 7.2 presents the detection failures for each subject spread along the time progress. The subjects with low detection failure rates above 1.70% have many failures spread along the video progress. And as expected, subject 32 also presents many spread failures, consistently with the yet high failure rate. All videos had a few failures at the beginning of the recording because of the camera initialization: it took about 1 second, period in which a black screen was captured.

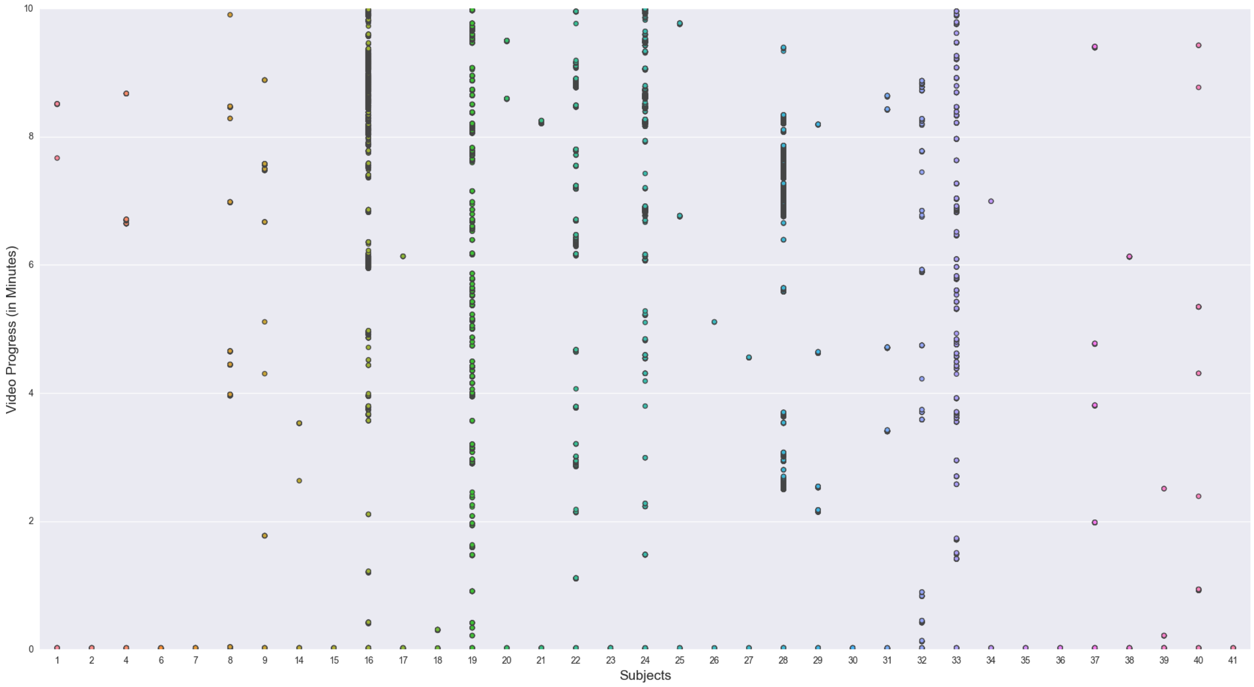


Figure 7.2: *Failures in face detection of the videos collected in the experiment*

A low detection failure rate is not enough to characterize the quality of the face tracking because it does not represent how well the face model fitting found the correct coordinates of the landmarks. Also, a quantitative evaluation on the quality of the face tracking is more difficult to achieve. The interface of the library used does not provide indications on that matter, and an estimation after the fitting is done would be equal to redo the fitting. Therefore, a qualitative analysis was performed based on the visual inspection of the results. Later on, with the estimations that depend on the positioning of the landmarks (i.e. the distance gradient and the blink rate), this qualitative analysis will be improved by a quantitative judgement of the quality of those estimated measurements.

Subject 1

Subject 1 is male, wears a beard and no glasses. His video was recorded at the music school. The face detection was very good, with just a few failures caused by occlusion from the hands. The tracking of landmarks was also very good, as shown in figure 7.3.

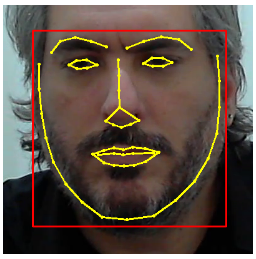


Figure 7.3: *Face detected in frame 9004 (05:00) of the video collected from subject 1*

Subject 2

Subject 2 is male, wears a beard and no glasses. His video was recorded at the music school. The face detection was very good, with no failures at all. The tracking of landmarks was also very good, as shown in figure 7.4.

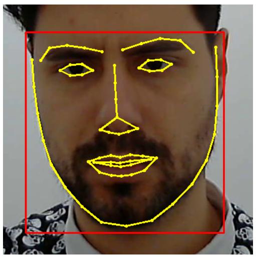


Figure 7.4: *Face detected in frame 10886 (06:02) of the video collected from subject 2*

Subject 4

Subject 4 is female and does not wear glasses. Her video was recorded at the music school. The face detection was good, even though the subject's face was not well framed due to a mistake in adjusting the camera during the set-up. The few failures that happened were because of the bad framing, but also because the session was interrupted (near 06:42) when a teacher entered the room with a child student. The subject turned her head backwards to look at the arriving people and the face detected and tracked the landmarks on the child's face. The tracking of the landmarks was regular, with poor fitting whenever the subject rotated her head, as seen in figure 7.5. This happened because the face was partially outside of the camera field of view.

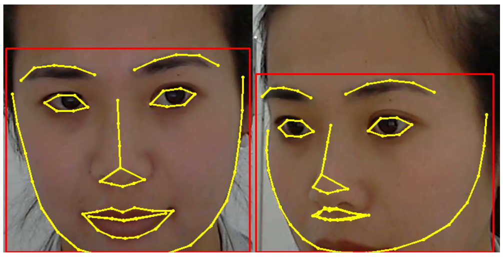


Figure 7.5: *Faces detected in frames 111 (00:03) and 15679 (08:42) of the video collected from subject 4*

Subject 6

Subject 6 is female and does not wear glasses. Her video was recorded at the music school. The face detection was very good, with no failures at all. The tracking of landmarks was also very good, as shown in figure 7.6.

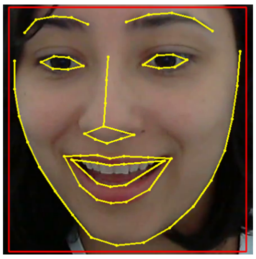


Figure 7.6: *Face detected in frame 13162 (07:18) of the video collected from subject 6*

Subject 7

Subject 7 is female and wears glasses. Her video was recorded at the music school. The face detection was very good, with no failures at all. The tracking of landmarks was also very good, as shown in figure 7.7.

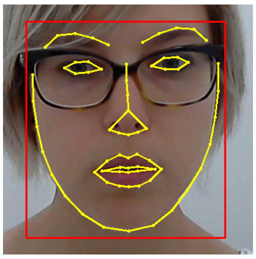


Figure 7.7: *Face detected in frame 1800 (01:00) of the video collected from subject 7*

Subject 8

Subject 8 is male, wears a beard and glasses. His video was recorded at the music school. The face detection was very good, with a few failures caused by the head being lowered or occlusion from the hands. The tracking of landmarks was regular, with difficulties when the head was lowered, as shown in figure 7.8. The quality was also regular because there was much flickering on the eye landmarks, due to the frame of the glasses.

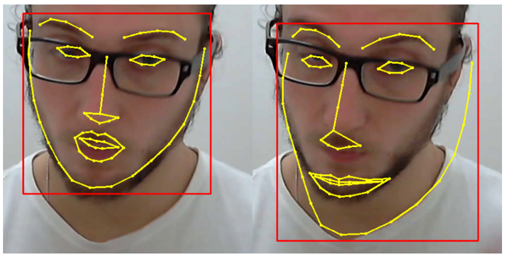


Figure 7.8: *Faces detected in frames 15218 and 15219 (08:27) of the video collected from subject 8*

Subject 9

Subject 9 is male and wears glasses. His video was recorded at the music school. The face detection does not produced many detection fails. However, the tracking of the landmarks had a very poor quality because of reflections in both lenses of the subject's glasses. The lenses were very shiny and reflected perfectly the computer screen, causing the algorithm to produce wrong fittings on a regular basis. The subject didn't authorize the reproduction of his facial images.

Subject 14

Subject 14 is male and does not wear glasses. His video was recorded at the University. The face detection was very good, with a few failures caused by the head being lowered or occlusion from the hands. The tracking of landmarks was good in general, with difficulties only when the head was lowered, as shown in figure 7.9.

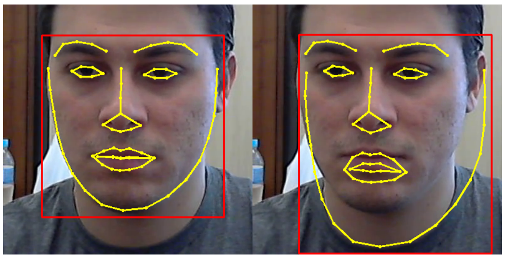


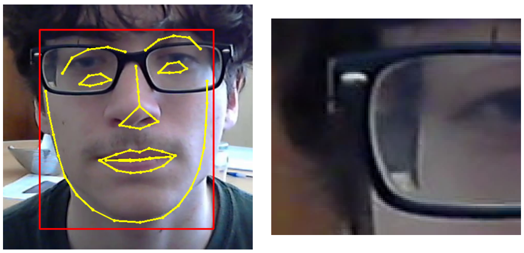
Figure 7.9: *Faces detected in frames 4035 and 4036 (02:14) of the video collected from subject 14*

Subject 15

Subject 15 is female and doesn't wear glasses. Her video was recorded at the University. The face detection was very good, with no failures at all. The tracking of landmarks was also very good. The subject didn't authorize the reproduction of her facial images.

Subject 16

Subject 16 is male and wears glasses. His video was recorded at the University. The face detection had many failures caused by the uneven illumination of his face. During this particular session the sun was very bright outside. The light could not be blocked from the blinds in the room, so it produced strong reflections on the left side of the subject's face. This happened particularly at the end of the gameplay (the last two minutes). The tracking of landmarks was, when it could be achieved, also had a very poor quality. His lenses were thick and caused refractions that induced important mistakes in the fitting, as it can be seen in figure 7.10.



(a) *Face detected in frame 14602 (08:06)*

(b) *Detail of the refraction in the lens*

Figure 7.10: *Images collected from subject 16*

Subject 17

Subject 17 is female and does not wear glasses. Her video was recorded at the University. The face detection was very good, even though the subject's face was not well framed due to the same problem that happened to subject 4. Only one failure happened because of occlusion from the hands. The tracking of the landmarks was regular, with a quality drop in the end of the recording when the subject moved more. The subject didn't authorize the reproduction of her facial images.

Subject 18

Subject 18 is male and wears glasses. His video was recorded at the University. The face detection was very good, with just a failure caused by occlusion from the hands. The tracking of the landmarks was also very good, with no difficulties. The subject didn't authorize the reproduction of his facial images.

Subject 19

Subject 19 is female and wears glasses. Her video was recorded at the University. The face detection was very poor, producing many failures because the subject lowered her head many times while playing the game. This happened while she was checking the game controls on the keyboard. The game assigned to her was the horror game (Kraven Manor), which had actions to interact with objects, run and turn on/off a torch, and she self-reported to usually play games, but only 0–2 hours per week. The subject didn't authorize the reproduction of her facial images.

Subject 20

Subject 20 is male and wears glasses. His video was recorded at the University. The face detection was very good, producing a couple of failures caused by occlusion from the hands. The tracking of landmarks was also very good, as shown in figure 7.11.

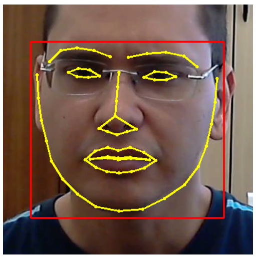


Figure 7.11: *Face detected in frame 10817 (06:00) of the video collected from subject 20*

Subject 21

Subject 21 is male, wears a beard and glasses. His video was recorded at the University. The face detection had just a couple of failures due to occlusion from the hands. The tracking of landmarks was also very good, as shown in figure 7.12.

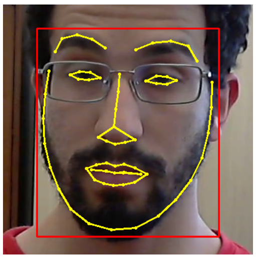
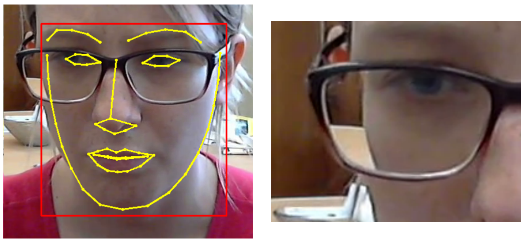


Figure 7.12: *Face detected in frame 1955 (01:05) of the video collected from subject 21*

Subject 22

Subject 22 is female and wears glasses. Her video was recorded at the University. The face detection had many failures because of the refraction of the lenses in her glasses. The glasses refracted the contour of the face in a way that the fitting in that region was always slightly incorrect, and depending on the position of her face would simply prevent the detection. Nonetheless, the tracking of landmarks was regular and worked relatively well when the face was detected. Figure 7.13 illustrates one of this occurrences, with the detail of the refraction of her facial line.



(a) *Face detected in frame 11068 (06:08)*

(b) *Detail of the refraction in the lens*

Figure 7.13: *Images collected from subject 22*

Subject 23

Subject 23 is female and wears glasses. Her video was recorded at the University. The face detection had no failures. The tracking of landmarks was also very good, as it can be seen in figure 7.14.

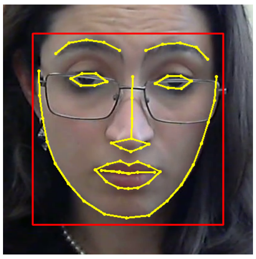


Figure 7.14: *Face detected in frame 3681 (02:02) of the video collected from subject 23*

Subject 24

Subject 24 is male and does not wear glasses. His video was recorded at the University. The face detection had many failures because the subject played with his head lowered most of the time. The tracking of features was also very poor because the folds in the subject's ears were frequently mistook with the outlines of the face laterals and the rim of his shirt collar with the jaw line contour. The subject didn't authorize the reproduction of his facial images.

Subject 25

Subject 25 is male and wears glasses. His video was recorded at the University. The face detection had just a few failures due to occlusion from the hands. The tracking of landmarks was also very good, as it can be seen in figure 7.15.

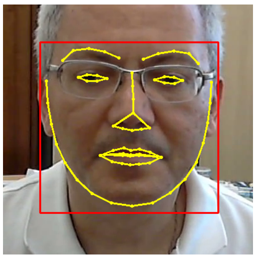


Figure 7.15: *Face detected in frame 7195 (03:59) of the video collected from subject 25*

Subject 26

Subject 26 is female and doesn't wears glasses. Her video was recorded at the University. The face detection had just one failure when the subject lowered her head. The tracking of landmarks was also very good, as shown in figure 7.16.

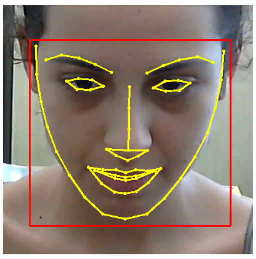


Figure 7.16: *Face detected in frame 9112 of the video collected from subject 26*

Subject 27

Subject 27 is male and wears glasses. His video was recorded at the University. The face detection had only a few failures because of occlusion from the hands. However, the tracking of landmarks was very unstable. While playing the subject kept his head slightly lowered. Also, the frame of his glasses were in front of the eyes, very close to where the eyelids were. Because of that, the fitting had minor variations from one frame to the next even if he didn't move his head, causing the landmarks to "flicker". Figure 7.17 illustrates two frames in which this flickering of the detected landmarks happened. The landmarks of the eyelid of the right eye detected in frame 6142 move up to the frame of the glasses in frame 6143. This was a problem particularly for the detection of blinks (discussed later in this chapter) because of the unexpectedly quick movement of the eye features.

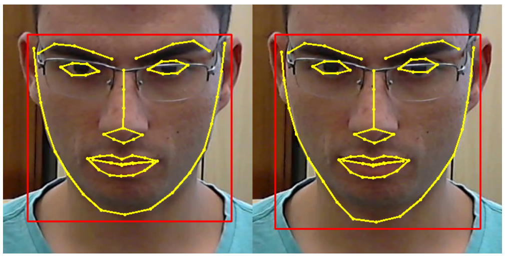


Figure 7.17: *Faces detected in frames 6142 and 6143 (03:24) of the video collected from subject 27*

Subject 28

Subject 28 is male and does not wear glasses. His video was recorded at the University. The face detection had many failures because the subject moved frequently, getting very close to the camera to the point of having his face partially outside of the field of view for long periods. Despite that, the tracking of landmarks was not bad when the face was detected. The poorest results only happened when the subject moved outside of the field of view, as seen in figure 7.18).

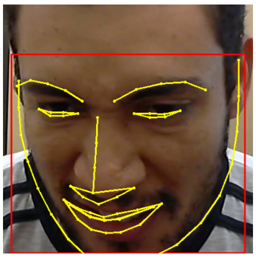


Figure 7.18: *Face detected in frame 12262 (06:48) of the video collected from subject 28*

Subject 29

Subject 29 is female and wears glasses. Her video was recorded at the University. The face detection had a few failures caused by occlusion from the hands and once by moving the face outside of the field of view (when the subject got scared by the horror game she played). However, the tracking of landmarks had regular results. When there was no occlusions a fitting was always achieved, but the eye and eyebrow landmarks were frequently misplaced, because the subject was wearing an eyeliner makeup. The frame of the glasses was easily mistook with the eyelids and the eyeliner with the eyebrows, causing the landmark positions to flick from one frame into the next. An example of this poor fitting can be seen in figure 7.19).

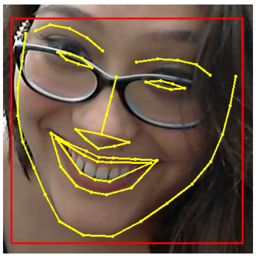


Figure 7.19: *Face detected in frame 4236 (02:21) of the video collected from subject 29*

Subject 30

Subject 30 is female and doesn't wears glasses. Her video was recorded at the University. The face detection was very good with no failures at all. The tracking of landmarks was also very good, as shown in figure 7.20.

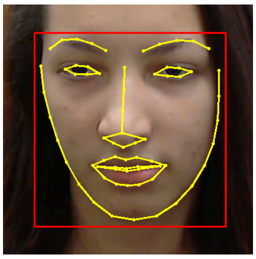


Figure 7.20: *Face detected in frame 1864 (01:02) of the video collected from subject 30*

Subject 31

Subject 31 is male, wears a beard but no glasses. His video was recorded at the University. The face detection was very good with just a few failures caused by occlusion from the hands. However, the tracking of landmarks had very low quality. The landmarks of the jaw line were frequently mistook with the rim of the shirt collar, even when the face was aligned with the camera, as it can be seen in figure 7.21.

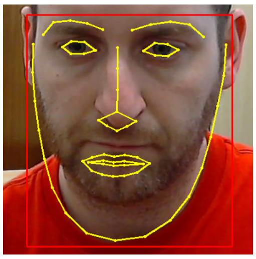


Figure 7.21: *Face detected in frame 8662 (04:48) of the video collected from subject 31*

Subject 32

Subject 32 is male, wears a beard and glasses. His video was recorded at the University. The face detection had many failures and the cause was similar to what happened with subject 19: he frequently lowered his head to look at the keyboard to check the controls. He also played Kraven Manor and reported to usually play games only 0–2 hours per week. The tracking of landmarks, however, was reasonably good when the detection was achieved. There was some flickering on the eye landmarks because of the very thin frame of his glasses, but overall the fitting was very good, as seen in figure 7.22.

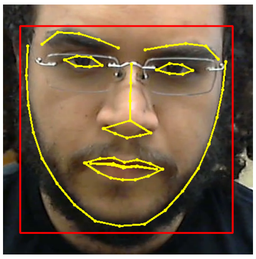


Figure 7.22: *Face detected in frame 15206 (08:26) of the video collected from subject 32*

Subject 33

Subject 33 is female and doesn't wear glasses. Her video was recorded at the University. The face detection had many failures and the cause was similar to what happened with subjects 19 and 32: she frequently lowered her head to look at the keyboard to check the controls. The game played was the Melter Man, which does not have many control options as Kraven Manor. But the subject self-reported to not usually play games, what might explain her difficulties with the game controls. But differently from the other subjects, when a face was detected the tracking of the landmarks had very good results, as shown in figure 7.23.

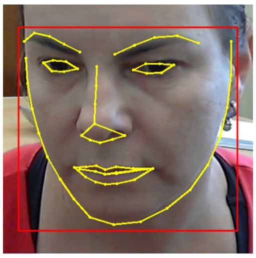


Figure 7.23: *Face detected in frame 14467 (08:02) of the video collected from subject 33*

Subject 34

Subject 34 is female and doesn't wear glasses. Her video was recorded at the University. The face detection was very good, with just a single failure that happened when she lowered her head too much. The tracking of landmarks was also very good, as shown in figure 7.24.

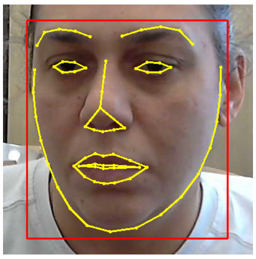


Figure 7.24: *Face detected in frame 16542 (09:11) of the video collected from subject 34*

Subject 35

Subject 35 is male, wears a beard but no glasses. His video was recorded at the University. The face detection was very good with no failures at all. However, the tracking of landmarks had very low quality for a similar reason as with subject 31. The landmarks of the jaw line were frequently pulled down towards a shadow region near his chest area (the shirt collar was distance in this case, but the subject kept his head slightly lower than subject 31). The subject didn't authorize the reproduction of his facial images.

Subject 36

Subject 36 is female and wears glasses. Her video was recorded at the University. The face detection was very good with no failures at all. The tracking of landmarks had regular results. The fitting worked well in general, as shown in figure 7.25, but the eye features were frequently mistaken with the frame of the glasses, causing flickering in that region.

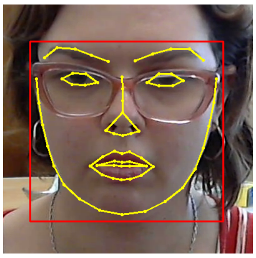


Figure 7.25: *Face detected in frame 1956 (01:05) of the video collected from subject 36*

Subject 37

Subject 37 is male, wears a beard and glasses. His video was recorded at the University. The face detection was very good, with a few failures caused by occlusion from the hands. However, the tracking of landmarks had regular results because there was much flickering in the landmarks of the jaw line and the eyes, as shown in figure 7.26. The beard, the frame of the glasses and specially the thick lenses caused these variations, which didn't prevent the detection as it did with subject 22 because subject 37 moved much less while playing.

