FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Affective Game Design: Creating Better Game Experiences based on Players' Affective Reaction Model

Rúben Pinto Aguiar

Mestrado Integrado em Engenharia Informática e Computação
Supervisor: Prof. Rui Rodrigues (PhD)
Co-Supervisor: Pedro Nogueira (MSc)

Junho de 2014

Affective Game Design:
Creating Better Game Experiences based on Players' Affective Reaction Model

Rúben Pinto Aguiar

Mestrado Integrado em Engenharia Informática e Computação

Aprovado em provas públicas pelo Júri:

Presidente: António Augusto de Sousa (Prof. Associado)

Vogal Externo: Pedro Miguel do Vale Moreira (Prof. Adjunto)

Orientador: Rui Pedro Amaral Rodrigues (Prof. Auxiliar Convidado)

____

20 de Julho de 2014

Abstract

Current industry approaches to game design improvements rely on gameplay testing, an iterative process that follows a *test, try and fix pattern*. This process has its foundation on target audience feedback, obtained via standard questionnaires. Because of its nature, it is a highly subjective and time consuming stage. In this work, a generalizable approach for building predictive models of players' affective reactions is presented, allowing a more precise tuning of game parameters in order to increase the players' gaming experience. This method aims to be used across a wide range of games and genres.

Two high-level distinct goals are targeted. First, to allow game developers the usage of these affective reaction models to more accurately and easily predict players' emotional responses, aiming to augment players' gaming experiences. Lastly, to provide the capability of using these models as a basis for adaptive and parameterisable affective gaming.

The work presented describes a novel, physiological-based method for profiling players' emotions. Three main phases exist: creation of more accurate affective reaction models based on non-diffuse metrics, exploration of the existent correlation between the biofeedback affective data and the subjective experience, and a mechanism for adapting level design parameters to a desired gaming experience.

The usage of biofeedback to create players' affective reaction models and their posterior use to adapt game design to the desired gaming experience are intended to be a proof of concept applicable in several other domains and problems.

Resumo

Atualmente, a abordagem industrial corrente para melhorar o design de jogo baseia-se em testes de jogabilidade, uma fase iterativa que segue o padrão de *testar*, *experimentar* e *corrigir*. Este processo baseia-se no retorno obtido da audiência alvo através de questionários standardes. Neste trabalho é apresentada uma generalista de construir modelos predictivos da resposta afectiva dos jogadores. Este método tem como objectivo ser usado numa vasta gama de jogos e géneros.

É pretendido atingir dois grandes objectivos. Primeiro, dar aos desenvolvedores de jogos a possibilidade de usar estes modelos de reacção afectiva para mais eficientemente e facilmente prever as reacções emocionais dos jogadores, com o intuito de exponenciar a experiência de jogo do jogador. Por último, providenciar a capacidade de usar estes modelos como base para jogos afectivos adaptativos e parametrizáveis.

O trabalho apresentado descreve um novo método, baseado em dados fisiológicos para fazer o *profiling* emocional dos jogadores. Este processo encontra-se dividido em diversas fases: a criação de modelos afectivos baseados em métricas não difusas mais fiáveis, exploração das relações existentes entre os dados afectivos provenientes de *biofeedback* e a experiência subjectiva, e um mecanismo para adaptar os parâmetros de design dos níveis para uma experiência emocional desejada.

O uso de *biofeedback* para criação dos modelos de reacção afectiva e o seu posterior uso para adaptar o design do jogo para a experiência de jogo desejada têm como objectivo ser provas de conceito aplicáveis em vários outros domínios e problemas.

Acknowledgements

I would like to thank everyone that has helped me one way or the other through this long road. First, a special thanks to Pedro Nogueira for putting up with me over the course of this work. Then the biggest of thanks to all my family. My mother, father, brother, grandmother and everyone that has known me since I've walked in this world. They have supported me over the years and made me who I am.

To all my friends, whether they are in Germany or Portugal, whether they're human or animals, whether they're here or left too early, thank you for making all my years so far worth it. Thank you for the Doto carry, thank you for the personal support and most of all thank you for all the laughs. Thank you 09.

Rúben Pinto Aguiar

“*Stop Feed*”

Fuso

Contents

List of Figures

List of Tables

Abbreviations

Ai	Initial Arousal - E$\{r_A\}$
BF	BestFirst
BVP	Blood Volume Pressure
dA	Delta Arousal - ∇A
dV	Delta Valence - ∇V
ECG	Electroencardiography
EMG	Electromyography
ER-IBF	Emotional Regulated Indirect Biofeedback
GODx	Game Optimal Design Experience
GS	Genetic Search
IBF	Indirect Biofeedback
LFS	Linear Forward Search
LR	Linear Regression
M5P	M5 Model Trees and Rules
MLP	Multilayer Perceptron
NBF	Non Biofeedback
NN	Neural Network
PV	Parameter Values
RS	Random Search
RSP	Respiration
Vi	Initial Valence - E$\{r_V\}$
SC	Skin Conductance
SD	Standard Deviation

Chapter 1

Introduction

Over the years, videogames have propelled innumerous breakthroughs in various fields - computer graphics, artificial intelligence, interaction techniques, physics simulation to name a few. These advances arose from the need of a more realistic game experience, reflected itself on better scenarios, more believable artificial behavior, incredibly realistic audio-visual effects and several other factors that bring to the player an ever growing level of immersion.

A wide range of emotions can arise from the act of playing a game. Players may become sad with a beloved character's death, relieved with the ending of a confrontation, scared with the sound of a distant creature or even frustrated with repeated defeat. This subjective experience has its inception on the game designers that aim to convey to the player these desired emotions and experiences through the act of playing a game.

1.1 Context

Nowadays, gaming industry has been slowing shifting its focus from the technological department, and invested its resources on underexplored areas of gameplay experience. The search for the reasons that lead people to play (Ryan, Rigby, & Przybylski, 2006) and why it is a pleasurable experience (Ermi & Mäyrä, 2005) are subjects that have been vastly studied over the years. A converging thought has been presented many times: video games must provide an engrossing experience, taking the player from the real world and plunging him into the virtual world. Understanding and improving on this immersion is the key to produce better gaming experiences.

Although a fuzzy subject, immersion has been vastly referred to as the degree of envelopment the user has established with the virtual world. How "detached" he has become from the real world and believes he is in the virtual world (Jennett et al., 2008). How captivating a particular challenge is and how much emotion certain events arouse in the player.

1.2 Motivation and Objectives

The search for better levels of user experience (UX), lead to the potential use of Affective Computing and the detection of emotions on players. Ways to use these perceived emotions in-game to make the best possible experience to the player is of utmost relevance. The prediction of players' behavior and emotional reaction can give game developers the tools to create much more immersive and entertaining experiences. It provides a way of assessing if the user experience of the target audience is the one game designers intended when creating the game. Even adaptive games can be vastly improved if the content generated in real-time takes into account the current emotional state of the player.

The primary objectives of this work are: obtain a dataset for the extraction of emotional results, Creation of players' affective reaction models to a pre-determined number of events and the creation of a medium/high level simulator where the previously created models are used to identify the ideal game parameters (possible incorporation of real-time mechanisms).

1.3 Dissertation Structure

Beyond this introduction, this dissertation contains 5 more chapters. In chapter 2, the state of the art is described and related works are presented. In chapter 3, the affective reaction models are constructed and validated. In chapter 4, a possible correlation between physiological/demographic data and reported game experience. Chapter 5 consists the global scope of the work is detailed and explained. The last chapter presents a global overview and some conclusions of the work done.

Chapter 2

State of the Art

In this chapter, the state of the art is described and related works are presented in order to showcase what exists in the same domain and what are the problems faced. Section 2.1 will present current methods of emotional recognition through psychophysiological data. Subsequently, several works regarding the modelling of players' experience are presented. Afterwards, the subject of affective gaming is discussed, in order to show how to augment players' gaming experiences. We conclude the chapter with some final remarks.

2.1 Psychophysiological Emotion Detection

Recognizing human emotions through the study of physiological data is a subject that has been researched numerous times. The investigation of physiologically-controlled biofeedback techniques for gaming purposes dates back to late 1970s and early 1980s (Stern, R., Ray, W., & Quigley, 2000). In fact, physiological metrics seem to be the most popular choice, possibly due to their nature that allows the collection of continuous and unbiased data. One of the early works is "The Atari Mindlink", an unreleased device that allowed to map traditional controllers using the users' forehead muscles. The Japanese version of the title Tetris 64, released in 1998 for the Nintendo 64, included a biosensor that would change the game speed based on the user's heart rate. Overall, these systems failed to achieve a better gaming experience and were seen as simple technological demonstrations. In the last decade however, the industry has shown a growing interest in the use of physiological signals to improve gamers' immersion and experience (Kalyn, Mandryk, & Nacke, 2011).

A vast number of successful attempts have been made in the field of emotion recognition using physiological metrics. For instance (Haag, Goronzy, Schaich, & Williams, 2004) have proposed that emotional states represented in the circumplex model presented by (Posner, Russell, & Peterson, 2005) can be modelled through Electromyography (EMG), Skin Conductance (SC), skin temperature, blood volume pressure (BVP), electroencephalographic (ECG) and respiration (RSP) sensors, reporting an accuracy of about 63% for valence and 89% for arousal (with a 10% error margin). A Neural Network (NN) classifier was used to predict both classes. Applying a similar NN classifier, research by (Leon, Clarke, Callaghan, & Sepulveda, 2007) has discretized valence in three different levels obtaining a recognition level of 71.4%. It uses Heart Rate (HR), BVP, Skin Resistance (SR) and two additional estimated parameters, the time gradient of SR (GSR) and its' derivative. (Drachen & Nacke, 2010) showed that features extracted from SC and HR measures are highly correlated with the reported affection ratings obtained through a seven dimension In-Game Experience Questionnaire (iGEQ). On a similar note, (G. Yannakakis & Hallam, 2008) presented proof of correlation between BVP, HR and SC measures and high-level concepts such as "fun" in a game environment. A later study by (Martínez, Garbarino, & Yannakakis, 2011) reported that with only measures of HR and SC, they were able to predict affective states across games of different genres and dissimilar game mechanics.

Works on possible real-time recognition of emotion have also emerged (Mandryk & Atkins, 2007; Nogueira, Rodrigues, Oliveira, & Nacke, 2013a, 2013b). These take into consideration the possibility of a real world scenario usage. In order to provide continuous classification of a persons' emotional state, a low computational cost is necessary. Furthermore, the usage of a small number of sensors is desirable to assist in their insertion during real gameplay.

2.2 Player Modelling

Parallel to biofeedback related emotion detection techniques, some research points to other ways to model player experience (G. N. Yannakakis & Togelius, 2011). The most direct and simple way is to ask the subjects themselves and build a model based on this data. Although this process may create very accurate models (Georgios N. Yannakakis, 2009), the human factor can lead to some problems. The vast presence of experimental noise (derived from human error in self-judgment, memory, etc.), the intrusiveness of the method among other factors can lead to some difficulty to analyze the data. Works such as (Tognetti, Garbarino, Bonarini, & Matteucci, 2010), have shown how self-reports can be successfully used to capture aspects of player experience. Other works on this area also model the users' experience on an emotional basis (Shaker, Yannakakis, & Togelius, 2009). By observing crowd-sourced playing styles and features of level design, models were constructed that predicted player experience on different emotional dimensions: fun, challenge and frustration. Posterior work on the same data (Pedersen, 2010) added three more dimensions to the prediction. Although these models provide some insight to a players' affective state, the low granularity of dimensions involved does not allow to capture all the nuances of human emotions and affection. A less diffuse way to describe the emotional state of the player is desired.