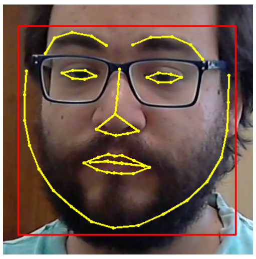


Figure 7.26: *Face detected in frame 7148 (03:58) of the video collected from subject 37*

Subject 38

Subject 38 is female and doesn't wear glasses. Her video was recorded at the University. The face detection was very good, with a single failure caused by occlusion from the hands. The tracking of landmarks was also very good, as shown in figure 7.27.

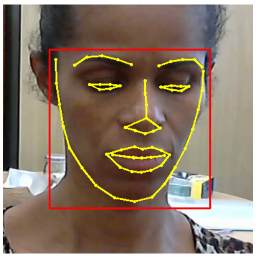


Figure 7.27: *Face detected in frame 10793 (05:59) of the video collected from subject 38*

Subject 39

Subject 39 is female and doesn't wear glasses. Her video was recorded at the University. The face detection was very good, with just two failures caused by lowering of the head (which the subject didn't do much). The tracking of landmarks was also very good. The subject didn't authorize the reproduction of her facial images.

Subject 40

Subject 40 is female and wears glasses. Her video was recorded at the University. The face detection was very good, with a few failures caused by occlusion from the hands. The tracking of landmarks was also very good, as shown in figure 7.28.

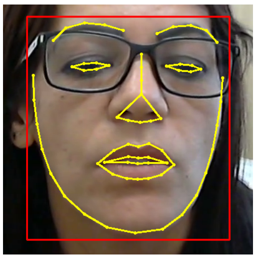


Figure 7.28: *Face detected in frame 7838 (04:21) of the video collected from subject 40*

Subject 41

Subject 41 is female and doesn't wear glasses. Her video was recorded at the University. The face detection was very good, with no failures at all. The tracking of landmarks was also very good, as shown in figure 7.29.

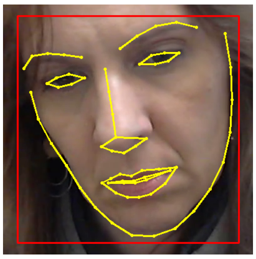


Figure 7.29: *Face detected in frame 3052 (01:41) of the video collected from subject 41*

Overview of the Tracking Quality

Based on the previous quality analysis, but also on the analyses described in the next sections, the videos collected from the subjects have been separated in two categories regarding their judged utility for the assessment of fun:

1. **Unusable**: The videos of subjects 8, 9, 16, 19, 24, 28, 29, 31, 35 and 36 were considered unusable because of the detection failures, but particularly because the very low quality of the tracking of landmarks. The videos of subjects 8, 29 and 36 were included in this group because of the flickering that happened in the eye landmarks due to the glasses. The low quality of their tracking is confirmed by the analysis of results from the blink detection.
2. **Usable**: The videos from all other subjects were considered usable. The videos of subjects 4, 22, 27, 32, 33 and 37 had detection failures and some tracking difficulties (in which the flickering of the eye landmarks are the most relevant ones), but they had many frames in which good tracking results were obtained. The videos of subjects 1, 2, 6, 7, 14, 15, 17, 18, 20, 21, 23, 25, 26, 30, 34, 38, 39, 40 and 41 presented the best detection and tracking qualities.

The data from all subjects was used for the detection of emotions and the estimation of the distance gradient and blink rate, but only the data from the usable subjects (25 out of the 35 subjects, that is, 71% of the collected data) was used in the detection of frustration, immersion and fun: 1, 2, 4, 6, 7, 14, 15, 17, 18, 20, 21, 22, 23, 25, 26, 27, 30, 32, 33, 34, 37, 38, 39, 40 and 41. The use of all data for the estimations of the basic features was required to help the analysis for separating the data that was good to use for the training of the classifiers.

7.1.2 Gradient of the Face Distance

Distance Estimation

The estimated distances between the face and the camera extracted from the collected data of all subjects are presented in figure 7.30. Each graph belongs to a subject whose number is placed above it. The video progress is in minutes and the distances are in centimetres. The estimation of the distance depends on the estimation of the pose of the subject's face, which in turn depends on the quality of the facial detection and tracking of landmarks. So whenever a detection failure occurs, the distance estimation was set to 0. To avoid polluting the graphs with the drops to 0, the last estimated distance value was used in the cases of failure.

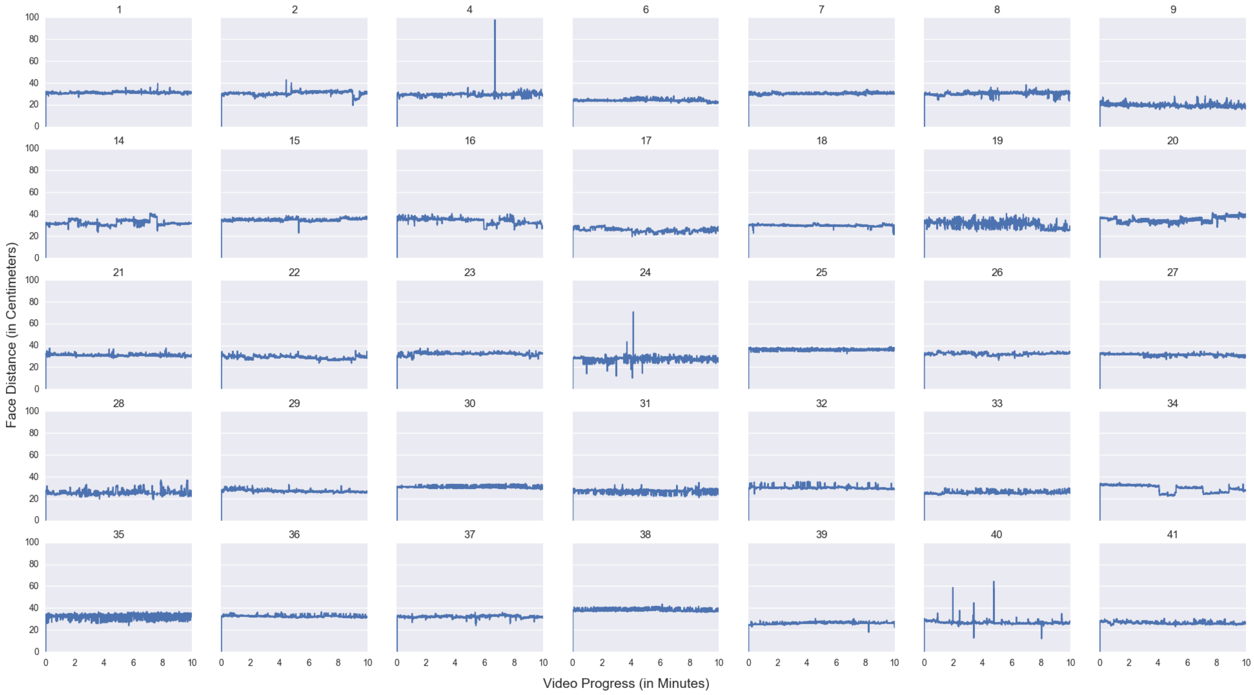


Figure 7.30: *Facial distances estimated from the videos of the subjects*

As expected, subjects 19, 24, 28, 31 and 35, previously described to be unusable, produced much noise due to the lower quality in the tracking of the landmarks. From the other subjects, a few positive peaks (with a rapid increase of the distance) can be observed in the graphs of subjects 4 and 40.

The very high positive peak in the video of subject 4 was caused by the gameplay interruption, when the software detected and tracked the landmarks of the face of the child that entered the room. Because of that the estimation jumped from about 28 to about 96 centimetres. Some noise, particularly at the end, are due to the incorrect framing of the face on the video. Subject 40 had very good results both in face detection and landmarks tracking, so the few high peaks in the distance estimation were not expected. Figure 7.31 presents the two frames in which one of these peaks occurred.

The reason for the bad estimation is related to the choice of the 3D face model and the facial expression displayed. The 3D face model, previously illustrated in figure 5.6, characterized the face with just five points, including the landmarks of the corners of the mouth. The facial expression displayed by the subject, which involved tighing and twisting the lips, caused a horizontal compression of the mouth landmarks on the x axis of the image plane. The proximity of these two bottom points, while the two top points (from the landmarks of the corner of the eyes) were kept apart, caused the pose estimation to infer that the face was in a negative slop – that is, bent away from the camera in the pitch axis. The distance is measured from the tip of the nose, so the estimated value doubled from the previous frame (going from 28.22 to 57.61 centimetres). The exactly same scenario happened again in frames 6141 (03:24) and 8592 (04:06)of the video of this subject, with the display of the same labial compression. The few negative peeks on this subject were also caused by similar deformations produced from partial occlusion of the mouth from the hands.

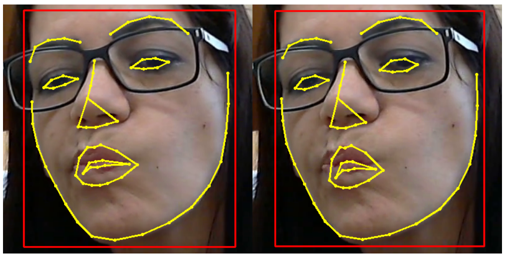


Figure 7.31: *Faces detected in frames 3567 and 3568 (01:58) of the video collected from subject 40*

In general, even when the tracking quality was poor the variations in the distance estimation were consistent. This is the case of subject 9, for instance. This subject is the one with the poorest fitting because of the reflections on the lenses of his glasses. Nonetheless, the deformations seemed to occur in both axis of the image plane, not producing much variation in depth.

For the cases in which the detection and the tracking were very good, the estimation results also had high quality. For instance, subject 34 had larger movement variations that were accurately captured by the distance estimation. In the graph of this subject there is a big valley starting right after minute 4. Figure 7.32 illustrates this moment in a two-seconds short clip, showing that the subject indeed moved much closer to the camera.

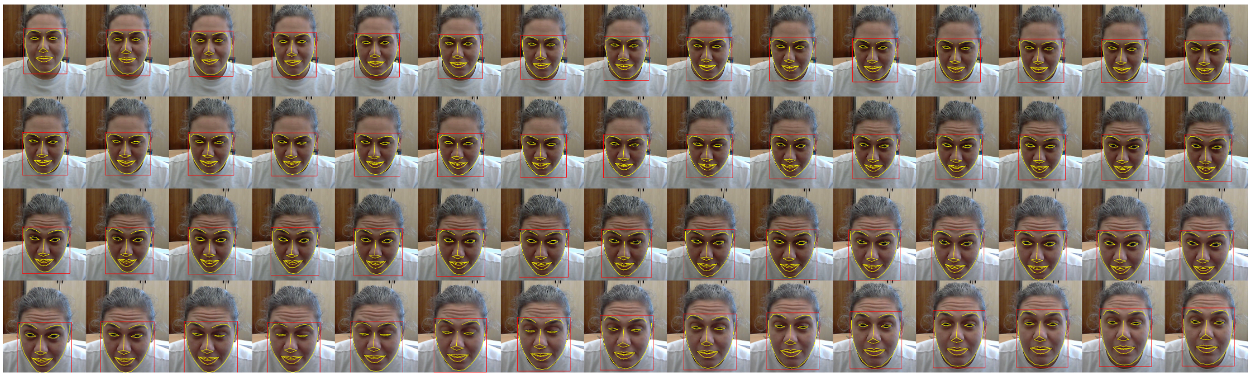


Figure 7.32: *A short clip of 60 frames captured from frame 7290 (04:03) to frame 7350 (04:05) of the video collected from subject 34*

Calculus of the Gradient

The gradients calculated from the estimated distances are illustrated in figure 7.33, with the y axis presenting the value of the distance gradient in centimetres. The missing values from the detection failures (when the distance estimation dropped to 0) were handled by reusing the last valid gradient previously calculated. This was done to improve the classifier when the data is used, assuming that the subject would have not moved in the interval otherwise the detection would possible have returned to work.

As it can be seen, some subjects had many noisy responses consistently produced across the video progress, with the gradient values getting above 3 or 4 centimetres in some cases. Consistent noise values during all the video progress is expected for subjects with a poor tracking quality, and the estimations presented corroborate the qualitative analysis that previously indicated subjects 9, 16, 19, 24, 28, 31 and 35 as unusable. The graphs of subjects 22 and 37 have some noise, but not with the same degree of the others. This also corroborates the analysis that they are useful despite a slightly lower quality in tracking than general.

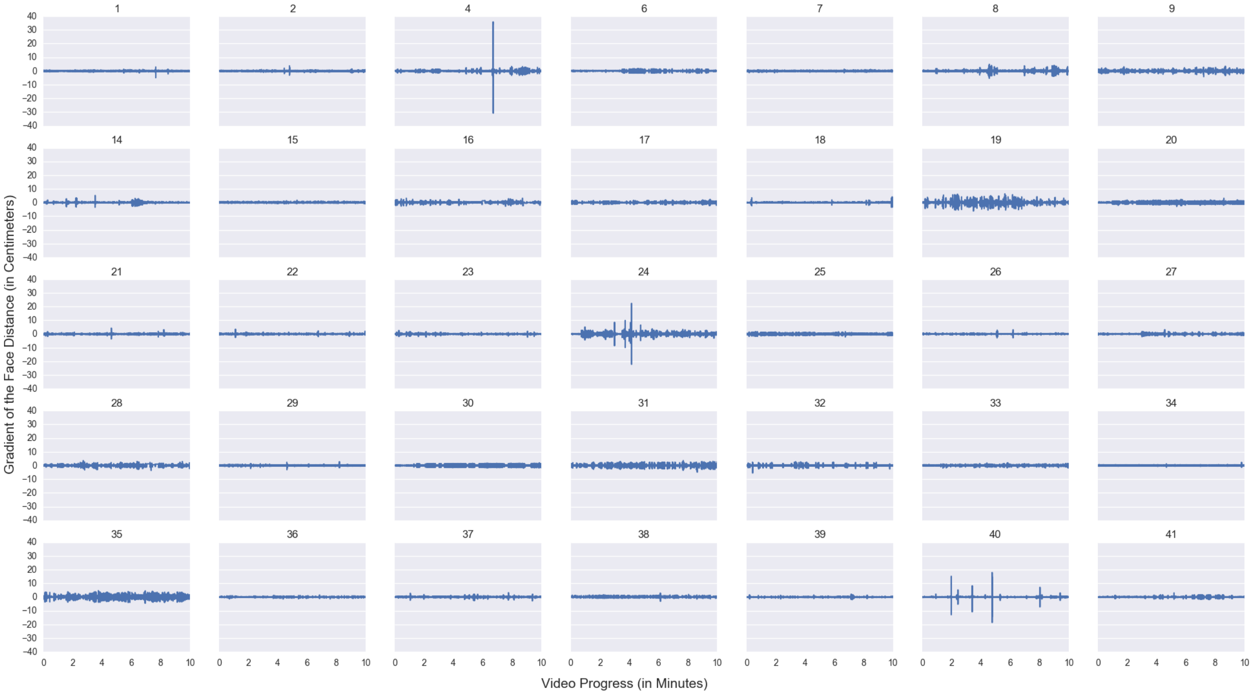


Figure 7.33: *Gradients of the facial distances of the subjects*

7.1.3 Blink Rate

The blink rate, measured in blinks per minute, was obtained by counting the number of detected blinks in the last minute of the video progress. In other words, the blinks are counted inside an one-minute time window calculated at each frame and going back 1800 frames (because the recording of the videos used a frame rate of 30 frames per second). Figure 7.34 presents the values produced from the videos of each subject, as well as the accumulated count of blinks (only needed to aid in the analysis of results). The green line is the estimated frame rate at a given time and the blue line is the total number of blinks accumulated until a given time.

As with the estimation of distance, the blink detection greatly depends on the quality of the tracking, particularly the tracking of the eye landmarks. The mean blink rate in adult humans varies from 2 to 50 blinks per minute, depending on the behavioural state (Miller, 1980). Therefore, estimations that are much above this mean range should be analysed as possibly wrong because of a tracking with poor quality.

Subjects 8, 9, 16, 19, 29 and 36 produced many estimations way above 100 blinks per minute, what it is another strong indication of the poor quality of their tracking (in this case, particularly regarding the region of the eye landmarks). The poor estimations of subjects 8, 29 and 36 confirmed that the flickering was worse than previously evaluated, so they were defined as unusable.

Subject 27 had a few estimations above 100 blinks per minute because of the unstable tracking of landmarks, but that situation didn't occur during the whole video, as it is also confirmed by the small slope of the accumulated blink count. Subjects 4, 32, 37 and 38 had large but localized

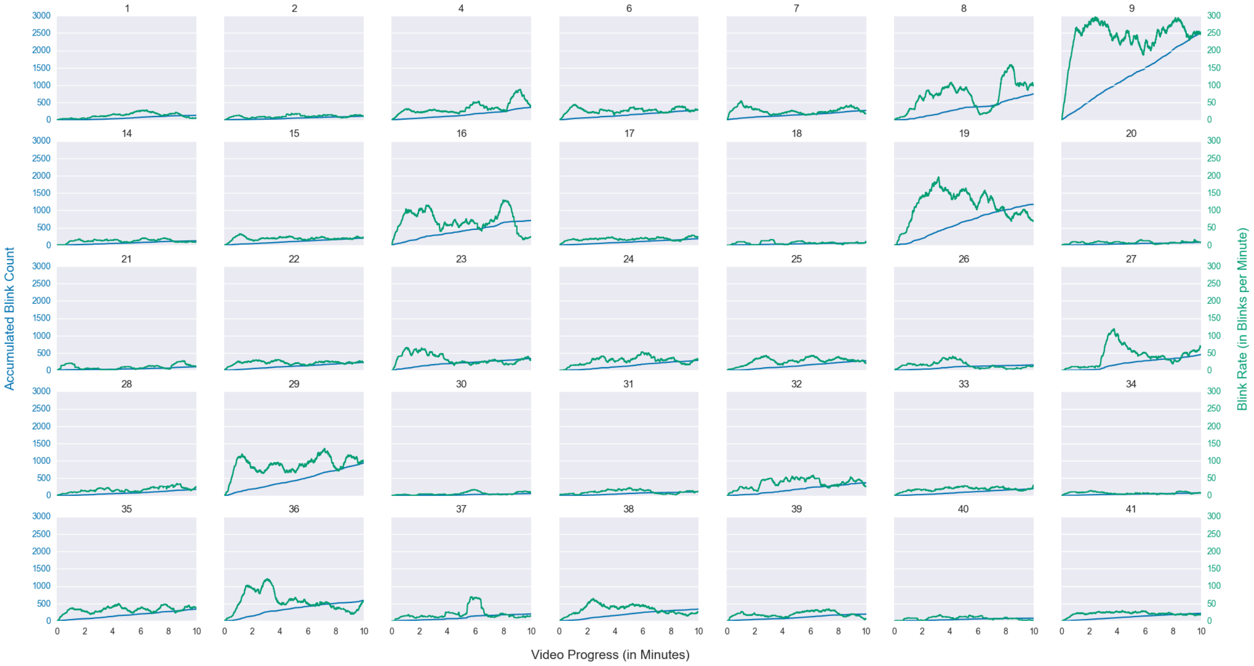


Figure 7.34: *Accumulated count of the detected blinks and blink rate from the subject videos*

estimations between 50 and 100 blinks per minute, but their better quality in general blink rate detection is also indicated by the smaller slope of their accumulated blink count.

It is important to notice that other subjects that have been defined as unusable did not produced poor estimations of blink rate. That was the case of subjects 24, 28 and 31, indicating that there wasn't much flickering of the eye landmarks in these subjects. However, their videos can not be moved into the usable groups because the tracking of landmarks still had significant difficulties that not only produced worse estimations of the distance gradient but will also affect the extraction of Gabor features – in the videos of subjects 8 and 31 the landmarks were frequently moved to the rim of the shirt collar, and in the video of subject 28 the face is frequently outside of the camera field of view and hence not completely captured).

7.1.4 Prototypic Emotions

Neutral Expression

Figures 7.35 presents the probabilities of the neutral expression detected on the videos of each subject (whose number is indicated above each graph). The x axis is the video progress (in minutes) and the y axis is the probability of the neutral expression in range $[0, 1]$. As it can be observed, the probability of the neutral expression is high most of the time in all of the videos. Subjects 1 and 30 were the participants that essentially displayed neutral expressions during all their gameplay sessions. This can be verified by a visual inspection of the videos: those were the subjects with the most serene faces, displaying more subtle expressions that changed very little during the whole session.

Other subjects presented a greater degree of variation of the neutral expression, though its probability is often above 70%. This is expected as the display of the prototypic emotions should occur as responses to meaningful situations in the gameplay instead of being maintained during the whole session. Also, since the classifier was trained to perform predictions on a frame basis (i.e. without considering the structured evolution from neutral to apex of an emotion), the detection of the prototypic emotions should be much more punctual than the neutral expression. Nonetheless, the larger number of samples of the neutral class in the datasets used for training the classifier might also have influenced this result.

Noticeable drops in the probability of the neutral expression are observed in subjects 2, 15, 16, 17, 18, 25, 28, 34, 36, 38, 39 and 41, caused by the detection of other emotions with higher probabilities. But even if there is no visible margin for other emotions, this does not mean that the neutral was the only expression detected. The graph of subject 30, for instance, seems completely taken by the neutral expression but this is a limitation of the graphic view because of the large amount of fine grained data. As it will be seen in the following, there were also other emotions detected in the video of subject 30, as well as in the other subjects.

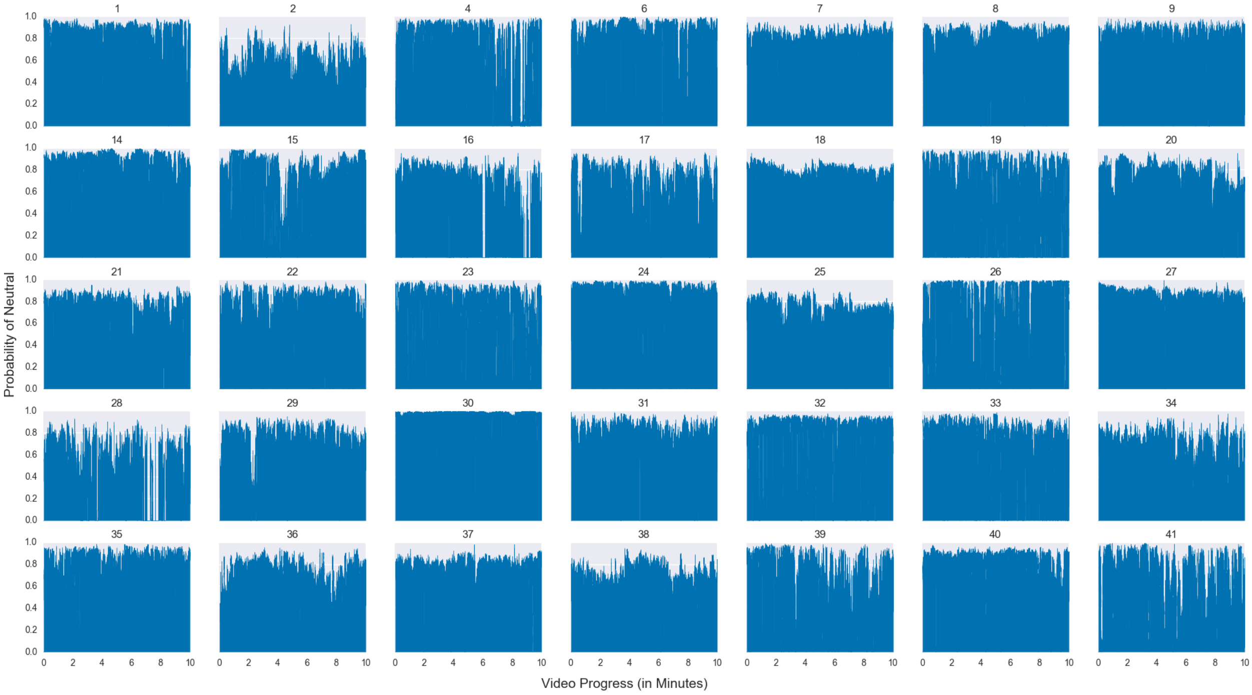


Figure 7.35: *Probabilities of the neutral face in the videos of the subjects*

Happiness

Figure 7.36 presents the probabilities of happiness detected on the videos of each subject. Differently than what happened with the neutral expression, happiness had larger variations. The higher probabilities of happiness occur when a subject tights the eyes and stretches the mouth, with peaks of more than 90% happening when there is a broad smile, as illustrated in figure 7.37c from a facial expression of subject 41. But even subtle smiles caused a higher detection of happiness. For instance, facial expression from subject 30 illustrated in figure 7.37a was classified as happiness, even though this expression may better indicate contempt (an emotion with labelled images available in the CK+ dataset, but which was not used in this work). In a matter of fact, the importance of the mouth region for the detection of the emotions is clearly observable in the results, as false positives were produced when the mouth was occluded. For instance, the facial expression of subject 40 illustrated in figure 7.37b was wrongly classified as happiness because of she covered her mouth with her fingers.

Sadness

Figure 7.38 presents the probabilities of sadness detected on the videos of each subject. Even though it is impossible to tell if a subject was really feeling sad (without having asked her), the indications of this emotion seem consistent with facial expressions displaying loose eyelids and closed mouth, as

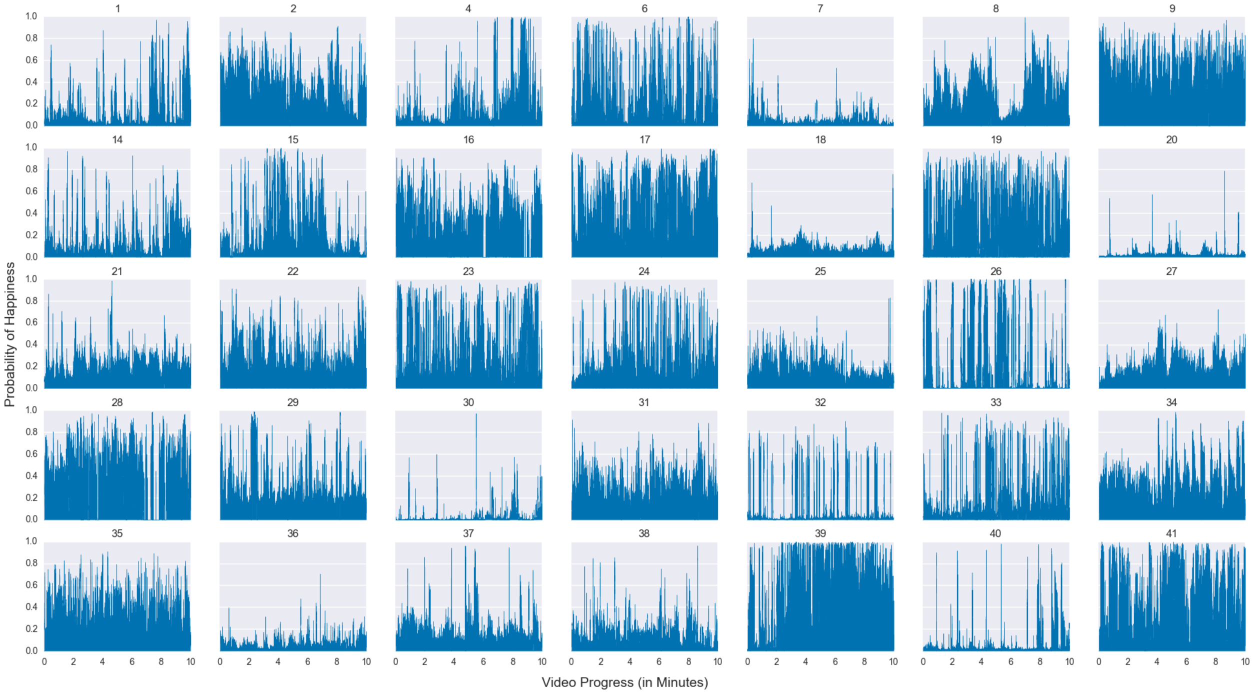
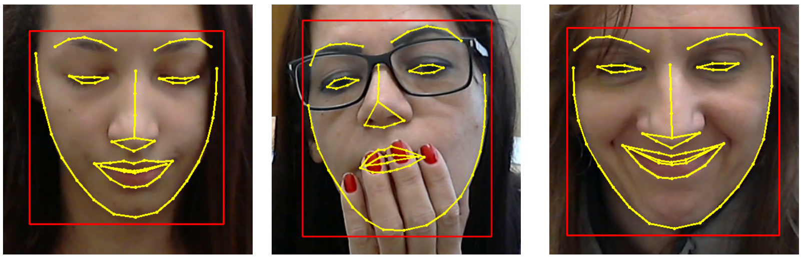


Figure 7.36: *Probabilities of happiness in the videos of the subjects*



(a) *Facial expression from subject 30* **(b)** *Facial expression from subject 40* **(c)** *Facial expression from subject 41*

Figure 7.37: *Examples of facial expressions classified as happiness*

illustrated in figure 7.41. The detection of this emotion was very robust, sometimes having a similar probability of the neutral expression (figure 7.39a, with a probability of 34% for both the neutral and the sadness expressions) and mainly producing false positives when the landmark tracking failed (figure 7.39b).

Anger

Figure 7.40 presents the probabilities of anger detected on the videos of each subject. This emotion was less detected than the others, but it also seemed very consistent with tightened lips or eyebrows pulled in (as illustrated in figures 7.41b and 7.41c). But as with happiness, the detection of anger was sensitive to occlusion of the mouth (as illustrated in figure 7.41a, with a false positive due to partial occlusion of the lips).

Fear

Figure 7.42 presents the probabilities of fear detected on the videos of each subject. This is one of the least detected emotions, what is expected since fear should be elicited mainly by the horror game (Kraven Manor). The subjects with the highest detections are subjects 4 and 41, both which

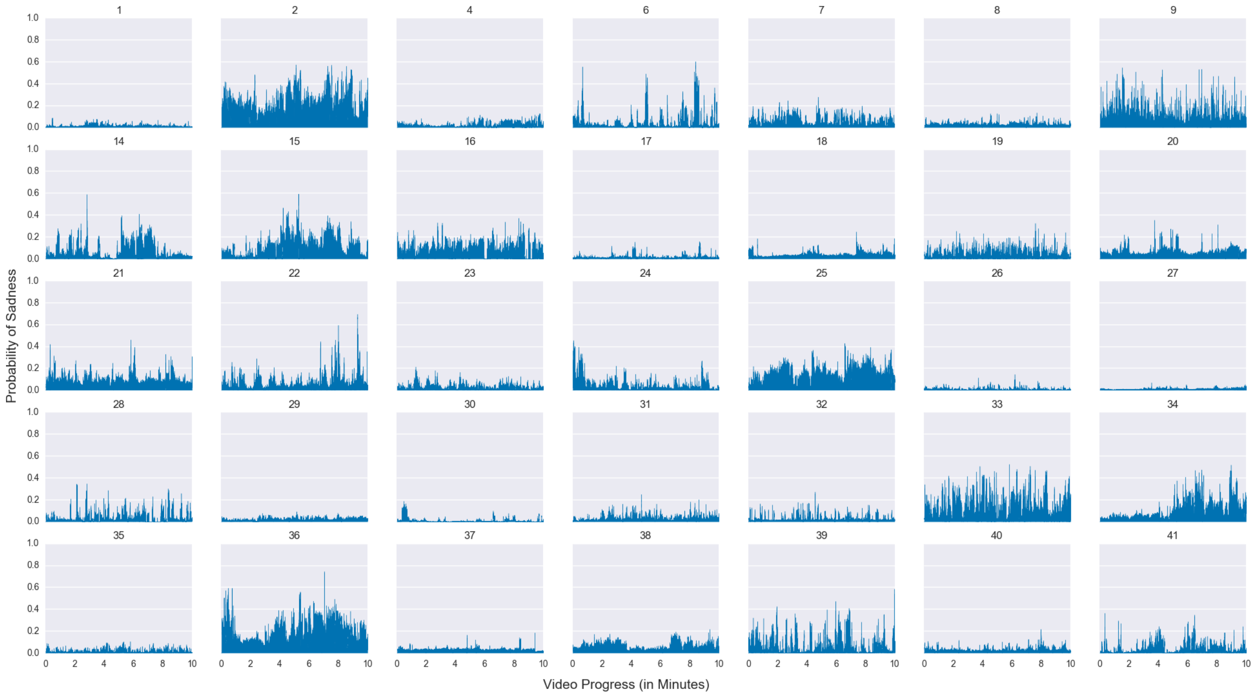
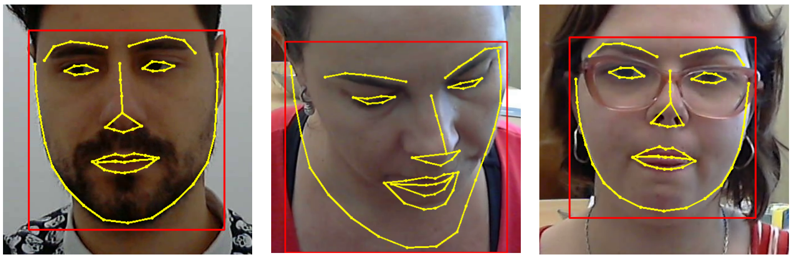


Figure 7.38: *Probabilities of sadness in the videos of the subjects*



(a) *Facial expression from subject 2*

(b) *Facial expression from subject 33*

(c) *Facial expression from subject 36*

Figure 7.39: *Examples of facial expressions classified as sadness*

played the horror game. Subject 4 had consistent detections of fear despite having hear face out of the camera's field of view (as illustrated in figure 7.43a). The detection of fear also seems correct in the case of subject 41, as depicted in figure 7.43c. Subject 6 also had higher detections of fear, as it is illustrated in figure 7.43b. This subject played the action-platform game Melter Man instead, so the emotion might have been elicited from the fear of making mistakes and falling from a platform. In a matter of fact, other subjects that played Melter Man also had detections of fear slightly higher than the rest of the participants (as it was the case of subjects 2 and 15, for instance).

Surprise

Figure 7.44 presents the probabilities of surprise detected on the videos of each subject. This was the second most difficult emotion to detect, as it is very sensitive to the difficulties in the tracking of landmarks. It produced many false positives due to poor tracking quality in subjects 9, 28 and 35, for instance, and due to partial occlusion of the mouth in subjects 17 and 39 (when they got their faces partially out of the camera's field of view for a brief moment). Also, it produced false positives for subject 32 when he lowered his head and the mouth landmarks were detected as if the mouth was opened. This also corroborates the importance of the region of the mouth for good detections of emotions. However, there were consistent detections of surprise related to eyebrows and eyelids

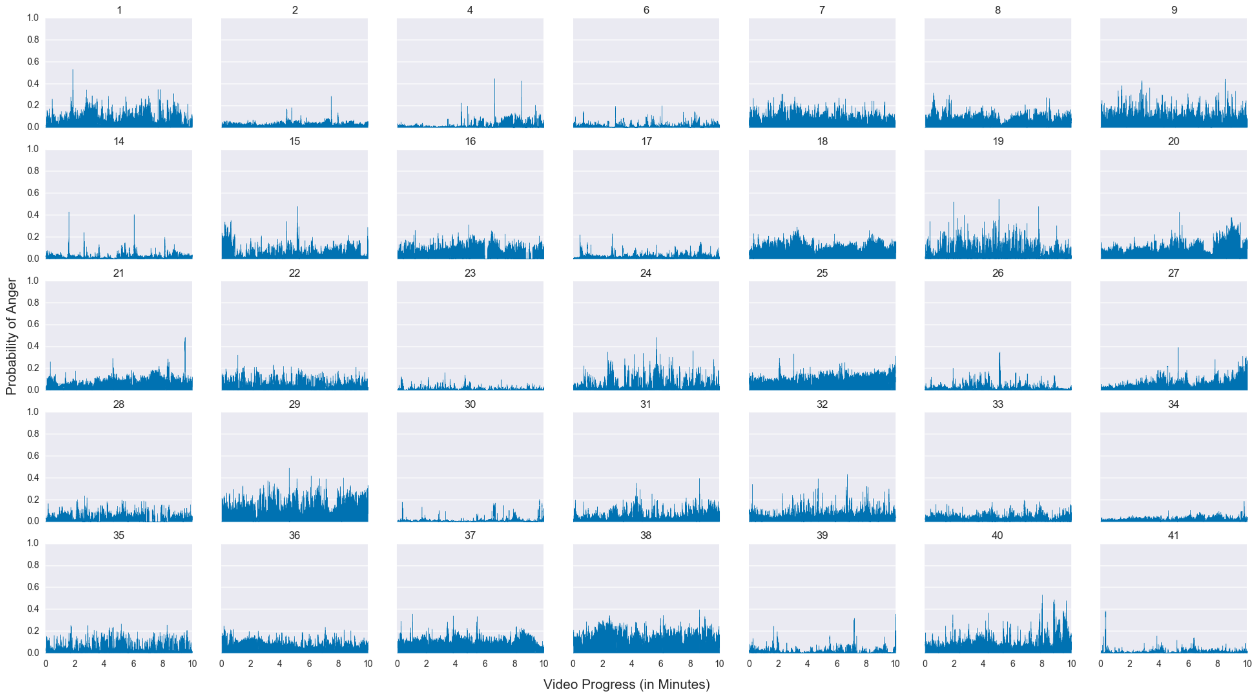
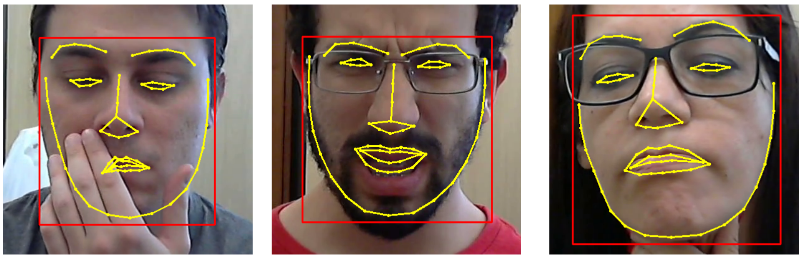


Figure 7.40: *Probabilities of anger in the videos of the subjects*



(a) *Facial expression from subject 14* **(b)** *Facial expression from subject 21* **(c)** *Facial expression from subject 40*

Figure 7.41: *Examples of facial expressions classified as anger*

pulled up together with mouth slithly opened. This is illustrated for subjects 1 in figure 7.45a), for subject 4 in figure 7.45b and for subject 23 in figure 7.45c. It is curious to notice that the biggest probability of surprise detected in the video of subject 4 occurred exactly when a teacher and a child student entered the room unannounced.

Disgust

Figure 7.46 presents the probabilities of disgust detected on the videos of each subject. This was the most difficult emotion to detect, as it was very sensitive to different issues. For instance, in the case of subjects 14, 29, 32 and specially 40 the false detections were due to the occlusion of the mouth. In the case of subject 23 the false positives were due to the lowering of the head. There were some true positives consistent with narrowed eye brows, curled upper lip and wrinkled nose – as it is illustrated in the faces of figure 7.47. Nonetheless, the probabilities of these detections tend to be just a little above the probability of the neutral expression. This is somehow expected, since the games employed didn't have many aesthetic elements that would specifically elicit disgust. The horror game Kraven Manor has some signs of violence (like blood on some of the walls), but from the illustrated only subject 4 played this game (the other two played Cogs).

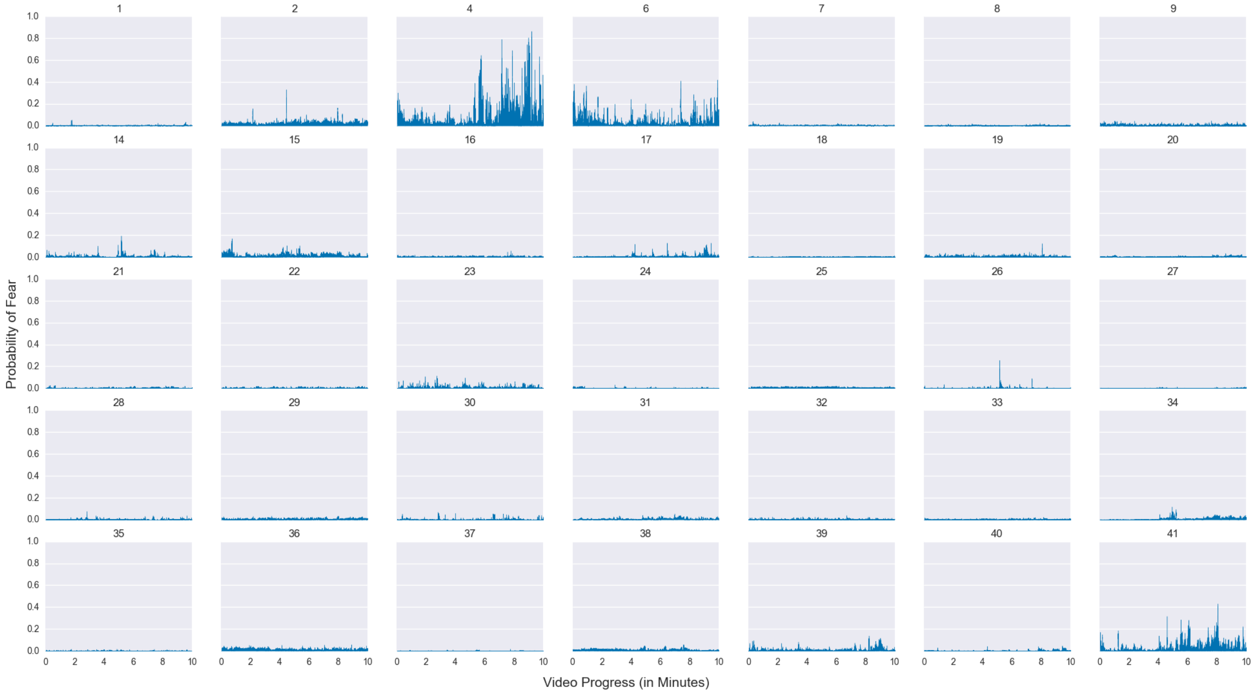
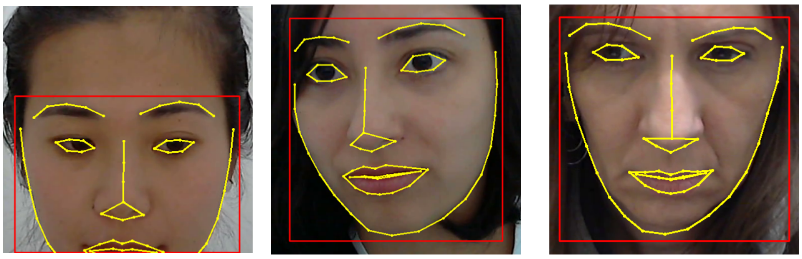


Figure 7.42: *Probabilities of fear in the videos of the subjects*



(a) *Facial expression from subject 4*

(b) *Facial expression from subject 6*

(c) *Facial expression from subject 41*

Figure 7.43: *Examples of facial expressions classified as fear*

Emotions per Game

The figure 7.48 presents the counting of each detection of the prototypic emotions plus the neutral expression in all videos, whenever the probability detected was higher than 50% (i.e. the emotion might not be the most probable, but it was significantly influential) and 70% (i.e. the emotion was the most probable). The counting is separated by game in order to produce an overview of how each different game might have influenced the elicitation of the emotions.

The neutral is by far the most detected expression, seconded by happiness. They both have been much more detected than the other emotions. This confirms the influence of the large number of samples from these classes in the datasets used for training the SVM classifier. The weight balancing procedure employed during the training helped, though. The datasets employed had almost 10% more samples of happiness than of the neutral expression, and yet the neutral expression was much more frequently detected with much higher probability than happiness (about 93% more frequently). Also, whenever the other emotions were detected, they usually had higher probabilities than neutral and happiness as well. As it has been discussed in the previous chapter, the classifier built gives more relevance to precision than to recall, meaning that it was not able to find all true instances of the emotions (particularly the less evident ones, that is, other than neutral and happiness), but had good precision when they were found.