Another existent method, is the use of gameplay data to try and build these models. The main assumption is that player actions and real-time preferences are linked to player experience, making possible the inferring of the player's emotional state by studying patterns of the interaction (Conati, 2002; Gratch & Marsella, 2005). This method is the least intrusive one, becoming a candid possibility to real world usage. However as (G. N. Yannakakis & Togelius, 2011) state, the models are often based on several strong assumptions that relate player experience to gameplay actions and preferences, resulting in a low-resolution model of playing experience and its affective component.

(Leite & Pereira, 2010) exhibited a social robot that could recognize the user's affective state and display empathic behavior. The users' affective state is inferred through the current state of the game and interpreted according to an empathic behavior model. Complex game aspects such as storyline have also been shown to be dynamically adaptable to individual players, in such a way that a pre-determined reaction is achieved. (Figueiredo & Paiva, 2010) described a small study where by using an expert source manipulation, were able to dynamically adapt the storyline to the player, making him follow a pre-determined path. (Bidarra, 2013) based on actions performed by the player, and created classes of players with different characteristics.

Moreover, solutions that try to combine the previous forms are also frequent, resulting on a hybrid-approach. (Pedersen, 2010; Shaker et al., 2009) implemented gameplay and subjective player emotion models.

The presented work also uses a hybrid approach, using both psychophysiological data and subjective player emotion models. With this method we believe a more effective solution for modelling player experience is created.

2.3 Affective Gaming

The previous topics discussed the works done to detect and predict the affective reaction experienced by the players when playing a game. However, to create more engaging and overall better gaming experiences, changes to the actual game must be made. Having as basis the players' models discussed, it becomes possible to use a player's current emotional state to manipulate gameplay, corresponding to a new form of gameplay, presented by (Gilleade, Dix, & Allanson, 2005) as "Affective Gaming". This process of improving game design is done by shifting the focus from static games with fixed contents to more dynamic systems. The presence of player models in game development allows the game developers to do just that, make informed decisions to elicit the desired emotions and affections on the player. The challenge resides in being able to model player behaviors and experiences and adapt the games' content accordingly (Bidarra, 2013).

One of the early demonstrations of game enhancement (Bersak, McDarby, & Augenblick, 2001) presented a two-player competitive game where the speed of the avatar (dragon) is controlled by the users' relaxation levels, measured through GSR. The more relaxed the user is, the faster the dragon becomes. This seems a common biofeedback game by mapping input controls to physiological data. However, the way the implementation was done counters this. By making the dragon speed increase when the player is relaxed and due to the competitive side of the game, players that became aroused started to lose, and because they were losing, they became even more aroused. By adapting itself to the players' state, it falls on the "Affective Gaming" category. Yet, the moment the user becomes able to control their physiological data to influence the game outcome, the game transforms into a simple biofeedback game (Gilleade et al., 2005).

A significant work on the matter of affective gaming was presented by (Dekker & Champion, 2007). In it several subjects played a modified version of Half-Life 2 on a survival and horror based level. The difference to the original version was that, during gameplay, the game was dynamically modified by the player's biometric information in an attempt to increase the "horror" experience. These changes reflected on audiovisual changes: dynamic changes in the game shaders, screen shake, dynamic changes in the background music, heartbeat sounds among others; and gameplay changes: new zombie spawning points, 'bullet time' effects, weapon damage, stealth mode etc. The results were encouraging, a vast majority of subjects liked the biometric-driven events, and nearly all of them acknowledged their potential.

More recent works have been done at Valve, (Ambinder, 2011) has presented several experiments using a players' physiological data. One of them consisted of a mod to the popular title "Alien Swarm", a top-down, team-based action shooter. The procedure was to index the players' arousal, measured through SC levels, to the countdown timer. When high levels of arousal where detected, the timer speeds up. This created a more frenetic experience, raising even further the arousal levels, similar to (Bersak et al., 2001) experiment. Another experience tried to gain a rudimental understanding of the players' affective reactions. By modifying the Left 4 Dead 2 AI Director. The AI Director is responsible for creating dynamic and variable experience by modifying game events, enemy spawns, health and weapon placement, boss appearances, etc. By determining the in-game encounters based on estimated arousal levels instead of predicted ones, greater values of enjoyment were reported. This lead to some insight into events which elicit enjoyment. In this work, an intriguing question was posed, "Can we determine optimal arousal patterns? ", "do we know the best way to model the players' affective states?" (Kalyn et al., 2011) presented a mixed-methods study to discover the best use for direct (user controllable) and indirect (hard to influence) physiological control in games. It had a basis on a side-scrolling platform shooter game that used a traditional game controller as primary input. Via physiological sensors, the traditional interaction was augmented. Participants played with three combinations of physiological and traditional input. As (Nijholt & Tan, 2007) showed, satisfaction was reported by players out of learning to control their biofeedback through indirect physiological control. Moreover, the physiological augmentation of the game controllers provided a more fun experience. A clear distinction was made by direct and indirect signals, players' reported that physiological controls worked most effectively and were most enjoyable when they were appropriately mapped to game mechanics. On the other hand, indirect control was perceived as best used as a dramatic device in games to influence features altering the game world. Similar methods to shape players' affective experience are presented by (Nogueira, 2013). In it, the adaptive design has its basis on a set of target emotional states and the usage of their emotional reactions to game events.

The design of affective games has also been a target of some approaches. (Gilleade et al., 2005) presented an approach to game design based on high-level design heuristics: assist me, challenge me and emote me (ACE). These can be used to create several different gaming experiences. 'Assist me' proposed a solution to players' frustration (arising from missing clues, inability to advance due to difficulty, etc.) by measuring it using physiology signals and, combined with knowledge of the game context, provide mechanisms to identify this situation and adjust the game itself accordingly. Results gathered from their own affective game showed that casual gamers were the most sensitive to these changes. 'Challenge me' had its inception due to the difficulty provided by commercial games. Usually only three or more levels (easy, medium, hard) are presented and it is the user himself that indicates their perceived expertise, hoping it matched the game designers intent. This leads to inefficient challenges presented and subsequent lack of engagement. The solution is to dynamically alter the challenge provided by the game based on the user's arousal, thus creating a more personalized gameplay experience. 'Emote me' refers to the emotional experiences players' are provided with and the best way to deliver them. By determining the current users' emotional state, and the intended one by the game designers, the game must modify its content to provoke the desired emotions.