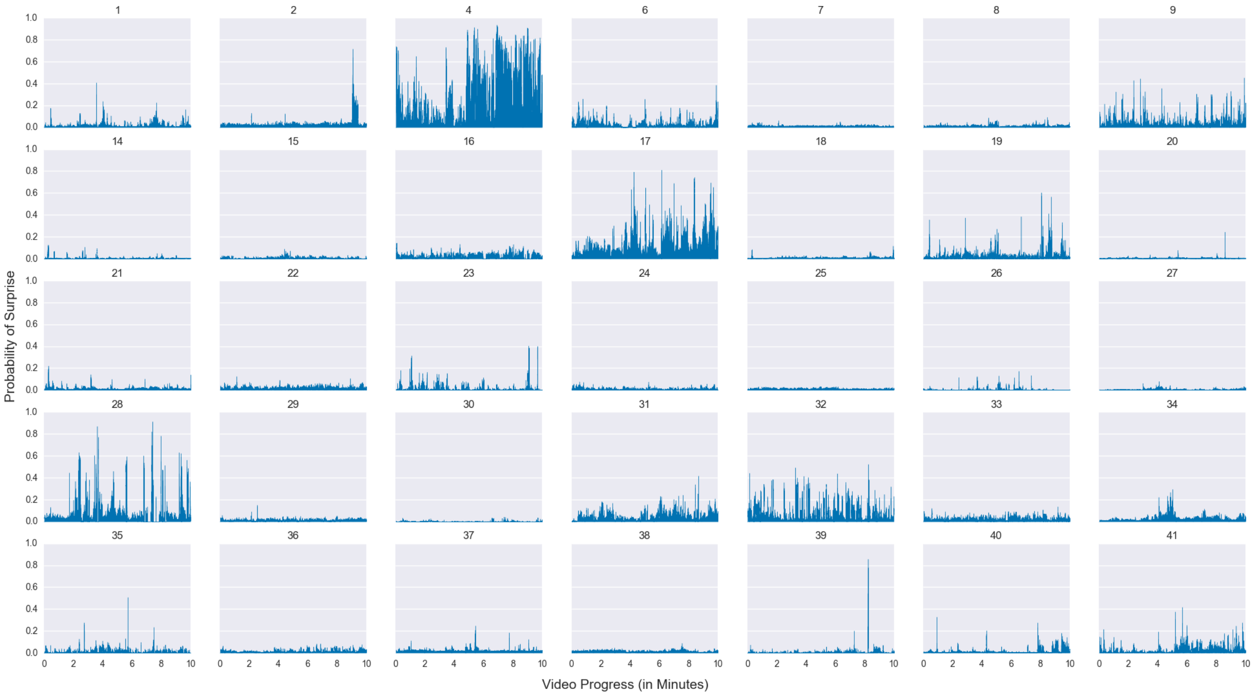
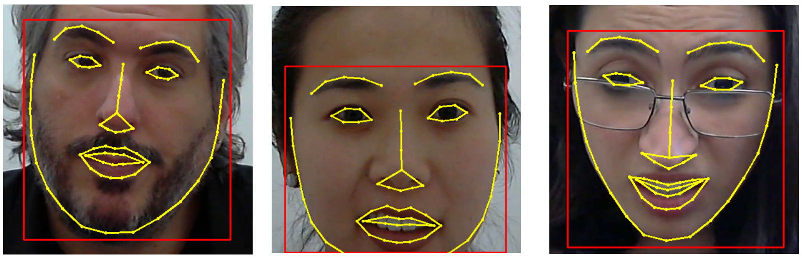


Figure 7.44: *Probabilities of surprise in the videos of the subjects*



(a) *Facial expression from subject 1*

(b) *Facial expression from subject 4*

(c) *Facial expression from subject 23*

Figure 7.45: *Examples of facial expressions classified as surprise*

Also, this figure is composed of all estimations from the videos of all subjects, including the unusable ones (those in which the detection failed consistently or the tracking quality was very poor). The detection of the prototypic emotions is very dependent on the mouth landmarks, so the large differences from the detection of happiness to the detection of the other emotions were also caused by the tracking difficulties.

Regarding other emotions than neutral and happiness, there is a visible tendency of Kraven Manor to elicit more surprise, disgust and fear, as it would be expected for a horror game. Melter Man also seems to elicit more sadness, what can be explained by the fact that this was the only game in which the avatar could die (when the player needed to restart the level after the death of her avatar in the game). Anger was almost not detected in any of the games.

As an extra validation step, an SVM classifier was trained from the probabilities of the prototypic emotions estimated from the videos of only the usable subjects, in order to verify if it would be possible to predict the game played from the emotions experienced. The evaluation of the trained model with K-Fold cross-validation using $k = 5$ folds produced an precision of just 0.35 with a 95% confidence interval of $\pm$ 0.18. By ignoring the two highest predicted expressions (neutral and happiness), a slightly better precision of 0.38 with a 95% confidence interval of $\pm$ 0.16 was achieved. Even though it was not possible to differentiate the games from the detected emotions, the increased in precision confirms the influence of the unbalanced samples in the training of the classifier. Perhaps

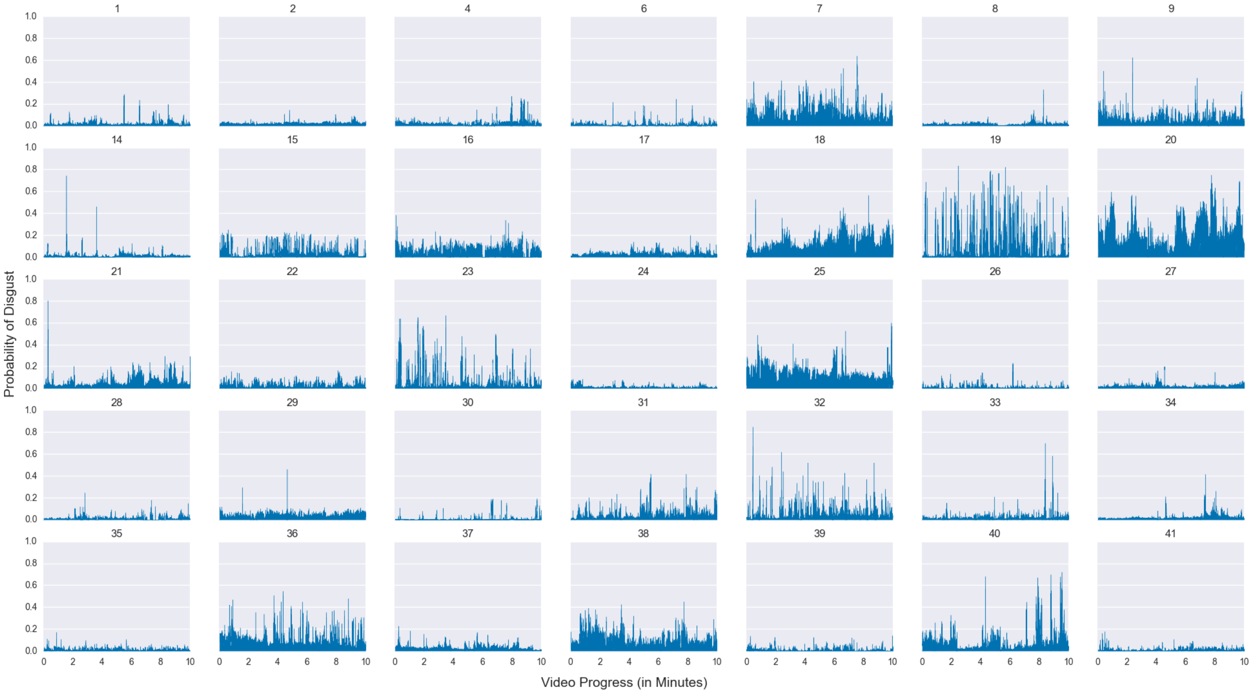
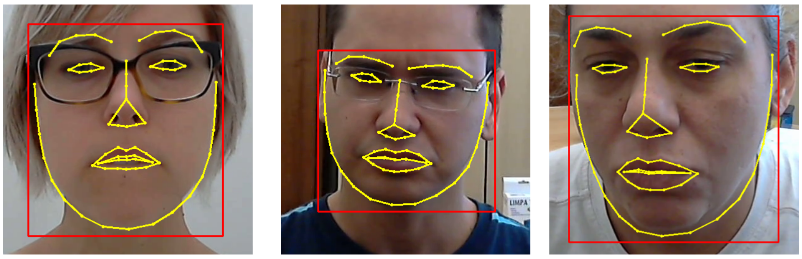


Figure 7.46: *Probabilities of disgust in the videos of the subjects*



(a) *Facial expression from subject 7*

(b) *Facial expression from subject 20*

(c) *Facial expression from subject 34*

Figure 7.47: *Examples of facial expressions classified as disgust*

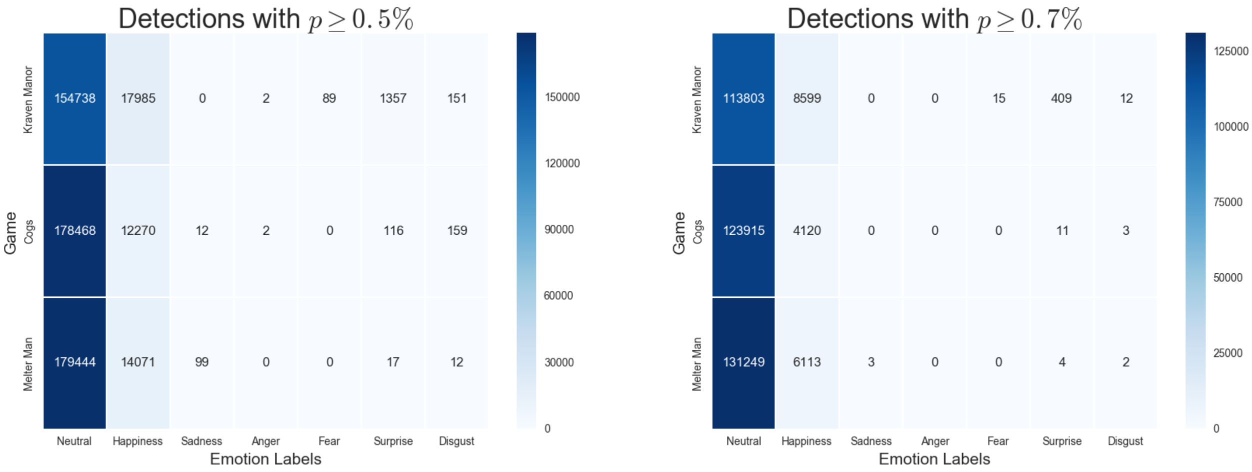


Figure 7.48: *Occurrences of the prototypic emotions with a probability higher or equal to 50% on each game*

by using only the probability estimation of a specific emotion (such as fear, for instance), it might be possible to better predict a game genre (horror versus others, in case of fear). But this is beyond the scope of this work and it is left to be verified on another occasion, as a future work.

7.2 Assessment of Fun

7.2.1 Detection of Frustration

Figure 7.49 presents the precision and recall scores obtained from the Leave-One-Out cross-validation, performed with the frustration classifier trained from the prototypic emotions (one classifier trained with all of the emotions, and another trained only with fear, sadness and anger – the prototypic emotions supposedly more related to frustration) detected in the videos of the usable subjects. As it can be observed, the best predictions (with both precision and recall above 0.80) were obtained from subjects 2 and 33 with the classifier trained with all emotions, including also subjects 21 and 40 with the classifier trained only with fear, sadness and anger. The worse predictions (with both precision and recall very close to 0.0) were obtained from subjects 4, 7, 17, 18 and 26, with both versions of the classifier. The recall of subjects 21, 23 and 40 was largely increased by using only fear, sadness and anger. All other subjects had relatively low precision and recall (near or bellow 0.5).

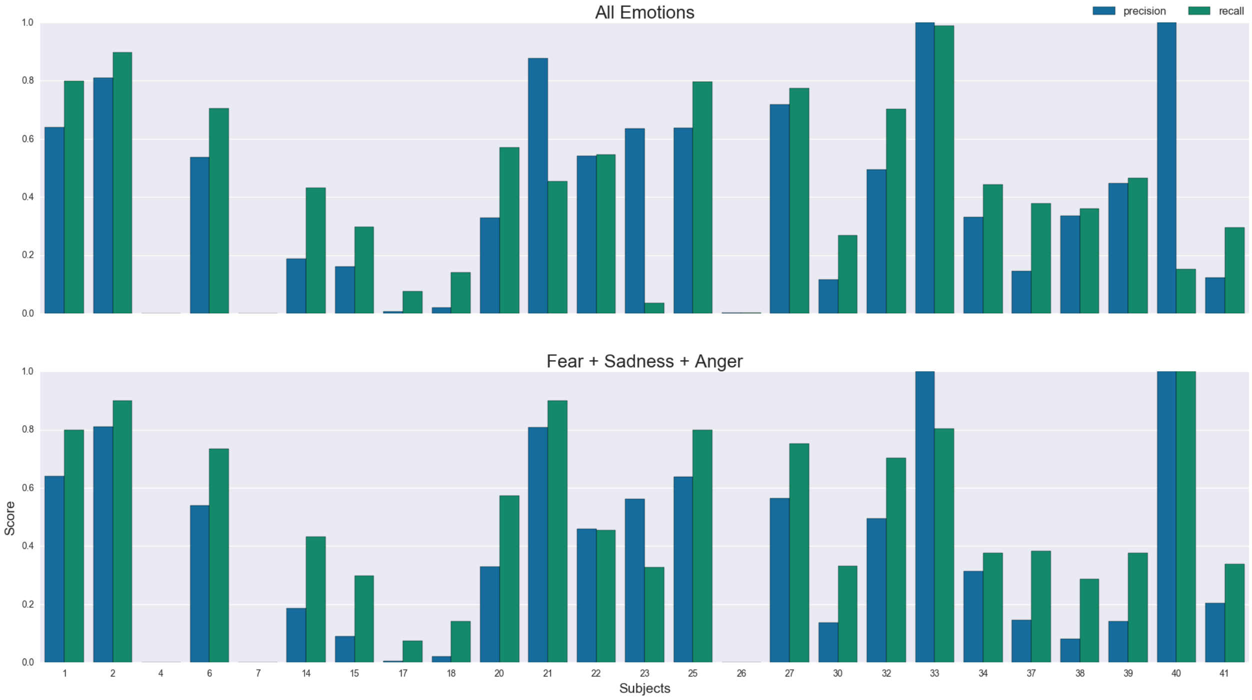


Figure 7.49: *Precision and recall scores from the frustration classifier*

The general tendency of the frustration classifier built is to have a low precision with a slightly higher recall, meaning that the classifier had many errors, producing more false positives than false negatives. This is clearly indicated in the box plots of figure 7.50. The classifier trained with all emotions achieved a precision median of 0.33, with Interquartile Range (IRQ) varying from 0.12 (25th quartil) to 0.64 (75th quartil), and a recall median of 0.43, with IQR varying from 0.15 to 0.70. The classifier trained with only fear, sadness and anger achieved a precision median of 0.31, with IQR varying from 0.09 to 0.56, and a recall median of 0.38, with IQR varying from 0.30 to 0.75. The conclusion is that the classifier built was not able to detect frustration with an accuracy sufficiently higher than chance. The causes are related to the unbalanced classes in the data used to train the emotion classifier and also to the lack of variance in the frustration data collected from the participants in the experiment.

The unbalanced number of samples used to train the emotion classifier made it more difficult to detected other expressions than neutral and happiness. This is demonstrated by the results from the classifier trained with only fear, sadness and anger, illustrated in the lower graph of figure 7.49 and the green box plots in figure 7.50. By using only the probabilities of these three emotions

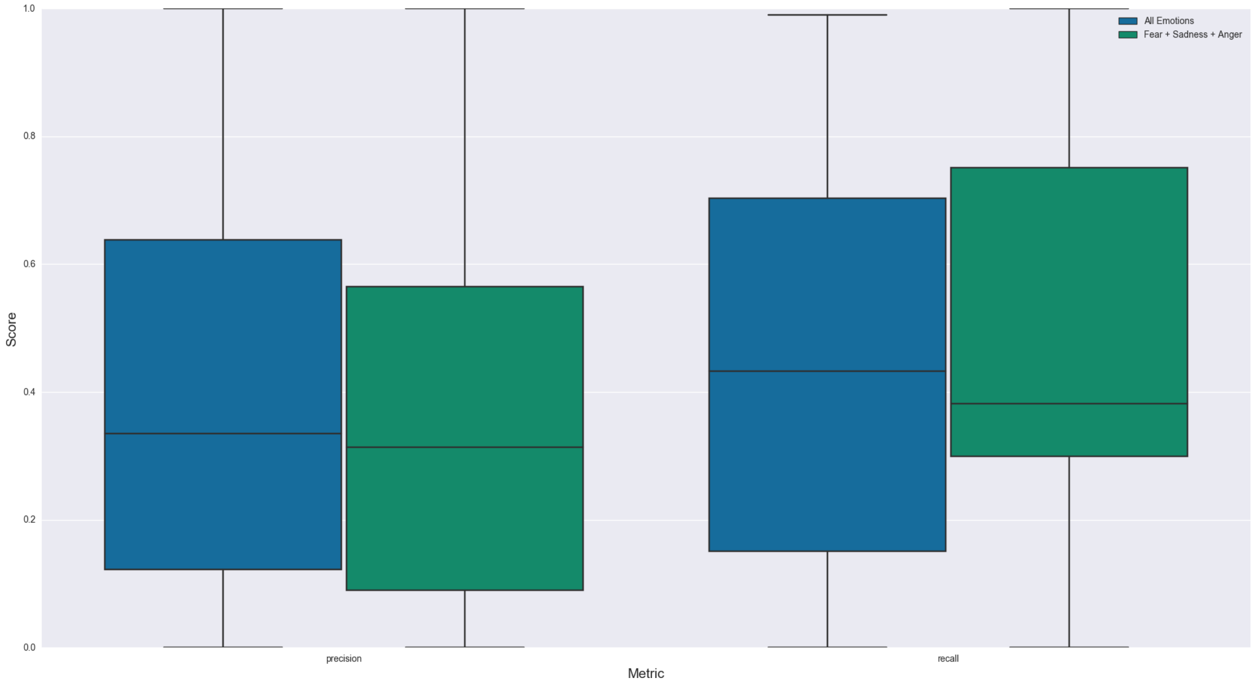


Figure 7.50: *Box plot of the scores for the frustration classifier*

the classifier results had a slightly reduction on the precision (IQR decreased from 0.12/0.64 to 0.09/0.56) but a greater increase on the recall (IQR increased from 0.15/0.70 to 0.30/0.75). The fact that almost no anger was detected by the emotion classifier also influenced the detection of frustration. So this classifier could have been improved if other emotions, particularly anger, had been detected with better recall.

The influence of low variability in the frustration data collected is observed in the comparison presented in figure 7.51. This figure indicates the self-reported levels of frustration (considered as the ground truth) and the predicted levels of frustration (using the classifier trained with all emotions). Only the levels reported (and predicted) at the moments of capture (i.e. when the questionnaire was presented to the players) are included, for brevity.

The most notable results are from subjects 4 and 26. They both self-reported the highest levels of frustration, statistically significant in comparison with the GEQ scores. Since they had more frustration reported, it would be expected for the frustration classifier to find more evidences of this affect in their videos. However, that was not the case, and the classifier completely failed to predict the reported levels of frustration by large margins of error.

It is also also noticeable that the predicted answers tend to be very low (near levels 0 and 1) in all of the subjects, with very little variation, indicating that the classifier is biased towards lower responses. In a matter of fact, many subjects reported very low levels of frustration, and hence the data used to train this classifier does not have great variability. For instance, subjects 1, 2, 7, 17, 18, 20, 21, 22, 23, 25, 27, 32, 33, 34, 39, 40 and 41 (68% of the usable subjects) have all reported levels bellow 2, with more than half of their responses bellow level 1. A better frustration classifier might have been produced if additional data was collected from more challenging or poorly designed games, in which higher levels of frustration were intentionally elicited in the players.

7.2.2 Detection of Immersion

Figure 7.52 presents the precision and recall scores obtained from a Leave-One-Out cross-validation, performed with the immersion classifier trained from the gradient of the face distance and the blink rate (in different combinations of these features) estimated from the videos of the usable subjects.

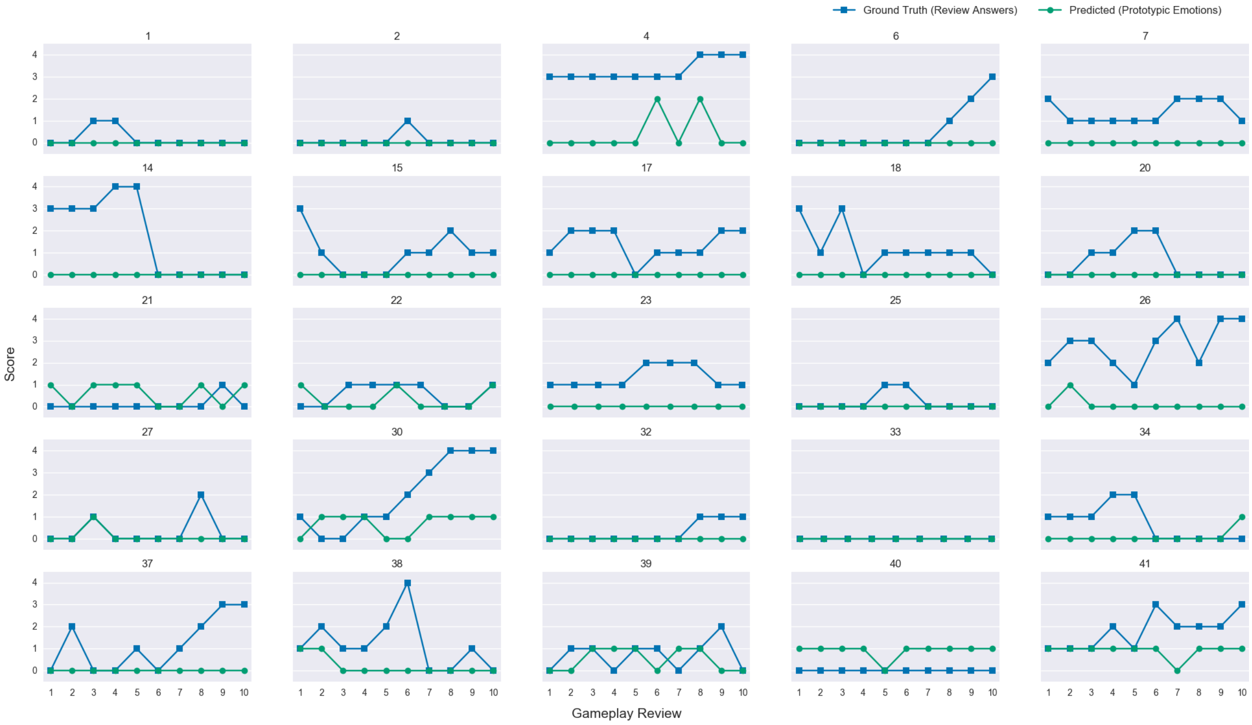


Figure 7.51: *Predictions of frustration for the reviewed moments of gameplay*

First of all, it can be observed that subjects 1, 4, 6 and 22 produced the worse results, in which the classifier was totally unable to predict the levels of immersion. Despite subject 4, the other subjects reported high levels of immersion, statistically significant in comparison with the GEQ responses. The best results were obtained from subjects 40 and 41, with precision above 0.8, but in general the classifier precision tends to be near or above 0.6, while the recall tends to be very low. This indicates an accurate classifier that is not capable of detecting many occurrences of immersion.

The blink rate seems to be a slightly better feature to characterize immersion. This is very noticeable in the results of subjects 7, 17, 23, 26 and 33, for instance, where both scores were null when the distance gradient was used alone to train the classifier but an improvement was obtained when the blink rate was used instead.

This is also observable in the box plots in figure 7.53. The distance gradient produced a lower precision, with median 0.2 and IQR varying from 0.0 to 0.64, and a higher recall, with median 0.45 and IQR varying from 0.0 to 0.8. On the other hand, the blink rate produced a higher precision, with median 0.42 and IQR varying from 0.12 to 0.73, and a lower recall, with median 0.30 and IQR varying from 0.12 to 0.43. So while the distance gradient tends to favour recall, the blink rate tends to favour precision. However, their combination did not improve the results, but yielded the worse precision and recall of the two features.

Even though the IQR ranges of both precision and recall are large, there still were many scores above 0.5 indicating that the precision of the classifier is generally better than chance. Also, it is important to consider that the metrics of precision and recall account for exact matches between the reported and the predicted levels. Since the landmark tracker had imprecisions (like the eye landmarks shaking between frames) and the word used to describe the affect ("involvement") might not have been properly understood by all participants, variations with a small margin would still indicate a classifier with good quality.

Figure 7.51 presents a comparison of the reported and predicted levels of immersion (using the classifier trained with the blink rate only). Again only the levels at the moments of capture are included for brevity. Differently than of what happened with the frustration classifier, the immersion classifier was not biased to lower levels. It is easy to note that even when the predicted levels are not

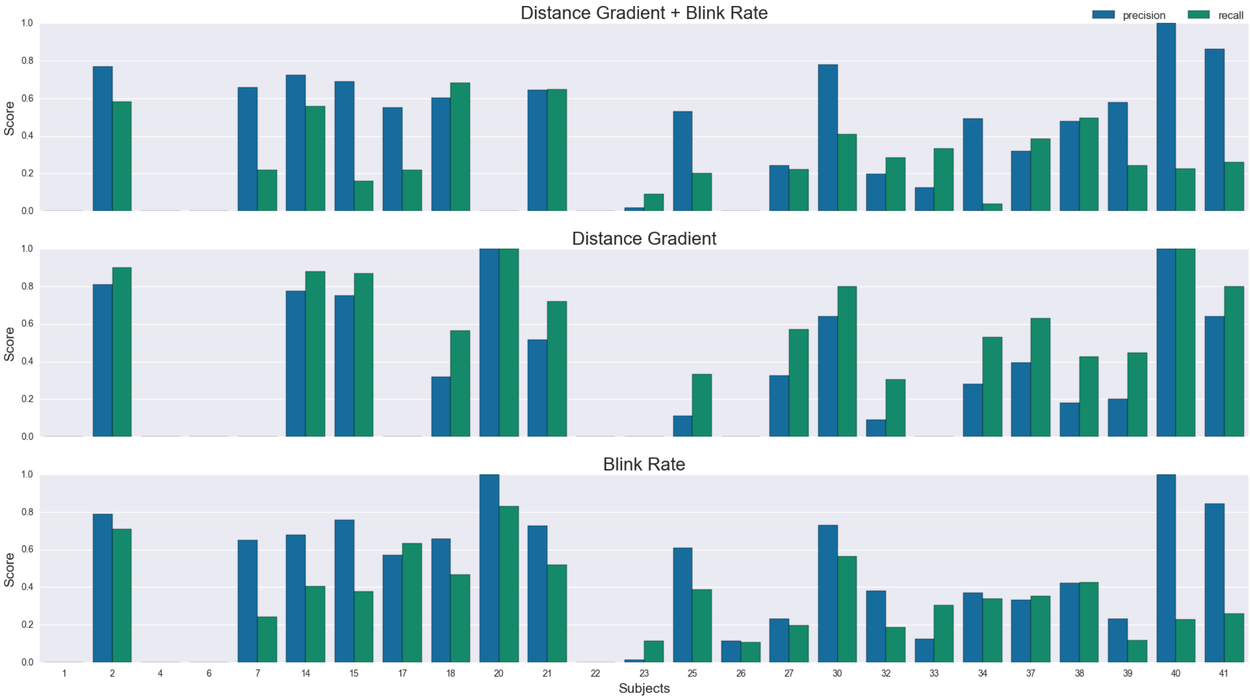


Figure 7.52: *Precision and recall scores from the Immersion classifier*

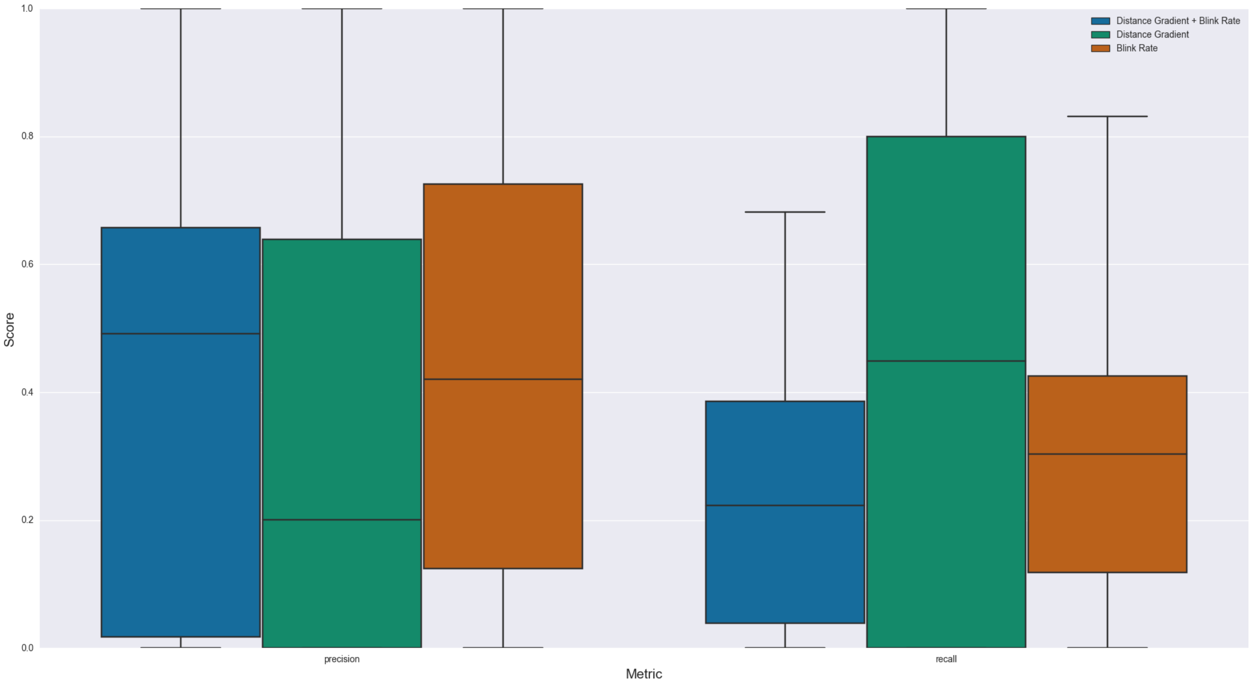


Figure 7.53: *Box plot of the scores for the Immersion classifier*

exactly the same as the reported levels, the variation along the gameplay (as the level of immersion increased or decreased) is almost the same, and with small margins.

Consider, for instance, the results of subject 18. While this subject didn't produce a high precision (because the reported and predicted levels only matched in two of ten reviews), the way that the immersion level increased and decreased in both the reported and predicted lines is very consistent (and the error margin is almost always of 1). Similar behaviour is observed in subjects 27, 32, 37 and 38, for instance. Also, whenever the subject reports didn't indicate variation (i.e. the reported level is constant during the gameplay), the classifier also tended to produce a constant prediction. That is observed in subjects 1, 2, 7, 20, 23 and 25, for instance.

Therefore, the results indicate that it is possible to assess immersion from the features studied, particularly by using the blink rate. Nonetheless, further studies are required, since the elimination of the minor difficulties with the landmark tracking and a better choice of 3D face model might improve the results with both these features.

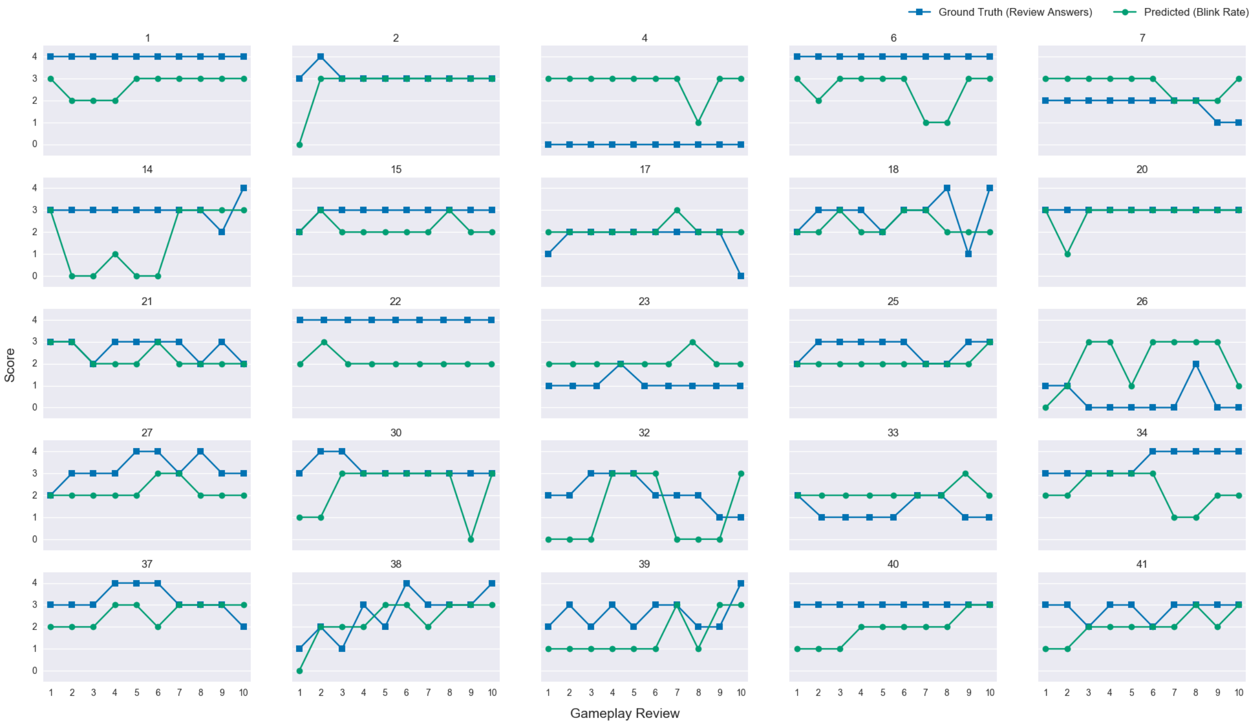


Figure 7.54: *Predictions of immersion for the reviewed moments of gameplay*

7.2.3 Detection of Fun

Figure 7.55 presents the precision and recall scores obtained from a Leave-One-Out cross-validation, performed with the fun classifier trained from the the prototypic emotions estimated from the videos of the usable subjects, and the frustration and the immersion levels reported by the subjects. Differently than with happened with immersion, the fun classifier trained with all features combined produced the highest precision scores, reaching values above 0.6 for many of the subjects. Other combination of features also yielded high precisions for some subjects, but it the best average results are definitely from all features (prototypic emotions, frustration and immersion) used together. On the other hand, the recall scores obtained varied among the different combination of features used.

Subjects 1, 15, 30 produced very low precision scores for all combinations of features. In case of subjects 15 and 30, the reason might be due to differences in the experienced and the reported levels of immersion, which were not statistically significant in comparison with the respective GEQ scores. The case of subject 1 is, however, intriguing. The face tracking for this subject was very good and the reports were statistically significant in comparison with the respective GEQ scores, and yet the classifier completely failed to detect fun. It might be the case of low expressiveness of this subject, since besides happiness and neutral, almost none of the other prototypic emotions were detected for this subject with probability bigger than 0.4.

Different than what happened with immersion, in which the combination of features did not improved the results, the fun classifier produced better results when all features were used: the prototypic emotions, the frustration and the immersion levels. Therefore, among the features used, the prototypic emotions seem to be the most important for both the precision and recall of the classifier produced. Frustration and immersion aided in recall depending on the subject, but in general they seem to contribute more to precision.

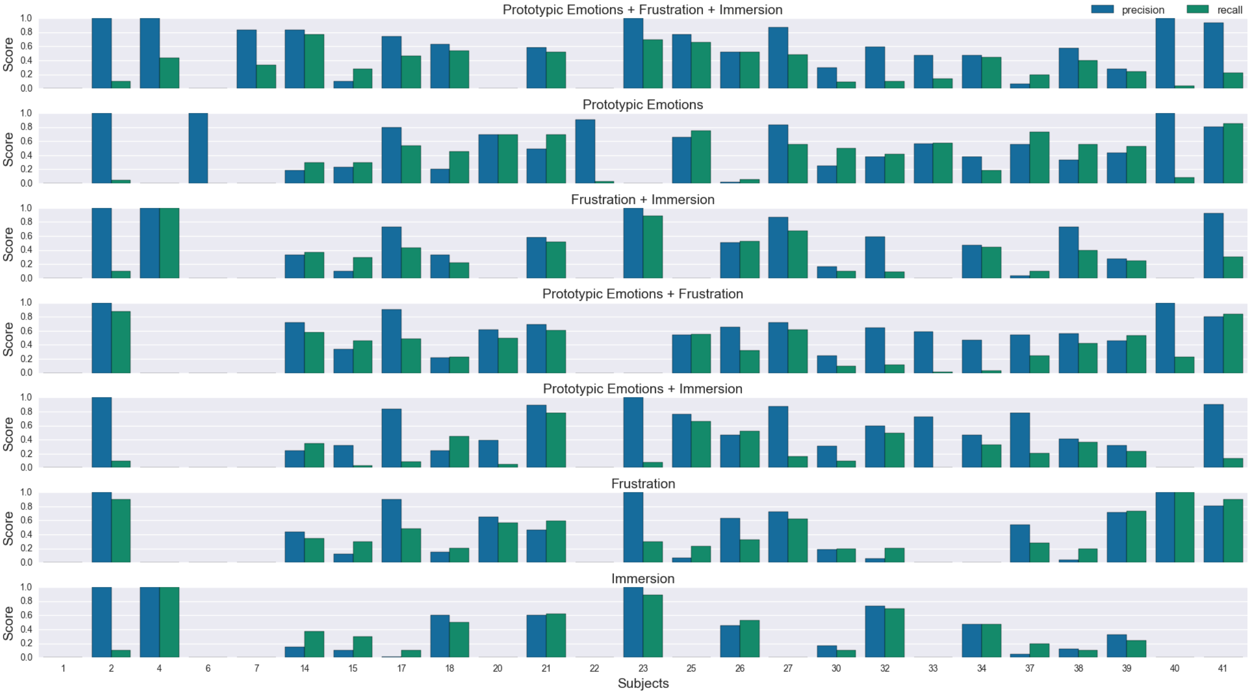


Figure 7.55: *Precision and recall scores from the Fun classifier*

This is easily observed in the box plots in figure 7.56. The use of all features largely improved the precision scores, yielding a median of 0.58 and an IQR varying from 0.28 to 0.84. By using only the prototypic emotions, there is a slightly reduction on the precision scores (all median and IQR values are reproduced in table 7.2 for easy reference). The recall of the classifiers produced tended to be low disregarding the features employed, with score values almost always bellow 0.5. Nonetheless, the highest median and IQR for the recall scores was definitely produced by the use of prototypic emotions only, followed by the combination of all features.

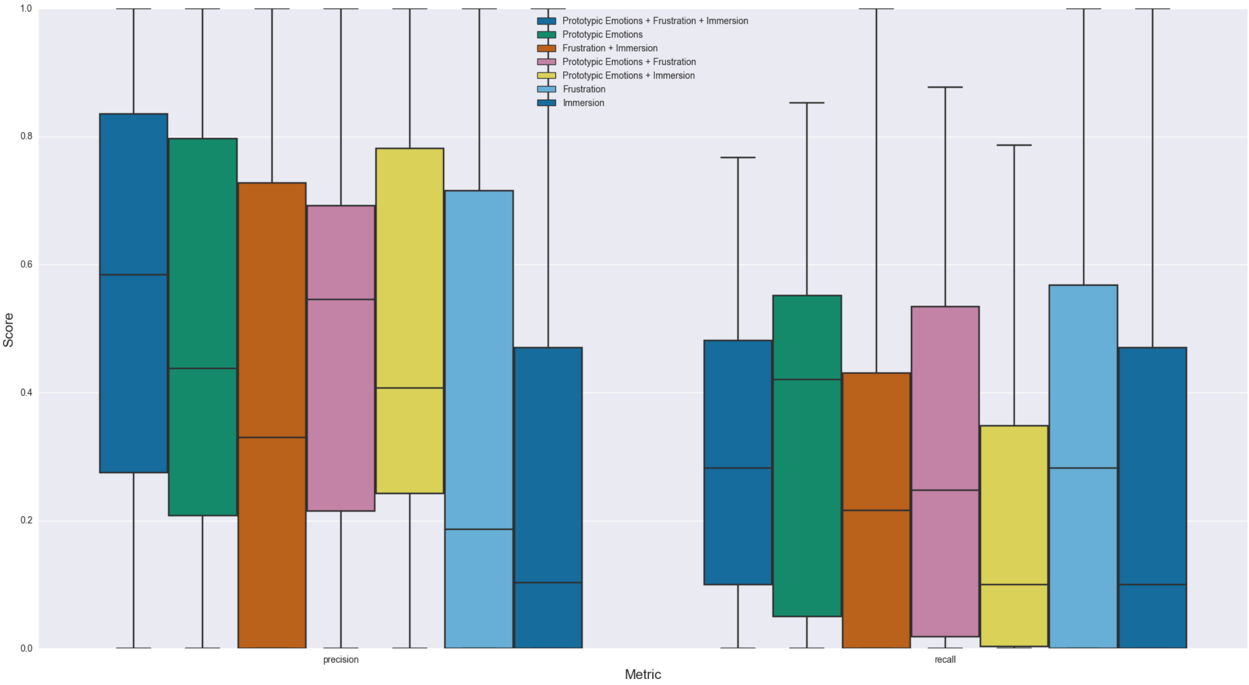


Figure 7.56: *Box plot of the scores for the fun classifier*

These precision and recall values indicate that the classifier was not able to identify many of the occurrences of the levels of fun, but when it did the precision was good and above chance. But

Table 7.2: *Score statistics for the fun classifier*

Features Used in Training	Precision		Recall	
	Median	IQR	Median	IQR
Prototypic Emotions + Frustration + Immersion	0.58	0.28 to 0.84	0.28	0.10 to 0.48
Prototypic Emotions	0.44	0.21 to 0.80	0.42	0.05 to 0.55
Frustration + Immersion	0.33	0.00 to 0.73	0.22	0.00 to 0.43
Prototypic Emotions + Frustration	0.55	0.22 to 0.69	0.25	0.02 to 0.53
Prototypic Emotions + Immersion	0.41	0.24 to 0.78	0.10	0.00 to 0.35
Frustration	0.19	0.00 to 0.72	0.28	0.00 to 0.57
Immersion	0.10	0.00 to 0.47	0.10	0.00 to 0.47

as with the other affects, these metrics are computed by counting the number of exact matches between the reported and predicted levels of fun, so it is important to also analyse the error margin between the reported and predicted levels.

Figure 7.57 presents the comparison of the reported and predicted levels of fun, from the classifier trained from all the features. As before, only the levels at the moments of capture are included for brevity. With the exception of subjects 1, 6, 20, 22 and 30, in which the error margin between the reported and predicted levels was very large, all other subjects had an average error margin very close to 1 and very similar patterns of variation. For instance, it is easy to observe in the graphs of subjects 7, 18, 32, 37, 38 and 39, that the predictions follow the increases and decreases of the reported levels of fun. This is an indication that the classifier was indeed able to predict the way that the level of fun varied during the gameplay, with a small error margin.

The levels of frustration and immersion used to train the fun classifier was not detected, but instead used directly from the reported levels. Also, the understanding of the word “involvement” (used to represent immersion in the experiment for data collection) might not have been consistent among all subjects, and the estimated probabilities of prototypic emotions were produced by a unbalanced classifier (that detected neutral and happiness expressions easier than the other expressions). All these characteristics might have influenced the results. In a real life scenario, in which the levels of frustration and immersion would be detected instead of being reported, a better precision may be achieved depending on the quality of these sub-classifiers.

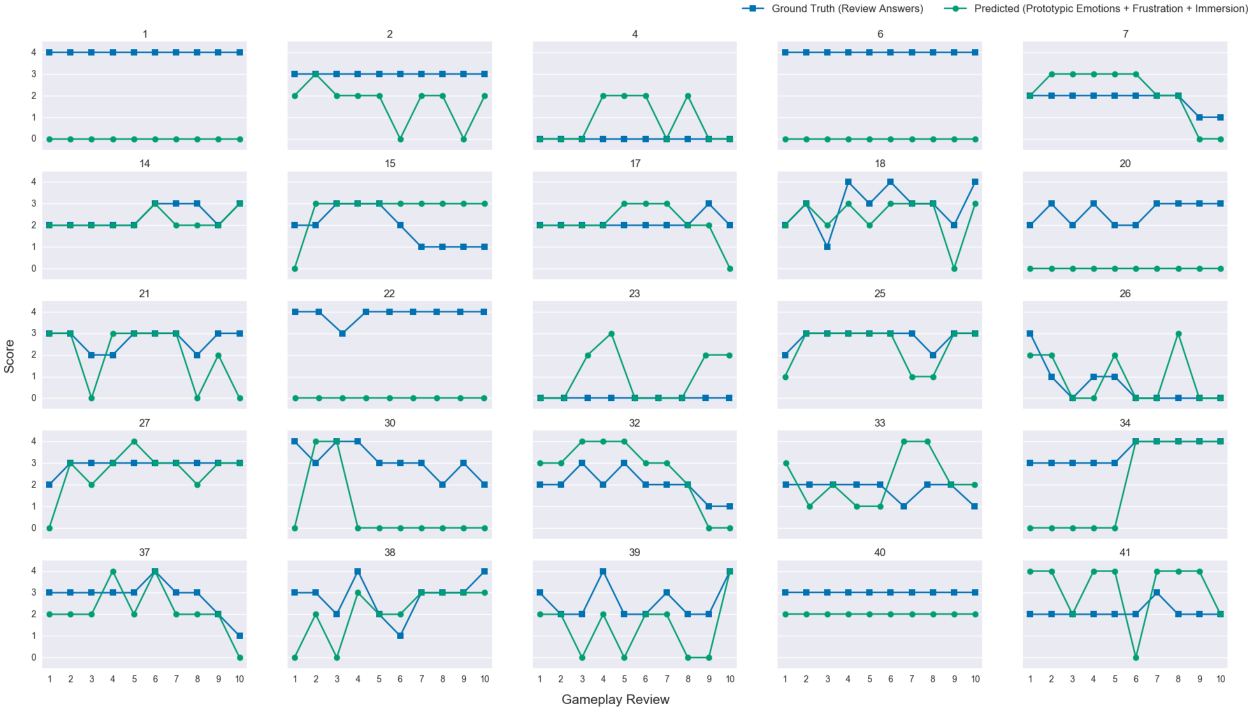


Figure 7.57: *Predictions of fun for the reviewed moments of gameplay*

Chapter 8

Conclusion

The objective of this work was to verify the possibility of assessing fun from just the analysis of facial images captured from players using common off-the-shelf cameras. That is, without requiring any other source of information, such as utilitarian metrics from the game or psycho-physiological measurements from sensors attached to the body of players. This objective was motivated by the formed understanding that the human face is a very important channel of communication for humans (Bettadapura, 2012; Koster, 2010; Matsumoto and Hwang, 2011), which conveys relevant data related to the affects involved in the experience of fun. It was not in the scope of the work to provide a self-adapting mechanism for games, but only to investigate if this channel could be used to aid game designers to produce better games.