Adding to the previous work, (Hudlicka, 2009) suggested a set of requirements for an affective game engine, with the purpose of allowing game developers the creation of better affective games. It presents a series of high-level requirements, not specifying their exact implementation. One of the central elements of this engine would be a knowledge-base that would contain information about emotions in general (their generation, influences, expression), and a depiction of the players' and other non-playing-characters' affective states. Four components with different functionalities would then be built that shared and changed this database: the recognition of the players' emotion, the expression of emotions by both the player avatar and the game characters, the dynamic construction and maintenance of the players' affective model (affective user models), and the modeling of emotion within the games' characters (Hudlicka, 2008).

Lastly, (Nogueira & Rodrigues, 2013) proposed an implementation of these high-level abstract requirements through a psychophysiological approach nicknamed Emotion Engine (E^2) biofeedback loop system. Figure 1 presents a high-level representation of this system architecture.

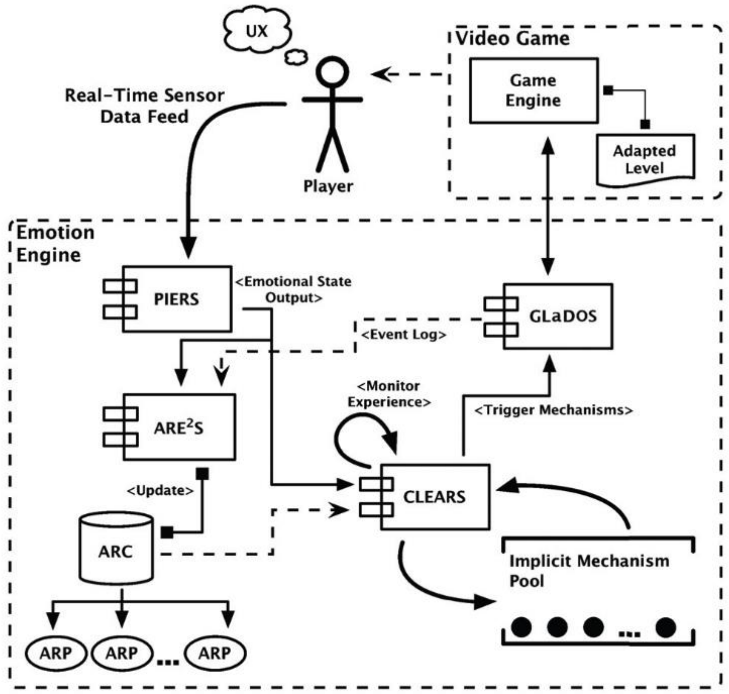

Figure 1: The Emotion Engine ($\boldsymbol{E^2}$) architecture.

2.4 Summary

Over the years, the usage of both direct and indirect biofeedback in games has gained increased attention of researchers. Real time psychophysiological data provides several possibilities for improving gaming experience, whether being new ways of input or enhancing immersion levels via affective gaming. Our goal is to extend current work on the field by presenting a way to create affective reaction models from this psychophysiological data and use them to enhance gaming experiences, through the use of adaptive and parameterisable affective gaming. In addition, previous publications made by the author to some renowned journals present some relevant information (Nogueira, Aguiar, Rodrigues, & Oliveira, 2014a, 2014b).

Chapter 3

Affective Reaction Models

One of the main aims of this work is to create individual player models for the prediction of their respective emotional responses to a predetermined set of game events. This means that for each subject, given an initial emotional state and game event, their emotional reaction in both arousal and valence dimensions is predicted. As such, these models should obey Equation 1:

$$\phi : \Lambda X \Omega \rightarrow \vec{w}$$

Where Λ is the set of possible emotional states and Ω the set of possible events. Thus, function Φ receives an emotional state λ, such that $\lambda \in \Lambda$ and an event ϖ, such that $\varpi \in \Omega$, and outputs a vector $\vec{w}$ that contains the emotional reaction. This vector can have several dimensions, being their total number defined by the space used to define an emotional state. This work uses the circumplex model of affect as presented by (Posner et al., 2005). This space has two dimensions, Arousal and Valence. Valence depicts the nature of an emotion, lower values mean sadder emotions, higher values happier emotions. Arousal measures the level of excitement, how stron is the emotion. As such, the above vectors present some constraints.

$$\forall\, q \in [1,2]: \overrightarrow{w_q} \in [0, 10]$$

3.1 Emotional Reactions Feature Extraction

For the creation of these affective emotional reaction models, an extraction of real-world emotional reactions was performed. In this study, these were extracted from 72 gameplay sessions of an indie horror game denominated Vanish. A total of 24 participants were present throughout this experience.

Vanish is a survival horror videogame where the player must escape a series of tunnels. This network of maze-like tunnels is procedurally generated. In order to escape, a series of key items must first be found and picked up by the player, only then being allowed to escape. At gameplay-time, a monstrous creature stalks and preys on the player, forcing him to avoid her at all costs. Several events happen in-game, both visual and audio, in an attempt to engage and involve the player in the game's atmosphere. These events range from lights failing, pipes bursting or even the creature's distant howl/cries. All these events, along with death, the locating of new items and creature encounters are tracked and constitute the whole set of considered game events.