The work started with an extensive review of the literature regarding what fun is and how it can be (and has been) assessed, described in chapters 2 and 3. The understanding formed and summarized in those chapters is that fun is an affect mainly characterized by immersion and emotions (Brown and Cairns, 2004; Calvillo-gámez *et al.*, 2010; Canossa *et al.*, 2011; Douglas and Hargadon, 2000; Jennett *et al.*, 2008; Lazzaro, 2010; Mäyrä and Ermi, 2011; Nacke *et al.*, 2008; Scherer, 2005). The former is related to the increasing level of involvement with a task or object, which requires the focus of a limited attention. The latter is related to the physical and psychological responses linked to the preparation of the body for action and to the signalling of similarly choices as preferred or undesirable. The experience of such affects together is pleasurable by nature and produce many changes in the body, including changes in the sensory organs and muscles of the face. The level of attention is linked to the movement of the sensory organs or sensory-related actions, such as blinks, pupil dilation and leaning the body towards an object of interest. Therefore, immersion can be assessed from the inspection of these changes. On the other hand, facial expressions are also strongly related to the display of emotions and natural to the point of being acknowledged across cultures. Hence at least the prototypic emotions can be assessed from the inspection of changes in the face regarding the movement of permanent features (such as eyes, eyebrows and lips) or changes that produce transient features (such as furrows, wrinkles and lines).

The following chapters described the approach used in this work, in which the measurements of the gradient of the distance between the face and the camera and measurements of the average number of blinks per minute were used as features to assess immersion. These features seemed to be the most robust and general, disregarding the type of the games involved. The work also employed responses of a bank of Gabor filters to represent the texture changes in the face images as the features to assess the prototypic emotions. Texture analysis has been pointed in the literature as more robust and producer of better results than geometric analysis. These estimations

Separated estimations of the prototypic emotions, the level of frustration and the level of immersion were obtained using a Support Vector Machine and a Structured Perceptron trained from the sequential data extracted from the videos collected. Those affects are all important aspects of fun, which by their own could also be very useful to support game design. But the main intention was to use their estimations as the features to estimate fun. Frustration was included as a more specific emotion related to the anxiety involved in challenge. So texture features extracted from the images were used for the estimation of the probabilities of the prototypic emotions and also for the estimation of the level of frustration. The rate of blinking in blinks per minute and the gradient of the estimated distance of the face from the camera were used as features for the prediction of immersion. Finally, the prediction of fun was attempted using the different combinations of these features, in order to verify which would produce better results. Statistical comparisons were performed with the self-reported data provided by the participants in the experiments for data collection.

The data collection used three popular independent games of different genres (horror, action/platformer and puzzle) that were played by volunteers during a short session of ten (10) minutes. The video of their gameplay was reviewed by the volunteers in the following using the well-known Game Experience Questionnaire and a custom questionnaire intended to collect the answers to be used by the classifiers to be trained. The short game sessions were an experiment design choice: the intention was to avoid tiring the players with a too long review (which could bias the answers) while providing enough time for the players to get immersed by the gameplay.

8.1 Overview of the Results

From the intended number of 60 participants, only 41 volunteers played the games. Some participants (about 30%) knew the researcher for being fellow students at the University, which freely applied to join the experiment in response to the e-mail invitations. Due to technical difficulties (explained in chapter 4) only the data from 35 of the volunteers was recorded. All games were played and no participant reported to have played the assigned game before. The option to have 10-minute sessions seemed appropriated for all games but the horror one: no volunteer that played that game advanced further than the first room of the manor or had the chance to encounter any of the points of "jump scare". Nonetheless, they reported to have felt tense, anxious and afraid of the environment and music.

The face tracking system worked well most of the time but it still had important flaws. Thick lenses and lighting variations were a recurrent problem, as well as "flickering": even when the detection and tracking was good there were minor variations in the position of the detected features from one frame to the next which induced mistakes in the detection of blinks. The lowering of one's head was also a cause for many tracking issues, particularly for players that are not used to the game interface (the keyboard and mouse were used). The estimations of blinks and the face gradient were very good considering the quality of the face tracking. And the estimation of the probabilities of the prototypical emotions were good, but biased towards neutral and happiness expressions due to the databases employed in training this classifier.

The estimation results of frustration were inconclusive, with no prediction above chance. The reason might have been related to the choices of games and to the play time in the experiment. The games employed were very popular and well-known to be fun by the gaming community (not by the players, which reported to not have played them before – as already mentioned). Also, the game sessions were limited in order to avoid tiring the players during the gameplay review. Therefore, the collected data didn't have much variation regarding this affect, as it was observed in the answers to the review questionnaire (which were almost all comparable with statistical significance to the GEQ answers). This might have impaired the classifier produced, biasing it to low-level responses (i.e. near no frustration at all). Better results might have been obtained if games well-known *to not* be fun would have also been used or if more difficult levels had been defined for the volunteers to play.

The results regarding immersion were slightly better. With the classifier produced from the blink rate and the face distance gradient, it was possible to predict immersion in many subjects with a precision above 0.5, even though the classifier was not able to detect all relevant levels of immersion that possible occurred (i.e. it had a general low recall). Also, it was observed that the blink rate seems to be a better feature for this affect, yielding slightly better results. The choice of the word "involvement" (to represent immersion in a way that was thought to be easier to understand by laymen), might not have been adequate and perhaps influenced the results. Immersion can be related to sensory and fantasy involvement as well as the consumption of attention in the execution of tasks, and some subjects didn't have statistically significant similarity with the scores of the Game Experience Questionnaire (GEQ).

Finally, the results regarding fun were much better. The classifiers were not trained with the detected levels of frustration and immersion, in order to avoid having their specific difficulties impairing the results. Instead, the reported levels of these affects were used, since almost all subjects had statistically significant and consistent responses in comparison with the responses to GEQ. The use of all features, that is, the prototypic emotions and both the levels of frustration and immersion, yielded the best results, with precision median of 0.6 and many predictions with a precision 0.8. This classifier was also not good in finding all occurrences of the significant levels of fun (i.e. it had a low recall), but that is attributed to the difficulties in the tracking of landmarks and to the bias towards neutral and happiness that occurred in the emotion classifier.

The most important observation is that the immersion and the fun classifiers produced very consistent variation patterns of the respective levels. That is, they were able to detect the increases and decreases of the levels of immersion and fun in the subjects in a very close fashion to what has been self-reported by players, with an average margin of error close to 1 level. This is a strong indication that if the quality of the base classifiers (the classifier of the prototypic emotions and the classifiers of frustration and immersion), much better results can be obtained by improving both precision and recall.

8.2 Discussion of the Results and Future Work

This work indicates that the assessment of fun based on the analysis of digital images captured from the faces of human players is possible. However, there were three important issues that must be properly resolved in order to allow for better predictions. They are:

1. **The quality of the face tracking**. One of the most sensitive aspects for the assessment of fun is the quality of the face tracker. Since all other estimations depend upon a good detection and tracking of the facial landmarks, failures in this activity will definitely impair the results. The main difficulties observed in this work were related to occlusion of the mouth from the hands and to framing problems (when the face is partially outside of the camera's field of view or the head is rotated so the face is not in a very frontal position). The lowering of the head happened frequently with casual players, because they constantly needed to look down to the keyboard, for instance. But even minor mistakes with the tracker may cause difficulties. A very popular library (DLib) was used in this work, capable of producing very good results in most of the cases. But still many failures were observed and illustrated in chapter 7. If the tracking of the landmarks is not *very good*, all the following efforts will be much more difficult because blink detection will not be precise enough or because the facial expressions might be represented by image regions that do not trully belong to the face.

2. **The quality of the prototypic emotion detection**. The prototypic emotions are very important for the detection of fun. Thus, a reliable estimation of their probabilities is important. Existing datasets have fewer samples of other expressions than neutral and happiness, making it difficult to produce properly balanced estimations. Also, prototypic emotions should be considered in a time scale, and an structured prediction approach should be used instead of trying to predict them from single frames of video. The difficulty with this is that there are

no easily accessible datasets containing sequences of emotions, from onset, to apex and offset, captured in nature, that could be used to train a structured predictor of prototypic emotions.

3. **The reports on immersion**. Immersion is a trick word for questionnaires. The affect itself may be related to the engrossment with the fantasy, with the art, or with the involvement in a challenge. Indeed, GEQ does not ask for it directly, but uses 5 related questions (asking if the player felt imaginative, if she thought she could explore things, of if the game was aesthetically pleasing – since its concept of immersion is more related to the sensory and imaginative immersion). The choice for "involvement" was thought to be a good choice in the beginning of this work, but some subjects might have only consider it regarding the challenge point of view. So, that choice was probably wrong. Since the affects related to fun are subjective and dependent upon the interpretation of the participants, the collection of data should be performed with greater care. In a matter of fact, even the order in which the questionnaires were done (the players reviewed the video of the gameplay before answering GEQ) might have created biases. In future works, the order in which a questionnaire and the gameplay review are done shall be randomized.

As it has been discussed in chapter 3, there is no way to really know the emotional state of a person than to ask the individual to report on the true nature of her feelings (Scherer, 2005). That is so because of many reasons, but mostly because there are many factors actively influencing an individual emotions beyond the game. Also, the interpretation each person has on the physiological responses their body are producing depend upon their own previous experiences, so the reported and predicted levels of any affect might not always exactly match. With that said, any prediction of fun should not be intended to automate the design decisions based on what the players are supposedly experiencing, but instead should be used to provide useful information for designers to improve their products. In that line of reasoning, it is worthwhile to include other information that might help to assess fun.

Frustration, for instance, considering as the result from repeated failures in achieve goals, could be enhanced by game performance metrics. Nonetheless, its classifier might be improved by just employing data collected from intentionally poorly designed or difficult games, in which frustration is much more elicited. If immersion is to be separated into subcategories of this affect, like sensory, aesthetic, fantasy and challenge immersion, it might be easier not only to collect data from participants, but also to produce better and specific classifiers.

Appendix A

Free and Clarified Consent

This is the English version of the consent form that had to be signed by the volunteers as a requirement to participate in the experiment to collect the gameplay data.

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
Departamento de Ciência da Computação
Rua do Matão, 1010 05508-090 São Paulo, SP, Brasil

FREE AND CLARIFIED CONSENT TERM

Research title: **Assessment of fun in videogames from the analysis of facial expressions**
Main researcher: Prof. Dr. Flávio Soares Corrêa da Silva
Assistant researcher: Luiz Carlos Vieira

1. **Nature of the research:** you are being invited to participate of this experiment with the objective of helping a research intended in building a software tool that will measure the level of fun from facial images of players to support the design of digital games (videogames). During the experiment you will remain seated in front of a personal computer, alone in a private room, and will be invited to play 1 game randomly selected whilst a video camera records images of your face and the computer records images and sound effects of the game. After 10 minutes playing you will be requested to answer a first questionnaire with 38 simple questions regarding your general experience with games and your experience with the game you just played. Following that, you will be requested to watch a video clip of your playing session as you answer a new questionnaire with 3 questions regarding your satisfaction with the game. Only the images of your face (without voice) and the images and audio of the game will be captured for later analysis. This experiment shall take 20 minutes in total to be performed.

2. **Participants:** any person aged between 18 and 65 years old, with no visual, auditory or motor deficiencies that prevent them to use a personal computer, that are not sleepy or tired, and that do not have a history of Photosensitive Epilepsy or Repetitive Strain Injury (RSI).

3. **Involvement:** by participating of this study you allow the capture of images and sounds of the game selected for you to play, as well as the capture of your facial images as you play it. The data collected from the game includes audio (effects and music) and video, but the data collected from you includes only your facial images, without any auditory (voice), personal (name, document, nationality, address, email, etc) or contextual (date or local) information that might permit to identify you. By participating you also allow the researchers in charge to use the collected data *only for computational manipulation*, in which the images will not be directly handled or accessible by people using the software produced. Additionally, and in a totally independent fashion, you may allow – *if you wish* – the reproduction of your face images in articles, papers or academic presentations, and only for purposes of scientific divulgation of the methods employed (for instance, to illustrate a certain facial expression relevant for the detection of fun). In case you decide to agree with that use, you still have the guarantee that those divulgations will not permit to identify you directly, since no other identity data will be collected and the images will be referenced as “subject 001” or “subject 043”, for instance. Please be also aware that the option for not authorize that divulgation *does not prevent you* from participating of the experiment, in the way described earlier. That is, if you wish to not authorize

the use of your face images in publications, you can only allow for the data to be used in the computational processing primary intended. Finally, please remember that you have total freedom to refuse to participate and yet to refuse to continue participating during *any phase of the experiment*, *and without the need to provide any explanation*. Both the refusal and withdraw options do not bring you any loss, and the researchers in charge guarantee you that in any of those cases any data that might already have been captured from you will be eliminated immediately in a complete and irrecoverable form.

4. **Risks and discomfort:** the participation in this experiment does not bring you any legal complications and it does not offers any risks to your health and dignity. The duration of the game session is restricted in order to avoid risks related to the prolonged exposition to repetitive tasks and light patterns. The games employed are of free use and their contents do not discriminate race, colour, sexual orientation or religion. They might contain, however, violent or scary images or situations. The procedures adopted in this research follow the Ethics Criteria in Research with Human Beings, according to the Resolution of number 466/2012 and their complements, from the Brazilian National Health Council (CNS, *Conselho Nacional de Saúde*).

5. **Guarantee of compensation:** Despite the fact that this experiment does not bring you any legal complications and does not offer you any risks to health and dignity, you have the guarantee that any eventual damage will be duly compensated by the institution of the researchers in charge, according to the Brazilian law.

6. **Confidentiality:** all information collected in this study is strictly confidential. Only the researchers in charge will have access to the data, which will not be made publicly available at an Internet website or stored in any folder or server publicly accessible at the University. Other researchers that might be interested in using the data, only to scientific purposes, might have access to it *if, and only if, they formally commit to this exact rules*, in the way that you are authorizing and according to a rigorous control from the main researcher.

7. **Benefits:** you will not gain any direct benefit to participate of this experiment. We do hope, however, that the study will produce relevant information for the development of digital games (videogames), for educational, commercial and scientific fins.

8. **Expenses and payments:** you will have no cost or dispense to participate of this experiment, and you will not receive any form of payment or reimbursement for your participation.

9. **Following up:** Your participation will be resumed to this experiment, without the need of any future intervention or new collection of data from you. The results will be published after the conclusion of the research, by the University of the researchers in charge, even though the data are still confidential and controlled. In case the project is cancelled, whichever may be the reason, all collected data will be immediately discarded in a complete and irrecoverable way. To guarantee you that what has been accorded here will be fulfilled, you will receive a copy of this document signed by one of the researchers in charge – both duly identified at the end of this document. Besides that, you are welcome to request more information on this research by contacting us via e-mail or telephone. If you prefer or need, you can also directly contact the Ethics Committee via the following address: Av. Raimundo Pereira de Magalhães, 3305, Pirituba, São Paulo – SP, Brasil, CEP: 05145-200, or via the e-mail cep.uniansp@anhanguera.com, or still via telephone telefone +55 11 3512 8415.

After these clarifications, we kindly request your consent, of your own free will, to participate of this experiment. Please fill the following items, marking with an X the items bellow.

Obs.: Do not sign this document if you still have doubts.

Free and Clarified Consent

Considering the items presented above, of my free will and having clarified all my doubts, I manifest my consent in participate of the experiment. I declare that I have received a copy of this term and authorize the execution of the research and the publishing of only the results obtained in this study.

□ I authorize the collection and computational utilization of images of my face, of images and sound of the game I will play, and the usage of my responses to the questionnaires that I will answer.

I also express my choice regarding the use of images of my face in scientific publishing, considering that it will not be possible to identify me directly from them (**select only one option among the following choices**).

○ I also authorize the reproduction of images of my face for scientific fins, considering that it wil not be possible to identify me directly from them.

○ I do not authorize the reproduction of images of my face or the use for any reason other than the computational use primarily intended by this research.

Signature of the participant:

Signature: ____ Participant Number: ____________

Signature of the researcher in charge:

Name: ____

Signature: ____ Date: ____________________________

Main researcher: Prof. Dr. Flávio Soares Corrêa da Silva (fcs@ime.usp.br)

Assistant researcher: Luiz Carlos Vieira (lvieira@ime.usp.br)

Contact phones: +55 11 3091 6134 and + 55 11 3091 6135

Appendix B

Grid Search for SVM Parameters

Bellow are the scores (and 95% confidence intervals) obtained by the Grid Search procedure used to estimate the best values for the SVM parameters C and γ in the detection of emotions.

Precision Scores

	C							
	0.001	**0.01**	**0.1**	**1.0**	**10.0**	**100.0**	**1000.0**	**GAMMA (γ)**
Linear Kernel	0.37 (+/-0.04)	0.70 (+/-0.13)	0.71 (+/-0.15)	0.69 (+/-0.15)	0.69 (+/-0.15)	0.69 (+/-0.15)	0.69 (+/-0.15)	
RBF Kernel	0.07 (+/-0.00)	0.07 (+/-0.00)	0.20 (+/-0.08)	0.51 (+/-0.03)	0.78 (+/-0.19)	0.71 (+/-0.23)	0.71 (+/-0.22)	**0.001**
	0.07 (+/-0.00)	0.13 (+/-0.13)	0.21 (+/-0.06)	0.55 (+/-0.12)	0.62 (+/-0.12)	0.62 (+/-0.12)	0.62 (+/-0.12)	**0.01**
	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	0.13 (+/-0.14)	0.17 (+/-0.11)	0.17 (+/-0.11)	0.17 (+/-0.11)	**0.1**
	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	**1.0**
	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	**10.0**
	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	**100.0**
	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	0.07 (+/-0.00)	**1000.0**

Recall Scores

Linear Kernel	0.33 (+/-0.10)	0.51 (+/-0.19)	0.58 (+/-0.23)	0.57 (+/-0.22)	0.57 (+/-0.22)	0.57 (+/-0.22)	0.57 (+/-0.22)	**GAMMA (γ)**
RBF Kernel	0.14 (+/-0.00)	0.14 (+/-0.00)	0.21 (+/-0.10)	0.37 (+/-0.09)	0.55 (+/-0.21)	0.58 (+/-0.23)	0.58 (+/-0.24)	**0.001**
	0.14 (+/-0.00)	0.14 (+/-0.02)	0.22 (+/-0.08)	0.43 (+/-0.11)	0.46 (+/-0.14)	0.46 (+/-0.14)	0.46 (+/-0.14)	**0.01**
	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	**0.1**
	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	**1.0**
	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	**10.0**
	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	**100.0**
	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	0.14 (+/-0.00)	**1000.0**

Annex A

The Game Experience Questionnaire

The Game Experience Questionnaire (GEQ) was developed by IJsselsteijn *et al.* (2013) at the Game Experience Lab. The complete manual is available under request from their website at http://www.gamexplab.nl/. Here it is reproduced just the core module, for easy of reference, with the employed Portuguese translations in parenthesis.

A.1 Title

Please answer the questions bellow considering how you felt about your whole experience playing [game name].

(*Por favor, responda às questões abaixo considerando como você se sentiu sobre toda a sua experiência jogando [nome do jogo]*).

A.2 Likert Scale

not at all	slightly	moderately	fairly	extremely
(*de jeito nenhum*)	(*levemente*)	(*moderadamente*)	(*bastante*)	(*extremamente*)
0	1	2	3	4
< >	< >	< >	< >	< >

A.3 Questions

1. I felt content (*Eu me senti contente*)
2. I felt skilful (*Eu me senti habilidoso(a)*)
3. I was interested in the game's story (*Eu estava interessado(a) na história do jogo*)
4. I though it was fun (*Eu achei que foi divertido*)
5. I was fully occupied with the game (*Eu estava totalmente ocupado(a) com o jogo*)
6. I felt happy (*Eu me senti feliz*)
7. It gave me a bad mood (*O jogo me deixou de mal humor*)
8. I thought about other things (*Eu pensei sobre outras coisas*)

9. I found it tiresome (*Eu achei cansativo*)
10. I felt competent (*Eu me senti competente*)
11. I thought it was hard (*Eu achei que foi difícil*)
12. It was aesthetically pleasing (*O jogo foi esteticamente agradável*)
13. I forgot everything around me (*Eu esqueci de tudo ao meu redor*)
14. I felt good (*Eu me senti bem*)
15. I was good at it (*Eu fui bom(oa) no jogo*)
16. I felt bored (*Eu me senti entediado(a)*)
17. I felt successful (*Eu me senti bem-sucedido(a)*)
18. I felt imaginative (*Eu me senti imaginativo(a)*)
19. I felt that I could explore things (*Eu senti que eu pude explorar as coisas*)
20. I enjoyed it (*Eu gostei do jogo*)
21. I was fast at reaching the game's targets (*Eu fui rápido(a) em alcançar os objetivos do jogo*)
22. I felt annoyed (*Eu me senti irritado(a)*)
23. I felt pressured (*Eu me senti pressionado(a)*)
24. I felt irritable (*Eu me senti irritável*)
25. I lost track of time (*Eu perdi a noção do tempo*)
26. I felt challenged (*Eu me senti desafiado(a)*)
27. I found it impressive (*Eu achei o jogo impressionante*)
28. I was deeply concentrated in the game (*Eu estava profundamente concentrado(a) no jogo*)
29. I felt frustrated (*Eu me senti frustrado(a)*)
30. It felt like a rich experience (*O jogo pareceu uma experiência rica*)
31. I lost connection with the outside world (*Eu perdi a conexão com o mundo lá fora*)
32. I felt time pressure (*Eu senti a pressão do tempo*)
33. I had to put a lot of effort into it (*Eu tive de colocar muito esforço no jogo*)

Bibliography

Ahlberg(2001) Jrgen Ahlberg. CANDIDE-3 - an updated parameterized face. Technical report, Linköping University, Linköping, Sweden. URL http://www.icg.isy.liu.se/candide/main.html. Cited at page 70

Ambinder(2009) Mike Ambinder. Valve's Approach to Playtesting: the Application of Empiricism, 2009. URL http://www.gdcvault.com/play/1566/Valve-s-Approach-to-Playtesting. Cited at page 3

Aouaki(2013) Naoual Aouaki. *Eye blinks as an objective measurement of consumers emotional and motivational attitude towards brands.* master, Erasmus University Rotterdam. Cited at page 41

Baggio *et al.*(2012) Daniel Lélis Baggio, Shervin Emami, David Millán Escrivá, Khvedchenia Ievgen, Naureen Mahmood, Jason M. Saragih and Roy Shilkrot. *Mastering OpenCV with Practical Computer Vision Projects.* Packt Publishing. ISBN 9781849517829. Cited at page 66

Bainbridge *et al.*(2013) Wilma A. Bainbridge, Phillip Isola and Aude Oliva. The Intrinsic Memorability of Face Photographs. *Journal of Experimental Psychology: General*, 142(4):1323–1334. ISSN 1939-2222. doi: 10.1037/a0033872. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/a0033872. Cited at page 90

Baldwin(2011) Scott Baldwin. UX Ideas in the Cards, 2011. URL http://uxmag.com/articles/ux-ideas-in-the-cards. Cited at page 32

Bell(2010) Robert Charles Bell. *Board and Table Games from Many Civilizations.* Dover Publications, Mineola, NY, USA, kindle edition. Cited at page 1

Bentley *et al.*(2002) Todd Bentley, Lorraine J. Johnston and Karola L. Von Baggo. Putting Some Emotion into Requirements Engineering. In *Proceedings of the 7th Australian Workshop on Requirements Engineering*, pages 227–241, Melbourne, Victoria, Australia. Deakin University. Cited at page 24, 25, 29, 43

Berger(2004) Friedrich Berger. From Circle And Square To the Image of the World: A Possible Interpretation for Some Petroglyphs of Merels Boards. *Rock Art Research*, 21(1):11–19. Cited at page 1

Bernhaupt(2010) Regina Bernhaupt. User Experience Evaluation in Entertainment. In Regina Bernhaupt, editor, *Evaluating User Experience in Games: Concepts and Methods*, Human-Computer Interaction Series, chapter 1, pages 3–7. Springer London, London. ISBN 978-1-84882-962-6. doi: 10.1007/978-1-84882-963-3. URL http://link.springer.com/10.1007/978-1-84882-963-3. Cited at page 24, 37, 38

Bettadapura(2012) Vinay Bettadapura. Face Expression Recognition and Analysis : The State of the Art. *Computer Vision and Pattern Recognition*, 1203.6:1–27. Cited at page 4, 46, 74, 90, 129

Bidwell and Fuchs(2011) Jonathan Bidwell and Henry Fuchs. Classroom Analytics: Measuring Student Engagement with Automated Gaze Tracking. Technical report, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. Cited at page 40, 44

Bikić and Vuković(2010) Vesna Bikić and Jasna Vuković. Board Games Reconsidered: Mancala in the Balkans. *Issues in Ethnology and Anthropology*, 5(1):183–209. Cited at page 1

Boecker *et al.*(2015) Lea Boecker, Katja U. Likowski, Paul Pauli and Peter Weyersa. The face of schadenfreude: Differentiation of joy and schadenfreude by electromyography. *Cognition & Emotion*, 29(6):1117–1125. doi: 10.1080/02699931.2014.966063. URL http://www.tandfonline.com/doi/abs/10.1080/02699931.2014.966063. Cited at page 35

Bogart and Ort(2007) Bruce Ian Bogart and Victoria H. Ort. Introduction to the Peripheral Nervous System. In *Integrated Anatomy and Embryology*, chapter 2, pages 10—-21. Mosby Elsevier, 1 edition. Cited at page 20

Bradley and Lang(1994) Margaret M. Bradley and Peter J Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59. ISSN 00057916. doi: 10.1016/0005-7916(94)90063-9. URL http://linkinghub.elsevier.com/retrieve/pii/0005791694900639. Cited at page 44

Branco(2006) Pedro Sérgio Oliveira Branco. *Computer-based Facial Expression Analysis for Assessing User Experience*. Phd, University of Minho. URL https://repositorium.sdum.uminho.pt/bitstream/1822/8457/1/TesedeDoutoramentoPedroBranco.pdf. Cited at page 20, 21

Brown(2010) Emily Brown. The Life and Tools of a Games Designer. In Regina Bernhaupt, editor, *Evaluating User Experience in Games: Concepts and Methods*, Human-Computer Interaction Series, chapter 5, pages 73–87. Springer London, London. ISBN 978-1-84882-962-6. doi: 10.1007/978-1-84882-963-3. URL http://link.springer.com/10.1007/978-1-84882-963-3. Cited at page 2, 31, 38

Brown and Cairns(2004) Emily Brown and Paul Cairns. A Grounded Investigation of Game Immersion. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, pages 1297—-1300. ACM Press, New York, NY, USA. ISBN 1581137036. doi: 10.1145/985921.986048. URL http://portal.acm.org/citation.cfm?doid=985921.986048. Cited at page xv, 16, 17, 18, 24, 43, 129

Caillois(2001) Roger Caillois. *Man, Play and Games*. University of Illinois Press, Champaign, Illinois, USA. Cited at page xix, 15

Calvillo-gámez *et al.*(2010) Eduardo H Calvillo-gámez, Paul Cairns and Anna L Cox. Assessing the Core Elements of the Gaming Experience. In Regina Bernhaupt, editor, *Evaluating User Experience in Games: Concepts and Methods*, Human-Computer Interaction Series, chapter 4, pages 47–71. Springer London, London. ISBN 978-1-84882-962-6. doi: 10.1007/978-1-84882-963-3_4. URL http://link.springer.com/10.1007/978-1-84882-963-3. Cited at page 5, 8, 17, 18, 24, 29, 129

Canossa *et al.*(2011) Alessandro Canossa, Anders Drachen and Janus Rau Møller Sørensen. Arrgghh!!! - Blending Quantitative and Qualitative Methods to Detect Player Frustration. *Proceedings of the 6th International Conference on Foundations of Digital Games - FDG '11*, pages 61–68. doi: 10.1145/2159365.2159374. URL http://dl.acm.org.prox.lib.ncsu.edu/citation.cfm?id=2159365.2159374. Cited at page 24, 94, 129

Carroll(2004) John M. Carroll. Beyond fun. *Interactions*, 11(5):38–40. ISSN 10725520. doi: 10.1145/1015530.1015547. URL http://portal.acm.org/citation.cfm?doid=1015530.1015547. Cited at page 2, 29

Chen(2007a) Jenova Chen. *Flow in Games*. Master thesis, University of Southern California. Cited at page 3, 5, 42, 43

Chen(2007b) Jenova Chen. Flow in games (and everything else). *Communications of the ACM*, 50 (4):31–34. ISSN 00010782. doi: 10.1145/1232743.1232769. URL http://portal.acm.org/citation.cfm?doid=1232743.1232769. Cited at page 2, 3, 26

Chen and Epps(2013) Siyuan Chen and Julien Epps. Automatic classification of eye activity for cognitive load measurement with emotion interference. *Computer methods and programs in biomedicine*, 110(2):111–24. ISSN 1872-7565. doi: 10.1016/j.cmpb.2012.10.021. URL http://www.ncbi.nlm.nih.gov/pubmed/23270963. Cited at page 41, 67

Collins(2002) Michael Collins. Discriminative training methods for hidden Markov models. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, volume 10, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics. doi: 10.3115/1118693.1118694. URL http://portal.acm.org/citation.cfm?doid=1118693.1118694. Cited at page 85

Cootes *et al.*(2001) T. F. Cootes, G. J. Edwards and C. J. Taylor. Active Appearance Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681—-685. ISSN 01628828. doi: 10.1109/34.927467. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=927467. Cited at page 65

Cornelius(2000) Randolph R. Cornelius. Theoretical approaches to emotion. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 3–10, Lous Tourils, France. International Speech Communication Association. Cited at page 19

Cortes and Vapnik(1995) Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297. ISSN 0885-6125. doi: 10.1007/BF00994018. URL http://link.springer.com/10.1007/BF00994018. Cited at page 81, 82

Cox *et al.*(2013) M. Cox, J. Nuevo, Jason M. Saragih and Simon Lucey. CSIRO Face Analysis SDK. In *AFGR 2013*. Cited at page 66

Crowley(2004) James L. Crowley. FGNet Talking Face Video, 2004. URL http://www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html. Cited at page 66, 72

Csikszentmihalyi(1991) Mihaly Csikszentmihalyi. *Flow: The Psychology of Optimal Experience*. Harper Perennial, New York, New York, USA, 1 edition. Cited at page 2, 4, 9, 10, 12, 13

Damásio(2012) António R. Damásio. *O Erro de Descartes: Emoção, Razão e o Cérebro Humano*. Companhia das Letras, São Paulo, SP, Brazil, 3 edition. Cited at page 3, 19, 20, 21, 22

Darwin(1872) Charles Darwin. *The Expression of Emotion in Man and Animals*. John Murray, London, England. Cited at page 19, 45

Daugman(1985) John G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America*, 2(7):1160—-1169. Cited at page 74

Daumé III(2006) Harold Charles Daumé III. *Practical structured learning techniques for natural language processing*. Phd, University of Southern California Los Angeles. Cited at page 85, 87

Deterding *et al.*(2011) Sebastian Deterding, Rilla Khaled, Lennart E. Nacke and Dan Dixon. Gamification: Toward a Definition. In *CHI 2011 Gamification Workshop Proceedings*, pages 12–15, Vancouver. ACM Press. ISBN 9781450302685. Cited at page 15

Dix(2004) Alan Dix. Fun Systematically. In M. Blythe D.J. Reed, G. Baxter, editor, *Twelth European Conference on Cognitive Ergonomics*, pages 9–10, York. URL http://www.comp.lancs.ac.uk/~dixa/papers/ECCE-fun-2004/ecce-alan-fun-panel.pdf. Cited at page 7

Dix *et al.*(2003) Alan Dix, Janet E. Finlay, Gregory D. Abowd and Russell Beale. *Human-Computer Interaction*. Pearson Prentice Hall, Harlow, England, 3 edition. ISBN 9780130461094. Cited at page 3, 38, 39

Douglas and Hargadon(2000) Yellowlees Douglas and Andrew Hargadon. The Pleasure Principle: Immersion, Engagement, Flow. In *Proceedings of the eleventh ACM on Hypertext and hypermedia - HYPERTEXT '00*, pages 153–160, New York, New York, USA. ACM Press. ISBN 1581132271. doi: 10.1145/336296.336354. URL http://portal.acm.org/citation.cfm?doid=336296.336354. Cited at page 16, 18, 129

Dubberly *et al.*(2009) Hugh Dubberly, Paul Pangaro and Usman Haque. What is interaction? Are there different types? *interactions*, 16(1):69–75. ISSN 10725520. doi: 10.1145/1456202.1456220. URL http://portal.acm.org/citation.cfm?doid=1456202.1456220. Cited at page 8, 9

Ekman and Friesen(1971) Paul Ekman and Wallace V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124—-129. URL http://psycnet.apa.org/index.cfm?fa=buy.optionToBuy&id=1971-07999-001. Cited at page 19, 46

Ekman and Friesen(1978) Paul Ekman and Wallace V. Friesen. Facial action coding system: A technique for the measurement of facial movement. *From appraisal to emotion: Differences among unpleasant feelings. Motivation and Emotion.*, 12:271—-302. Cited at page 90

El-nasr and Yan(2006) Magy Seif El-nasr and Su Yan. Visual Attention in 3D Video Games. In *Proceedings of the 2006 ACM SIGCHI international conference on Advances in computer entertainment technology*, pages 1—-7, Hollywood, California, USA. ACM Press. Cited at page 67

Entertainment Software Association(2016) Entertainment Software Association. Essential Facts About Computer and Video Game Industry. Technical report, Entertainment Software Association, Washington, DC, USA. URL http://www.theesa.com/facts/pdfs/esa_ef_2013.pdf. Cited at page 1

Fasel and Luettin(2003) Beat Fasel and Juergen Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1):259–275. ISSN 00313203. doi: 10.1016/S0031-3203(02)00052-3. URL http://linkinghub.elsevier.com/retrieve/pii/S0031320302000523. Cited at page 74, 75

Federoff(2002) Melissa A. Federoff. *Heuristics and Usability Guidelines for the Creation and Evaluation of Fun in Video Games.* Phd, Indiana University. Cited at page 31, 32

Fierley and Engl(2010) Remigius Fierley and Stephan Engl. User Experience Methods and Games: Lessons Learned. In *Proceedings of the 24th BCS Interaction Specialist Group Conference*, pages 204—-210, Swinton, United Kingdom. British Computer Society. Cited at page 3, 29, 38, 39

Forlizzi and Battarbee(2004) Jodi Forlizzi and Katja Battarbee. Understanding experience in interactive systems. In *Proceedings of the 2004 conference on Designing interactive systems processes, practices, methods, and techniques - DIS '04*, pages 261—-268, New York, New York, USA. ACM Press. ISBN 1581137877. doi: 10.1145/1013115.1013152. URL http://portal.acm.org/citation.cfm?doid=1013115.1013152. Cited at page xix, 2, 5, 24, 27, 28

Freund and Schapire(1997) Yoav Freund and Robert E Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139. ISSN 00220000. doi: 10.1006/jcss.1997.1504. URL http://linkinghub.elsevier.com/retrieve/pii/S002200009791504X. Cited at page 64

Fullerton(2008) Tracy Fullerton. *Game Design Workshop: A Playcentric Approach to Creating Innovative Games.* CRC Press, 2 edition. ISBN 9780240809748. Cited at page xv, xix, 3, 5, 8, 15, 16, 24, 37, 38

Garcia and Sichman(2003) Ana Cristina Bicharra Garcia and Jaime Simão Sichman. Agentes e Sistemas Multiagentes. In Solange Oliveira Rezende, editor, *Sistemas Inteligentes: Fundamentos e Aplicações*, chapter 11, pages 269–306. Manole Ltda, Barueri, SP, Brazil, 1 edition. Cited at page 9

González *et al.*(2009) José González, Francisco Montero Simarro, Natalia Padilla Zea and Francisco Luis Gutiérrez Vela. Playability as Extension of Quality in Use in Video Games. In Silvia Mara Abrahão, Kasper Hornbæk, Effie Lai-Chong Law and Jan Stage, editors, *Proceedings of the CEUR Workshop*, pages 1—-6. CEUR-WS.org. Cited at page 29, 30

Gonzalez and Woods(2002) Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Prentice Hall, 2 edition. ISBN 978-0201180756. Cited at page 63

Goodman *et al.*(2012) Elizabeth Goodman, Mike Kuniavsky and Andrea Moed. *Observing the User Experience: A Practitioner's Guide to User Research*. Morgan Kaufmann, 2 edition. ISBN 9780123848697. Cited at page 29

Grigorescu *et al.*(2003) Cosmin Grigorescu, Nicolai Petkov and Michel a Westenberg. Contour detection based on nonclassical receptive field inhibition. *IEEE transactions on image processing*, 12(7):729–739. ISSN 1057-7149. doi: 10.1109/TIP.2003.814250. URL http://www.ncbi.nlm.nih.gov/pubmed/18237948. Cited at page 76

Groh(2012) Fabian Groh. Gamification: State of the Art Definition and Utilization. In Naim Asaj, Könings Bastian, Mark Poguntke, Florian Schaub, Björn Wiedersheim and Michael Weber, editors, *Proceedings of the 4th Seminar on Research Trends in Media Informatics*, pages 39–45. Institute of Media Informatics Ulm University. URL http://vts.uni-ulm.de/docs/2012/7866/vts_7866_11380.pdf. Cited at page xv, 15, 16

Hassenzahl(2004) Marc Hassenzahl. Emotions can be quite ephemeral. We cannot design them. *Interactions*, 11(5):46–48. ISSN 10725520. doi: 10.1145/1015530.1015551. URL http://portal.acm.org/citation.cfm?doid=1015530.1015551. Cited at page 3, 26

Hassenzahl(2005) Marc Hassenzahl. The thing and I: understanding the relationship between user and product. In Peter C. Blythe, Mark A. and Overbeeke, Kees and Monk, Andrew F. and Wright, editor, *Funology*, pages 31–42. Kluwer Academic Publishers, Norwell. URL http://dl.acm.org/citation.cfm?id=1139008.1139015. Cited at page xv, 2, 3, 24, 25, 26, 27

Hassenzahl *et al.*(2010) Marc Hassenzahl, Sarah Diefenbach and Anja Göritz. Needs, affect, and interactive products - Facets of user experience. *Interacting with Computers*, 22(5):353–362. ISSN 09535438. doi: 10.1016/j.intcom.2010.04.002. URL http://linkinghub.elsevier.com/retrieve/pii/S0953543810000366. Cited at page 19

Henriksen(2007) Jesper Juul Henriksen. *3D Tracking of Texture Point for Surface Approximation*. Thesis (PhD), Southern University of Denmark. Cited at page 75, 76

Hiller-Sturmhoefel and Bartke(1998) Susanne Hiller-Sturmhoefel and Andrzej Bartke. The Endocrine System: An Overview. *Alcohol Health & Research World*, 22(3):153—-164. Cited at page 20

Hollows(2011) Gregory Hollows. Distortion, 2011. URL http://www.edmundoptics.com/resources/application-notes/imaging/distortion/. Cited at page 68

Huizinga(2008) Johan Huizinga. *Homo ludens: o jogo como elemento da cultura*. Perspectiva, São Paulo, SP, Brazil, 5 edition. Cited at page 2

Hunicke and Chapman(2004) Robin Hunicke and Vernell Chapman. AI for Dynamic Difficulty Adjustment in Games. In *Proceedings of the Challenges in Game Artificial Intelligence AAAI Workshop*, pages 91—-96. AAAI Press. Cited at page 5, 42

Hunicke *et al.*(2004) Robin Hunicke, M LeBlanc and Robert Zubek. MDA: A formal approach to game design and game research. In *Proceedings of the Challenges in Games AI Workshop, Nineteenth National Conference of Artificial Intelligence*, pages 1–5. URL http://www.aaai.org/Papers/Workshops/2004/WS-04-04/WS04-04-001.pdf. Cited at page xix, 33

IJsselsteijn *et al.*(2013) W. A. IJsselsteijn, K. Poels and Y. A. W. de Kort. The Game Experience Questionnaire: Development of a self-report measure to assess player experiences of digital games. Technical report, Technische Universiteit Eindhoven, Eindhoven, Netherlands. Cited at page 43, 52, 139

Isbister(2006) Katherine Isbister. *Better Game Characters by Design: A Psychological Approach*. Morgan Kaufmann Publishers, San Francisco, USA. ISBN 9781558609211. Cited at page 40, 46

Isbister(2010) Katherine Isbister. Enabling Social Play: A Framework for Design and Evaluation. In Regina Bernhaupt, editor, *Evaluating User Experience in Games: Concepts and Methods*, Human-Computer Interaction Series, chapter 2, pages 11–22. Springer London, London. ISBN 978-1-84882-962-6. doi: 10.1007/978-1-84882-963-3. URL http://link.springer.com/10.1007/978-1-84882-963-3. Cited at page 3, 15, 39

ISO(2010) International Organization for Standardization ISO. ISO 9241-210:2010. Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems, 2010. URL http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=52075. Cited at page 26

Jääskö and Mattelmäki(2003) Vesa Jääskö and Tuuli Mattelmäki. Observing and probing. In *Proceedings of the 2003 international conference on Designing pleasurable products and interfaces - DPPI '03*, pages 126—-131, New York, New York, USA. ACM Press. ISBN 1581136528. doi: 10.1145/782924.782927. URL http://portal.acm.org/citation.cfm?doid=782896.782927. Cited at page 3, 26

Jackson and Marsh(1996) Susan A. Jackson and Herbert W. Marsh. Development and Validation of a Scale to Measure Optimal Experience: The Flow State Scale. *Journal of Sport & Exercise Psychology*, 18(1):17–19. Cited at page 42

James(1884) William James. What is an Emotion? *Mind*, 9(34):188—-205. Cited at page 19

Jennett *et al.*(2008) Charlene Jennett, Anna L. Cox, Paul Cairns, Samira Dhoparee, Andrew Epps, Tim Tijs and Alison Walton. Measuring and Defining the Experience of Immersion in Games. *International Journal of Human-Computer Studies*, 66(9):641–661. ISSN 10715819. doi: 10.1016/j.ijhcs.2008.04.004. URL http://linkinghub.elsevier.com/retrieve/pii/S1071581908000499. Cited at page 16, 18, 43, 129

Jordan(2002) Patrick W. Jordan. *Designing Pleasurable Products*. Taylor & Francis, Philadelphia, PA, USA. Cited at page 24, 25

Kanade *et al.*(2000) Takeo Kanade, Jeffrey F. Cohn and Yingli Tian. Comprehensive Database for Facial Expression Analysis. *FG*, pages 46—-53. Cited at page 90

Kazemi and Sullivan(2014) Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1867–1874. ISSN 10636919. doi: 10.1109/CVPR.2014.241. Cited at page 66

King(2009) Davis E. King. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, 10:1755–1758. URL http://jmlr.csail.mit.edu/papers/volume10/king09a/king09a.pdf. Cited at page 66

Klein *et al.*(2002) J. Klein, Y. Moon and R.W. Picard. This computer responds to user frustration: Theory, design, and results. *Interacting with Computers*, 14(2):119–140. ISSN 09535438. doi: 10.1016/S0953-5438(01)00053-4. Cited at page 94

Koster(2010) Raph Koster. *Theory of fun for game design*. O'Reilly Media Inc., Sebastopol, CA, USA, 1 edition. Cited at page 4, 9, 129

Koštomaj and Boh(2009) Mitja Koštomaj and Bojana Boh. Evaluation of User's Physical Experience in Full Body Interactive Games. *Haptic and Audio Interaction Design*, 5763:145–154. Cited at page 4

Laeng *et al.*(2012) B. Laeng, S. Sirois and G. Gredeback. Pupillometry: A Window to the Preconscious? *Perspectives on Psychological Science*, 7(1):18—-27. ISSN 1745-6916. doi: 10.1177/1745691611427305. URL http://pps.sagepub.com/lookup/doi/10.1177/1745691611427305. Cited at page 41

Lamar and Raz(2007) Melissa Lamar and Amir Raz. Neuropsychological Assessment of Attention and Executive Functioning. In Susan Ayers, Andrew Baum, Chris McManus, Stanton Newman, Kenneth Wallston, John Weinman and Robert West, editors, *Cambridge Handbook of Psychology, Health and Medicine*, pages 290—-294. Cambridge University Press, Cambridge, England, 2 edition. Cited at page 40