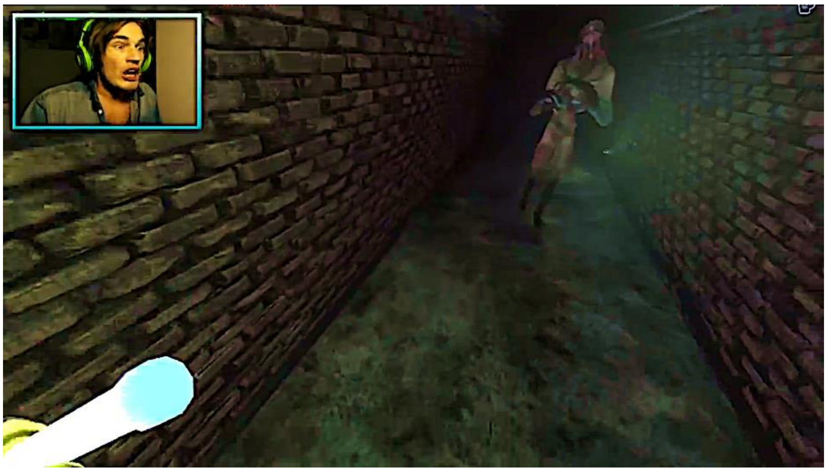

Figure 2: Screenshot of a creature encounter event on a Vanish gameplay session publicly available on Youtube

As previously mentioned, the collected dataset originated from 24 players over 72 gameplay sessions. Regarding the subjects, they were randomly selected from a pool of interested candidates (N=89) being that their ages varied between 19 and 28 years old (μ = 22.47, σ = 2.50). The physiological data was obtained via a range of sensors: Skin Conductance, Heart Rate and facial EMG. Although an hybrid approach of the work was used, a more in-depth analysis of the process of mapping physiological input to emotional states can be found in (Nogueira, Rodrigues, et al., 2013a) combined with rules suggested by (Mandryk & Atkins, 2007). Regarding the special placement of these sensors, HR was derived from BVP, SC was measured at the players' index and middle finger using two Ag/AgCL surface sensors snapped to two Velcro straps and facial EMG was measured at the zygomaticus major (cheek) and the corrugator supercilii (brow) muscles.

This physiological data is then processed, producing a 1:1 both arousal and valence ratings, being afterwards segmented by study participant. The automatically generated timestamps were then synchronized to these AV ratings in order to extract an emotional response. Singular emotional reactions were then extracted by using a time window of 0.5 seconds prior and 5 seconds after the correspondent timestamp. The contextualization of the players' immediate emotional response prior to occurrence of the game event and the analysis of his emotional reaction is possible due to this time window. Note that these values were not random, they are based on the detected physiological data and player perception delays of game events (Nogueira, Torres, & Rodrigues, 2013).

Additionally, a total of twelve features are extracted for each emotional reaction: six related to arousal and another six pertaining to valence levels. Both valence and arousal share the same feature extraction process. Onwards from the gameplay event timestamp, the following features are created:

- $E\{r\}$: Initial value, calculated as the average of the maximum and minimum values registered in the 0.5 seconds prior to the game event
$$avg(max\{r\}, min\{r\})$$
- $\mu\{r\}$: Mean of the signal
- $\sigma\{r\}$: Standard Deviation of the signal
- $M\{r\}$: Maximum Value of the signal
- $m\{r\}$: Minimum Value of the signal
- D_h: Absolute time period between minimum and maximum value
$$D_h = |\ t^h_{max}\{r\} - \ t^h_{min}\{r\}\ |$$
- $h_{in}\{r\}, h_{out}\{r\}\ and\ h_{ev}\{r\}$: Auxiliary features denoting the reactions beginning, ending and event timestamps.

The delta value of the reactions (ΔA , ΔV) are calculated as the greatest difference registered between the maximum and minimum values of the initial time window frame (0.5 seconds prior to the game event), and the maximum and minimum values of the remaining event time window.

Over 1400 (fourteen hundred) individual emotional reactions were recorded. However, a more in depth analysis to this data brought some questions to the surface.

First of all, one particular subject presented a greatly reduced number of events and emotional reactions. Moreover, nearly all of his emotional reactions were concentrated on a pair of events, resulting in insufficient data when looking at the full spectrum. As such, this subject has been completely removed from subsequent phases. Additionally, two subjects didn't have their emotional reactions recorded due to hardware failures making impossible their inclusion.

Lastly, a total of three events were not present in more than half of the input data, and were, as such, entirely eliminated from the dataset. Their presence would wrongly inflate the classifiers' performance. After all this filtering process, of both subjects and events, over 1160 emotional reactions are present in the full dataset.

An additional manual examination was made to the data. For each pair of subject and event, their emotional reactions were drawn along the initial arousal and valence values. This allowed to perceive outliers, possibly originated from another event that occurred at the same time. Only values that were vastly irregular with the other data were adjusted. These adjustments still preserved some of this point disparity, however their value was changed to better represent the overall players' response. This was done to preserve the maximum amount of information and because the detection of real outliers is a complex and difficult decision.

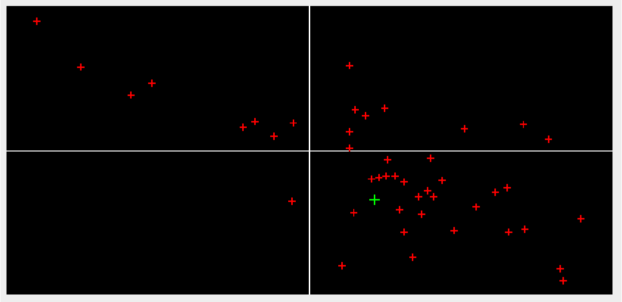

Figure 3: Representation of one of the plots used to readjust outliers

A general overview of the statistics of the final dataset is present in Table 1. To note that $Ai \equiv \mathrm{E}\{r_A\}, Vi \equiv \mathrm{E}\{r_V\}, dA \equiv \nabla A, dV \equiv \nabla V$ (see abbreviations)

	Mean	Median	Standard Deviation
Ai	6,4296	8,6054	0,6844
Vi	4,3369	5,8750	0,5520
dA	0,2388	0,2342	0,2168
dV	0,3356	0,3142	0,3226

Table 1: Global Dataset Statistics

As is easily noted, the average initial arousal level is larger than the baseline value (5), while its valence counterpart shows a lower value. This is probably due to the games nature. Being a horror game, players remain in a constant state of alert.