Lang *et al.*(2005) Peter J. Lang, Margaret M. Bradley and Bruce N. Cuthbert. International Affective Picture System (IAPS): Affective ratings of pictures and instruction manual. Technical report, University of Florida, Gainesville, Florida, USA. Cited at page 41

Lazzaro(2004) Nicole Lazzaro. Why We Play Games: Four Keys to More Emotion Without Story, 2004. Cited at page xix, 5, 33, 34

Lazzaro(2008) Nicole Lazzaro. Why We Play Games. In Tracy Fullerton, editor, *Game Design Workshop: A Playcentric Approach to Creating Innovative Games*, pages 258—-260. CRC Press, 2 edition. Cited at page 5, 34

Lazzaro(2010) Nicole Lazzaro. The future of ux is play: The 4 keys to fun, emotion, and user engagement, 2010. URL http://www.adaptivepath.com/ideas/nicole-lazzaro/. Cited at page 5, 22, 34, 35, 129

Levialdi *et al.*(2007) Stefano Levialdi, Alessio Malizia, Teresa Onorati, Enver Sangineto and Nicu Sebe. Detecting attention through Telepresence. In Laura Moreno, editor, *he 10th Annual International Workshop on Presence - PRESENCE 2007*, pages 233—-236, Barcelona, Spain. Starlab Barcelona S.L. Cited at page 40

Litman(2005) Jordan A. Litman. Curiosity and the pleasures of learning: Wanting and liking new information. *Cognition & Emotion*, 19(6):793–814. ISSN 0269-9931. doi: 10.1080/02699930541000101. URL http://www.tandfonline.com/doi/abs/10.1080/02699930541000101. Cited at page 2, 13

Lorena and de Carvalho(2007) Ana C. Lorena and A.C.P.L.F. de Carvalho. Uma Introdução às Support Vector Machines. *Revista de Informática Teórica e Aplicada*, 14(2):43–67. ISSN 00978418. doi: 10.1145/268085.268132. URL http://seer.ufrgs.br/index.php/rita/article/viewArticle/rita_v14_n2_p43-67. Cited at page 81, 83

Lucas and Kanade(1981) Bruce D Lucas and Takeo Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, volume 2, pages 674–679, Vancouver, Canada. ISBN 0769521754. doi: 10.1109/HPDC.2004.1323531. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.49.2019&rep=rep1&type=pdf. Cited at page 65, 71

Lucey *et al.*(2010) Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason M. Saragih, Zara Ambadar and Iain Matthews. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2010)*, pages 94–101, San Francisco, USA. IEEE. ISBN 978-1-4244-7029-7. doi: 10.1109/CVPRW.2010.5543262. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5543262. Cited at page 90

Lyons *et al.*(1998) Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi and Jiro Gyoba. Coding Facial Expressions with Gabor Wavelets. In *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205, Nara, Japan. IEEE. Cited at page 75, 78

Maitland(2010) Margaret Maitland. 'It's not just a game, it's a religion': games in ancient Egypt, 2010. URL http://www.eloquentpeasant.com/2010/10/14/its-not-just-a-game-its-a-religion-games-in-ancient-egypt/. Cited at page 2

Mallick(2016) Satya Mallick. Head Pose Estimation using OpenCV and Dlib, 2016. URL http://www.learnopencv.com/head-pose-estimation-using-opencv-and-dlib/. Cited at page xvi, 67, 68, 69, 70

Malone(1980a) Thomas W. Malone. *What Makes Things Fun to Learn. A Study of Intrinsically Motivating Computer Games.* Phd thesis, Stanford University. URL http://cci.mit.edu/malone/tmstudy144.pdf. Cited at page 2, 13, 14, 15, 18

Malone(1980b) Thomas W. Malone. What makes things fun to learn? heuristics for designing instructional computer games. In *Proceedings of the 3rd ACM SIGSMALL symposium and the first SIGPC symposium on Small systems - SIGSMALL '80*, volume 162, pages 162–169, New York, New York, USA. ACM Press. ISBN 0897910249. doi: 10.1145/800088.802839. URL http://portal.acm.org/citation.cfm?doid=800088.802839. Cited at page 2, 14, 31

Malone and Lepper(1987) Thomas W. Malone and Mark R. Lepper. Making Learning Fun: A Taxonomy of Intrinsic Motivations for Learning. In Richard E. Snow and Marshall J. Farr, editors, *Aptitude, Learning and Instruction - Volume 3: Conative and Affective Process Analysis*, pages 223—-253. Lawrence Erlbaum, London, England. Cited at page 31

Mandryk *et al.*(2006) Regan L. Mandryk, Kori M. Inkpen and Thomas W. Calvert. Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & Information Technology*, 25(2):141–158. ISSN 0144-929X. doi: 10.1080/01449290500331156. URL http://www.tandfonline.com/doi/abs/10.1080/01449290500331156. Cited at page 3, 39

Marcus(2007) Aaron Marcus. Fun! fun! fun! in the user experience we just wanna have fun...don't we? *interactions*, 14(4):48–55. ISSN 10725520. doi: 10.1145/1273961.1273988. URL http://portal.acm.org/citation.cfm?doid=1273961.1273988. Cited at page 29

Maslow(1943) Abraham H. Maslow. A theory of human motivation. *Psychological Review*, 50(4):411. Cited at page 25

Matsumoto and Hwang(2011) David Matsumoto and Hyi Sung Hwang. Reading facial expressions of emotion, 2011. URL http://www.apa.org/science/about/psa/2011/05/facial-expressions.aspx. Cited at page 4, 46, 129

Mauss and Robinson(2009) Iris B. Mauss and Michael D. Robinson. Measures of emotion: A review. *Cognition & Emotion*, 23(2):209–237. ISSN 1464-0600. doi: 10.1080/02699930802204677. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2756702&tool=pmcentrez&rendertype=abstract. Cited at page 3, 39, 45, 46

Mäyrä and Ermi(2011) Frans Mäyrä and Laura Ermi. Fundamental Components of the Gameplay Experience: Analysing Immersion. In Stephan Günzel, Michael Liebe and Dieter Mersch, editors, *DIGAREC Keynote-Lectures*, pages 88—-115. Potsdam University Press, Potsdam, Germany. Cited at page 18, 129

McCarthy and Wright(2004) John McCarthy and Peter Wright. *Technology as Experience.* The MIT Press, London, England. ISBN 0262134470. Cited at page 19, 24, 26

McDonald(1996) John K. McDonald. *House of Eternity: The Tomb of Nefertari*. Thames & Hudson Ltd, London, England. URL http://www.getty.edu/conservation/publications_resources/pdf_publications/pdf/house_eternity1.pdf. Cited at page xv, 2

McDuff *et al.***(2013)** Daniel McDuff, Rana el Kaliouby, David Demirdjian and Rosalind W. Picard. Predicting Online Media Effectiveness Based on Smile Responses Gathered Over the Internet. *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1—-7. doi: 10.1109/FG.2013.6553750. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6553750. Cited at page 47

McKinsey & Company(2015) McKinsey & Company. Global Media Report 2015 - Global Industry Overview. Technical report, McKinsey & Company, London, England. URL http://www.mckinsey.com/industries/media-and-entertainment/our-insights/global-media-report-2015. Cited at page 1

Mello and Perani(2012) Vinícius Mello and Letícia Perani. Gameplay x playability: defining concepts, tracing differences. In Maria das Graças Chagas and Tiago Barros Pontes e Silva, editors, *Proceedings of the SBGames 2012*, pages 157–164, Brasilia, Brazil. SBC. Cited at page 30

Michael and Chang(2013) David Michael and Jordan Chang. Dynamic Difficulty Adjustment in Computer Games. In *Proceedings of the Interactive Multimedia Conference*, pages 1—-6. University of Southampton. Cited at page 42

Miller(1980) David Miller. Long-term trends in human eye blink rate. *Survey of Ophthalmology*, 24(6):649. ISSN 00396257. doi: 10.1016/0039-6257(80)90131-9. URL http://linkinghub.elsevier.com/retrieve/pii/0039625780901319. Cited at page 112

Morkes *et al.***(1999)** John Morkes, Hadyn Kernal and Clifford Nass. Effects of Humor in Task-Oriented Human-Computer Interaction and Computer-Mediated Communication: A Direct Test of SRCT Theory. *Human-Computer Interaction*, 14(4):395–435. ISSN 0737-0024. doi: 10.1207/S15327051HCI1404_2. URL http://www.informaworld.com/openurl?genre=article&doi=10.1207/S15327051HCI1404_2&magic=crossref{%}7C{%}7CD404A21C5BB053405B1A640AFFD44AE3. Cited at page 29

Nacke(2013) Lennart E. Nacke. An Introduction to Physiological Player Metrics for Evaluating Games. In Magy Seif El-Nasr, Anders Drachen and Alessandro Canossa, editors, *Game Analytics: Maximizing the Value of Player Data*, pages 585–619. Springer London, London, England. ISBN 978-1-4471-4768-8. doi: 10.1007/978-1-4471-4769-5_26. URL http://link.springer.com/10.1007/978-1-4471-4769-5. Cited at page 3, 5, 39, 45, 46

Nacke and Lindley(2008) Lennart E. Nacke and Craig A. Lindley. Flow and Immersion in First-person Shooters: Measuring the Player's Gameplay Experience. In *Proceedings of the 2008 Conference on Future Play: Research, Play, Share - Future Play '08*, pages 81—-88, New York, New York, USA. ACM Press. ISBN 9781605582184. doi: 10.1145/1496984.1496998. URL http://portal.acm.org/citation.cfm?doid=1496984.1496998. Cited at page 18, 43

Nacke *et al.***(2008)** Lennart E. Nacke, Craig Lindley and Sophie Stellmach. Log Who's Playing: Psychophysiological Game Analysis Made Easy through Event Logging. *Fun and Games*, 5294:150–157. ISSN 0302-9743. doi: 10.1007/978-3-540-88322-7_15. URL http://link.springer.com/10.1007/978-3-540-88322-7. Cited at page 39, 129

Nacke *et al.***(2009)** Lennart E. Nacke, Anders Drachen, Kai Kuikkaniemi and Yvonne A W Kort. Playability and Player Experience Research. In *Proceedings of DiGRA 2009*, pages 1—-5. Authors & Digital Games Research Association. Cited at page 30, 38, 39

Nakamura and Csikszentmihalyi(2001) Jeanne Nakamura and Mihaly Csikszentmihalyi. The Concept of Flow. In C.R. Snyder and Shane.J. Lopez, editors, *Handbook of Positive Psychology*, pages 89–105. Oxford University Press, Oxford, United Kingdom. Cited at page xv, 10, 12

Niedenthal(2009) Simon Niedenthal. What We Talk About When We Talk About Game Aesthetics. In Tanya Krzywinska, Helen Kennedy and Barry Atkins, editors, *Proceedings of DiGRA 2009*, pages 1—-9, London, England. Digital Games Research Association. Cited at page 29

Nielsen(1994) Jakob Nielsen. *Usability Inspection Methods*. John Wiley & Sons, New York, New York, USA, 1 edition. Cited at page 31

Norman(1999) Donald A. Norman. Affordance, conventions, and design. *interactions*, 6(3):38–43. ISSN 10725520. doi: 10.1145/301153.301168. URL http://portal.acm.org/citation.cfm?doid=301153.301168. Cited at page 8

Norman(2002) Donald A. Norman. *The Design of Everyday Things*. Basic Books, New York, NY, USA, 2 edition. Cited at page 8, 14

Norman(2005) Donald A. Norman. *Emotional Design: Why We Love (or Hate) Everyday Things*. Basic Books, New York, New York, USA, 1 edition. ISBN 0465051359. Cited at page xv, 21, 22, 27

Olatsek(2013) Patrik P. Olatsek. Eye Blink Detection. In *Proceedings of the IIT.SRC 2013*, pages 1—-8. Slovak University of Technology in Bratislava. Cited at page 71, 73

Oster and Ekman(1978) Harriet Oster and Paul Ekman. Facial Behavior in Child Development. In W. Andrew Collings, editor, *Minnesota Symposia on Child Psychologic*, chapter 6, pages 231–276. Lawrence Erlbaum, Hillsdale, New Jersey, USA. Cited at page 52

Piaget(1952) Jean Piaget. *The Origins of Intelligence in Children*. International Universities Press, New York, NY, USA. Cited at page 13

Picard(1995) Rosalind W. Picard. Affective computing. Technical Report 321, MIT Media Laboratory, Cambridge. Cited at page 3, 19, 20

Piccione(1980) Peter A. Piccione. In Search of the Meaning of Senet. *Archaeology*, 33(33):55–58. Cited at page 1

Platt(1999) John C. Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*, pages 61—-74. MIT Press. Cited at page 84

Plutchik(2001) Robert Plutchik. The Nature of Emotions. *American Scientist*, 89(4):344–350. ISSN 0003-0996. doi: 10.1511/2001.4.344. URL http://www.americanscientist.org/issues/feature/2001/4/the-nature-of-emotions. Cited at page xix, 19, 22, 23

Preece *et al.*(2002) Jenny Preece, Yvonne Rogers and Helen Sharp. *Interaction Design: Beyond Human - Computer Interaction*. John Wiley & Sons, Inc, Chichester, United Kingdom, 1 edition. Cited at page 8, 24, 25, 29, 31, 37, 38, 39

Prensky(2001) Marc Prensky. Fun, Play and Games: What Makes Games Engaging. In *Digital Game-Based Learning*, chapter 5, pages 105–226. Paragon House. Cited at page 12, 19

Ravaja *et al.*(2006) Niklas Ravaja, Timo Saari, Mikko Salminen, Jari Laarni and Kari Kallinen. Phasic Emotional Reactions to Video Game Events: A Psychophysiological Investigation. *Media Psychology*, 8(4):343–367. ISSN 1521-3269. doi: 10.1207/s1532785xmep0804_2. URL http://www.tandfonline.com/doi/abs/10.1207/s1532785xmep0804_2. Cited at page 23

Read *et al.***(2002)** Janet Read, Stuart Macfarlane and Chris Casey. Endurability, Engagement and Expectations: Measuring Children's Fun. In *Interaction Design and Children*, pages 189—-198. Shaker Publishing. Cited at page 43

Rollefson(1992) Gary O. Rollefson. A Neolithic Game Board from 'Ain Ghazal, Jordan. *Bulletin of the American Schools of Oriental Research*, 1(286):1–5. URL http://www.jstor.org/stable/1357113. Cited at page 1

Rosenblatt(1957) F. Rosenblatt. The Perceptron: A Perceiving and Recognizing Automaton. Technical report, Cornell Aeronautical Laboratory, New York. Cited at page 85

Roshani(2011) Shila Roshani. *The effect of ocular surface conditions on blink rate and completeness*. Master, Queensland University of Technology. Cited at page 71

Russell and Norvig(2010) Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 edition. ISBN 9780136042594. Cited at page 88, 89

Sackett *et al.***(2010)** Aaron M Sackett, Tom Meyvis, Leif D Nelson, Benjamin A Converse and Anna L Sackett. You're having fun when time flies: the hedonic consequences of subjective time progression. *Psychological science*, 21(1):111–117. ISSN 1467-9280. doi: 10.1177/0956797609354832. URL http://www.ncbi.nlm.nih.gov/pubmed/20424031. Cited at page 12

Saragih *et al.***(2011)** Jason M. Saragih, Simon Lucey and Jeffrey F. Cohn. Deformable Model Fitting by Regularized Landmark Mean-Shift. *International Journal of Computer Vision*, 91(2): 200–215. ISSN 0920-5691. doi: 10.1007/s11263-010-0380-4. URL http://link.springer.com/10.1007/s11263-010-0380-4. Cited at page 66

Schell(2008) Jesse Schell. *The Art of Game Design: A book of lenses*. CRC Press, 1 edition. ISBN 9780123694966. Cited at page xv, 3, 29, 30, 32, 33, 38, 39

Scherer(2005) Klaus R. Scherer. What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729. ISSN 0539-0184. doi: 10.1177/0539018405058216. URL http://ssi.sagepub.com/cgi/doi/10.1177/0539018405058216. Cited at page xv, xix, 3, 5, 19, 20, 21, 22, 23, 24, 39, 45, 129, 132

Schindler *et al.***(2008)** Konrad Schindler, Luc Van Gool and Beatrice de Gelder. Recognizing emotions expressed by body pose: a biologically inspired neural model. *Neural networks : the official journal of the International Neural Network Society*, 21(9):1238—-1246. ISSN 0893-6080. doi: 10.1016/j.neunet.2008.05.003. URL http://www.ncbi.nlm.nih.gov/pubmed/18585892. Cited at page 46, 76

Schmidhuber(2010) Jürgen Schmidhuber. Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990-2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247. ISSN 1943-0604. doi: 10.1109/TAMD.2010.2056368. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5508364. Cited at page 9, 13

Schuytema(2008) Paul Schuytema. *Design de Games: Uma abordagem prática*. Cengage Learning, São Paulo, SP, Brazil. Cited at page 3

Shaker *et al.***(2010)** Noor Shaker, Georgios N. Yannakakis and Julian Togelius. Towards Automatic Personalized Content Generation for Platform Games. In *Proceedings of the Sixth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, pages 63—-68. AAAI Press. Cited at page 42

Sokolova *et al.***(2006)** Marina Sokolova, Nathalie Japkowicz and Stan Szpakowicz. Beyond accuracy, F-Score and ROC: A family of discriminant measures for performance evaluation. *Advances in Artificial Intelligence*, 4304(c):1015–1021. ISSN 0302-9743. doi: 10.1007/11941439_

114. URL http://link.springer.com/10.1007/11941439_114{%}5Cnhttp://dx.doi.org/10.1007/11941439_114. Cited at page 88, 89

Spielberger *et al.***(1970)** Charles D. Spielberger, R. L. Gorsuch, R. Lushene, P. R. Vagg and G.A. Jacobs. Manual for the State-Trait Anxiety Inventory (Self Evaluation Questionnaire), 1970. Cited at page 44, 45

Spronck *et al.***(2004)** Pieter Spronck, Ida Sprinkhuizen-Kuyper and Eric Postma. Difficulty Scaling of Game AI. In Abdennour El Rhalibi and Danny Van Welden, editors, *Proceedings of the 5th International Conference on Intelligent Games and Simulation*, pages 33—-37. EUROSIS. Cited at page 42

Stanley(2013) Darren Stanley. *Measuring Attention using Microsoft Kinect.* master, Rochester Institute of Technology. Cited at page 40

Sweetser and Wyeth(2005) Penelope Sweetser and Peta Wyeth. GameFlow: a model for evaluating player enjoyment in games. *Computers in Entertainment*, 3(3):3. ISSN 15443574. doi: 10.1145/1077246.1077253. URL http://portal.acm.org/citation.cfm?doid=1077246.1077253. Cited at page 2, 8, 32

Tan *et al.***(2012)** Chek Tien Tan, Daniel Rosser, Sander Bakkes and Yusuf Pisan. A feasibility study in using facial expressions analysis to evaluate player experiences. *Proceedings of The 8th Australasian Conference on Interactive Entertainment Playing the System - IE '12*, pages 1–10. doi: 10.1145/2336727.2336732. URL http://dl.acm.org/citation.cfm?doid=2336727.2336732. Cited at page 4, 39, 47

Theodoridis and Koutroumbas(2008) Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern Recognition.* Academic Press, 4 edition. ISBN 9781597492720. Cited at page 79, 81, 86, 87

Tsai(1993) D M Tsai. Optimal Gabor filter design for texture segmentation. *ICASSP-93., 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 5:37–40. Cited at page 75

Valiant(1995) Leslie G. Valiant. Rationality. In *Proceedings of the eighth annual conference on Computational learning theory - COLT '95*, pages 3–14, New York, New York, USA. ACM Press. ISBN 0897917235. doi: 10.1145/225298.225299. URL http://portal.acm.org/citation.cfm?doid=225298.225299. Cited at page 8

Viola and Jones(2001) P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 1:511—-518. doi: 10.1109/CVPR.2001.990517. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=990517. Cited at page 64

Viterbi(1967) Andrew J. Viterbi. Error bounds for convolution codes an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269. Cited at page 87

Wagner(1994) Ellen D. Wagner. In support of a functional definition of interaction. *American Journal of Distance Education*, 8(2):6–29. URL http://dx.doi.org/10.1080/08923649409526852. Cited at page 8

Walla *et al.***(2011)** Peter Walla, Gerhard Brenner and Monika Koller. Objective measures of emotion related to brand attitude: a new way to quantify emotion-related aspects relevant to marketing. *PloS one*, 6(11):1—-7. ISSN 1932-6203. doi: 10.1371/journal.pone.0026782. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3206826&tool=pmcentrez&rendertype=abstract. Cited at page 41

Watson *et al.***(1988)** D. Watson, L. A. Clark and A. Tellegen. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6):1063—-1070. ISSN 0022-3514. URL http://www.ncbi.nlm.nih.gov/pubmed/3397865. Cited at page 44

Watson and Clark(1994) David Watson and Lee Anna Clark. THE PANAS-X: Manual for the Positive and Negative Affect Schedule - Expanded Form, 1994. Cited at page 44

Weinberg and Marazita(2010) Seth Weinberg and Mary Marazita. 3D Facial Norms Database, 2010. URL https://www.facebase.org/facial_norms/summary/. Cited at page 77

Wood *et al.***(2004)** Richard T. A. Wood, Mark D. Griffiths and Virginia Eatough. Online data collection from video game players: methodological issues. *CyberPsychology & Behavior*, 7(5):511–518. doi: 10.1089/cpb.2004.7.511. URL http://online.liebertpub.com/doi/pdf/10.1089/cpb.2004.7.511. Cited at page 39

Yannakakis *et al.***(2007)** Georgios N. Yannakakis, John Hallam and Henrik Hautop Lund. Entertainment capture through heart rate activity in physical interactive playgrounds. *User Modeling and User-Adapted Interaction*, 18(1-2):207–243. ISSN 0924-1868. doi: 10.1007/s11257-007-9036-7. URL http://link.springer.com/10.1007/s11257-007-9036-7. Cited at page 46

Yizong Cheng(1995) Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799. ISSN 01628828. doi: 10.1109/34.400568. URL http://ieeexplore.ieee.org/document/400568/. Cited at page 65

Zhang *et al.***(1998)** Zhengyou Zhang, Michael Lyons, Michael Schuster and Shigeru Akamatsu. Comparison Between Geometry-Based and Gabor-Wavelets-Based Facial Expression Recognition Using Multi-Layer Perceptron. In *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 454–459, Nara, Japan. IEEE. ISBN 0-8186-8344-9. doi: 10.1109/AFGR.1998.670990. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=670990. Cited at page 74

Zhao(2014) Kai Zhao. Structured Prediction with Perceptron : Theory and Algorithms. Technical Report November, City University of New York. URL https://pdfs.semanticscholar.org/8f5c/5b3633195a51e0b6ac664b3073a816769145.pdf. Cited at page 85, 86, 87