3.2 Machine Learning

One of the most fundamental steps of this thesis is the creation of affective reaction models. The ability to predict the players' emotional responses is of utmost importance and relevance.

The first approach to the creation of these affective reaction models is the employment of machine learning. With this, a model is created that predicts the emotional response of a subject to a certain event along all emotional states spectrum. This whole process was segmented into several phases, namely: single classifiers, optimal feature selection algorithm and lastly the creation of these models.

3.2.1 Single Classifiers

To serve as a baseline and due to the exploratory nature of this work, the first models created used a single feature. This can lead to some conclusions and deductions that might prove valuable in later phases. Possible correlations between features and classes can also be discovered through this approach.

	Mean	Median	Standard Deviation
dA	0,2961	0,2687	0,2236
dV	0,4193	0,3524	0,3528

Table 2: Global RMSE Values

	Mean	Median	Standard Deviation
dA	0,5307	0,5444	0,4061
dV	0,5453	0,5675	0,3967

Table 3: Global Pearson Coefficient Values

Root Mean Squared Error (RMSE) gives us a solid way to evaluate the dimensionality of the errors involved in the classification. Moreover, due to its nature it penalizes the existence of very strong outliers, which is something beneficial viewing that large errors in classification can lead to extremely bad results later on. As one can see in Table 2, the error values presented are very large taking into consideration the range of values in the classes involved. These error rates are significantly larger than the original Standard Deviation, leading to the belief that the classification has poor results. Also note the vastly superior error rates in the Valence dimension. Both an increase in Mean error and its Standard Deviation is noticeable. This probably originates from lower volatility in estimating arousal, opposed to valence, as shown in (Nogueira, Rodrigues, et al., 2013a).

The same can be seen in the Pearson Correlation Coefficient presented in Table 3. This value, ranging from minus one to one, measures the linear correlation between variables, higher absolute values representing higher correlations. As one can see, the mean values presented are relatively small. Furthermore, the Standard Deviation is extremely large, indicating abnormal classifications throughout.

A closer look to the detailed RMSE values showed that Arousal related features provided less error values. The same happened in the Valence dimension. However, this difference is very small, not providing sufficient insight.

3.2.2 Optimal Feature Selection Algorithm

Because the classifying/regression approach is dependent on the features selected to create it, the selection of these features can vastly improve the viability of the models. As such, several feature search methods are tested to improve the overall results. Because of its proven reliability and results, the attribute evaluator used can be seen in (Hall, 1999). A total of four different search methods are employed, Best First (BF), Random Search (RS), Linear Forward Selection (LFS) and Genetic Search (GS). The RS method serves as baseline due to its random nature. The last ones are well accepted among the industry and have proven their values by having good results over a vast selection of fields and applications. Several other search methods were not used due to some constraints, some required an evaluator function that only deals with single features excluding subsets of features (similar to previous phase). Others, for example Exhaustive Search, required too much processing power, making them undesirable.

Each of the search methods used is then combined with three different classifiers (the reason for the usage of these classifiers is presented later) and are evaluated. The results obtained are shown in Tables 4 through 7:

	BestFirst			LinearForwardSelection		
	Mean	Median	SD	Mean	Median	SD
dA	0,28436	0,26446	0,21052	0,28438	0,26452	0,21054
dV	0,37308	0,30606	0,31844	0,37303	0,30555	0,31854

	GeneticSearch			RandomSearch		
	Mean	Median	SD	Mean	Median	SD
dA	0,28725	0,26209	0,21365	0,28848	0,26807	0,21246
dV	0,37680	0,31286	0,32002	0,38425	0,30938	0,33041

Table 4: RMSE Values

	BestFirst	GeneticSearch	LinearForwardSelection	RandomSearch
dA	2,11934	3,09053	2,11934	4,02058
dV	2,13580	3,06584	2,13580	4,11523
Total	2,12757	3,07819	2,12757	4,06790

Table 5: Average Number of Features

BF	Ai_sd	Ai	Vi_sd	Ai_Abs _t	Vi_Abs _t	Vi_max
	36,21%	30,86%	30,04%	22,22%	20,58%	14,40%
GS	Ai_sd	Vi_sd	Vi_Abs _t	Ai_Abs _t	Ai	Ai_max
	37,45%	37,04%	32,10%	30,04%	29,22%	23,46%
LFS	Ai_sd	Ai	Vi_sd	Ai_Abs _t	Vi_Abs _t	Vi_max
	36,21%	30,86%	30,04%	21,40%	20,99%	14,40%
RS	Ai_sd	Vi_sd	Ai_Abs _t	Vi_Abs _t	Ai_u	Ai_max
	51,03%	45,27%	41,15%	40,74%	35,80%	30,45%

BF	Ai_max	Vi	Ai_u	Ai_min	Vi_u	Vi_min
	11,93%	11,52%	9,05%	8,64%	8,23%	8,23%
GS	Vi_min	Ai_u	Ai_min	Vi	Vi_max	Vi_u
	23,05%	22,63%	21,81%	19,75%	19,34%	13,17%
LFS	Ai_max	Vi_min	Ai_u	Vi	Ai_min	Vi_u
	11,93%	11,52%	9,05%	9,05%	8,64%	7,82%
RS	Vi_min	Ai_min	Vi	Ai	Vi_max	Vi_u
	30,04%	27,16%	26,75%	25,51%	25,10%	23,05%

Table 6: Ordered Features Usage When Classifying dA

BF	Vi_sd	Ai_sd	Ai	Vi	Vi_Abs _t	Ai_Abs _t
	32,92%	30,86%	27,98%	22,63%	21,40%	19,34%
GS	Vi_sd	Ai_sd	Vi_Abs _t	Vi_min	Ai_Abs _t	Ai
	41,56%	33,33%	31,28%	28,81%	28,40%	25,93%
LFS	Vi_sd	Ai_sd	Ai	Vi_Abs _t	Vi	Ai_Abs _t
	32,51%	30,86%	27,98%	21,81%	20,99%	19,34%
RS	Ai_sd	Vi_sd	Vi_Abs _t	Ai_Abs _t	Vi	Vi_min
	56,38%	44,86%	43,62%	39,51%	37,45%	35,80%

BF	Vi_max	Vi_min	Ai_max	Vi_u	Ai_min	Ai_u
	14,40%	11,93%	11,52%	11,11%	4,94%	4,53%
GS	Vi	Ai_max	Vi_max	Vi_u	Ai_min	Ai_u
	24,69%	23,05%	20,16%	18,11%	17,28%	13,99%
LFS	Vi_max	Vi_min	Ai_max	Vi_u	Ai_min	Ai_u
	14,40%	13,58%	11,52%	11,11%	4,94%	4,53%
RS	Ai_max	Ai_u	Vi_u	Ai_min	Vi_max	Ai

	29,63%	28,40%	27,16%	25,10%	23,05%	20,58%

Table 7: Ordered Features Usage When Classifying dV

Analysis of the Global RMSE values on this approach bring forth some conclusions (Table 4). First, the use of FSA over Single Feature decreases the error value significantly. This means the usage of a subset of features over a single feature brings visible benefits. Second, Random Search has, as expected, the largest error. This is due to the random nature of the method. However, Genetic Search presents marginally better results. Best First and Linear Forward Selection have the lowest error values. They present similar and largely better results.

Table 5 presents the average number of features used by each FSA when trying to classify each dimension. A similar pattern is seen. Random Search uses approximately four features model, while LFS and BF only half. Genetic Search sits on the middle of this table. Some ratings can already be made to these FSA. Random Search has the highest error value and uses the most features. Next comes Genetic Search with better results. BF and LFS present similar results and expressively better than their counterparts. Because of this BF is the FSA used in posterior phases.

On Table 6 and Table 7 the usage of features per FSA is depicted. The presented ordered list allows for a quick inspection to the most used features to predict both dimensions. Similar to the single feature results, Arousal features tend to be chosen more often when classifying the emotional reaction in the Arousal dimension. The same phenomenon is manifested in the Valence dimension.

3.2.3 Model Creation

The final step is the creation of the models using the feature selection algorithm previously chosen. Viewing that this is a regression problem, several classifiers can be used to predict these emotional reactions. However due to some previous observations, only three classifiers were chosen. This relates to some of the patterns discovered in the previous phases, during the pre-processing of the data. In general, two strong patterns emerged from the visualization of the data, a linear model and a more quadratic and complex type. For this reason, three different classifiers were used and their results compared in order to obtain the best prediction: Linear Regression (LR), M5 Model Trees and Rules (M5P) and Multilayer Perceptron (MLP). Both the M5P and the MLP classifiers can easily handle complex behaviours. However, the last one can more easily fall in the pit of over fitting. Nevertheless, both are tested to ensure the best possible result. For each one of these three classifiers, a total of three evaluation modes are presented. Presented in order of preference: 10-fold cross-validation, 3-fold cross validation and the use of the whole testing set for training. Some key results are shown in Tables 8, 9 and 10:

	Average	SD
LR	0,311353	0,075435
M5P	0,29591	0,067145
MLP	0,238855	0,093251

Table 8: RMSE Global Value When Classifying dA

	Average	SD
LR	0,400987	0,138804
M5P	0,396201	0,127097
MLP	0,323791	0,139332

Table 9: RMSE Global Values When Classifying dV

	Average	SD
dA	0,230942	0,080196
dV	0,312054	0,123437

Table 10: RMSE Global Values for Best Classifier

As is seen in Table 8 and Table 9, a large difference exists in the error rates between classifiers. Linear Regression presents the worst results, with both high average and standard deviation error values. On the other hand, M5P shows better results. However, only a small decrease in error is seen. The decrease in the Standard Deviation of the errors' values is an encouraging result. Even so, it is still not a satisfactory solution. Moreover, due to the MP5 ability to mimic the LR and create more complex "functions", this increase in performance was expected. The Multilayer Perceptron classifier showcases largely better results. A vast difference in the average error is seen. However, the large increase in the Standard Deviation values brings suspicion to the validity of this solution. Overfitting may have occurred.

As a final global overview, the RMSE values when the same and best classifier per event is used are shown in Table 10. This approach makes the posterior comparison between subjects possible by making each related model based on the same classifier. Nevertheless, although the results are slightly better than previous ones, when looking at the global scope they are not adequate enough. Error rates are in the same magnitude as the Standard Deviation for the class in question, making the results not very positive.

3.3 Clustering Approach

The previous machine learning approach treated individuals as single entities, without any relation between them. A classifier predicted the players' affective reaction to an event based on the optimal subset of features. This however might not produce the best results. Viewing that we are modelling human behaviour, some patterns and relationships between subjects can help strengthen these predictions. As such, and with the intent of approximating the human world, a second approach was attempted where groups of people that present similar emotional responses are treated as whole. This is done via hierarchical clustering.

In order to make possible the comparison of models and their subsequent distances calculated, an initial stage of creating these models with the same domain is necessary. Only with these distances is the actual clustering possible. Due to some correlations presented between features and classes, the creation of these models was transformed into a three-dimensional space, with two of the axes representing the features used: initial arousal (Ai) and initial valence (Vi) and the third axis representing the expected emotional reaction in either arousal or valence levels (dA or dV) that the player experienced. The purpose is to find a relationship between the affective reaction (response variable) and the combination of the two features (predictors).

As in the previous approach, taking into consideration the perceived distributions of the reactions, the relationships were built through linear and non-linear regression models. Viewing that using an automatic approach to discover the regression model that produces better results will most certainly lead to overfitting and high-degree polynomials, a supervised approach was followed. After a manual analysis of a large number of these models, some conclusions were inferred. First, the regression models should not exceed a third degree polynomial. A bigger degree represents a negligible increase in the fitness of the model while showing a large increase in symptoms of overfitting. Lastly, due to the nature of a second degree polynomial, being characterized for its parabolic shape, the models created with this degree will present an ever growing emotional reaction either in positive or negative values. This is incongruent with common sense, which led to the decision of not using these models.

Ultimately, the regressions were produced using either linear or third degree regressions, depending on the number of points available for the model. Figure 4 illustrates a sample plot of a model. The upward axis represents the reaction. A Linear Regression is already present. Table 11 presents a small error analysis.

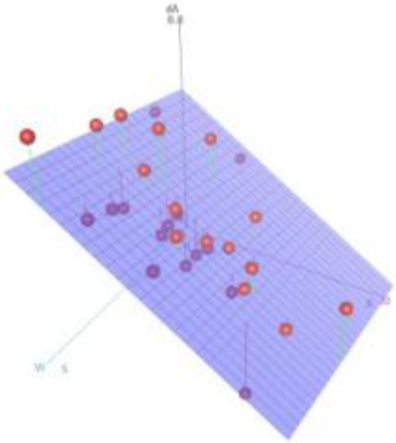

Figure 4: Representation of one model in the 3D Space.

	Average	Standard Deviation
R^2 Value	0,954267	0,208303
Adjusted R^2 Value	0,704619	0,509913

Table 11: Coefficients of Determination Values

Present in Table 11 are the coefficients of determination for the created models. The R-squared value ranges from zero to one, indicating how well data points fit a statistical model, in this case the regressions. A value of zero specifies that the model explains none of the variability of the response data around its mean. A value of one indicates all the variability is explained by the model. The adjusted R-squared is a modified version of R-squared that adjusts itself for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance.

As can be seen, the R-squared value shows a very high average value with relatively small standard deviation. On the other hand, the adjusted R-squared variable presents a moderately smaller mean, with higher standard deviation. This relates to the low number of points present in the models. Due to the difficulties of retrieving a high number of emotional reactions already discussed earlier, the number of emotional reactions per model is not very high. Because of that, some models created present some overfitting as can be seen by the high R-squared value. Furthermore, because the models are built upon a low number of points, the adjusted R-squared value penalizes heavily the usage of linear and even more the quadratic models.

Each one of these models predicts the emotional response of one subject to a certain event over one dimension (either arousal or valence). As such, a total of 32 regression surfaces are created that describe an individual player's emotional reactions over the AV emotional space. The next step is to create a distance matrix that depicts the differences between emotional responses. For that to happen, these models need to be compared. This is done via the mapping of these models over all the AV space. The generation of a hyper-dimensional matrix that ranges from [0, 10] for both feature response variables (initial arousal and valence) with a 0.1 increment. By standardizing these models representations, they can easily be compared and their differences evaluated.

The next step is the creation of a distance matrix from which the hierarchical clustering is done. For that, a way to quantify the distance between the maps created is needed. The current implementation provides three different distance calculations and an extra post-processing stage that scaled these distances over several functions. Relatively to the distance calculations, the first one is a Euclidean distance.

This a well known measure for distance that has proved its usefulness in various fields. All points of the models are compared and the average Euclidean distance between points is then used as the distance between models. This procedure of calculating the distance between all the models' points and then averaging this sum is used for all distance calculations.

The second calculation has its basis on an exponential distance function. This sanctions bigger distances even further, increasing their value. Additionally, similar models maintain a very low distance tightening their relationship.

Finally, the third method simply uses the normal distance, with no alteration to its original value.

$$Euclidean\ distance = \sqrt{\sum_{k=0}^{n} (P[k] - M[k])^2}$$

$$Exponential\ distance = \frac{\sum_{k=0}^{n} e^{P[k]-M[k]}}{n}$$

$$Normal\ distance = \frac{\sum_{k=0}^{n} P[k] - M[k]}{n}$$

$$n = number\ of\ points\ in\ sequence$$
$$P[x] = value\ of\ sequence\ P\ in\ the\ point\ x$$

A post-processing stage was made. It was an experiment that tried to modify the whole range of distances calculated in order to see if it would yield better results. The general idea was to penalize certain ranges of distances, for example differences in the higher distance ranges are attenuated. Sigmoidal, logarithmical and the original linear functions were used. However, it was later revealed that the original linear function produced better results, which led to no post-processing changes made in the final data.

At this stage, we have the tools to determine the distance between each pair of player/event regarding one emotional dimension. We need to correctly merge these values in order to create a distance between subjects. First, we create matrices holding the distances between subjects for a single game event over one emotional dimension. Since we have no evidence that any particular game event or emotional dimension has higher influence on the games' affective experience, we can merge this "partial" distance matrices by averaging all their cells, assuming the correspondent cell orders are preserved. This brings forth a new global matrix holding the distances between subjects over all game events and emotional dimensions. With this new matrix, a hierarchical clustering algorithm can be applied to cluster players.

The hierarchical clustering used employed the Ward's method (Joe H . Ward, 1963) for the criterion of choosing which clusters to merge, meaning the objective function is the error of sum of squares. Furthermore, a multi-scale bootstrap resampling process is used, allowing the assessment of uncertainty in this clustering approach. More specifically, Approximately Unbiased (AU) and Bootstrap Probability (BP) *p*-values are computed. Note however that the AU *p*-values, computed via multiscale bootstrap resampling, provide a better approximation to unbiased p-value than the BP value that is computed by normal bootstrap resampling. These values represent the confidence that a particular cluster is supported by the data, not simply caused by "sampling error" but may stably be observed if we increase the number of observations. As a more formal definition for these values, a cluster presenting an AU p-value of x has the null-hypothesis "the cluster does not exist" rejected with a significance level s = 100 – x. In sum, high AU p-values provide high confidence to the clusters found. With this in mind, all the results of the different distance matrices generated were manually analysed. The result was the selection of the exponential distance. The final result of this hierarchical clustering approach is presented in Figure 5.

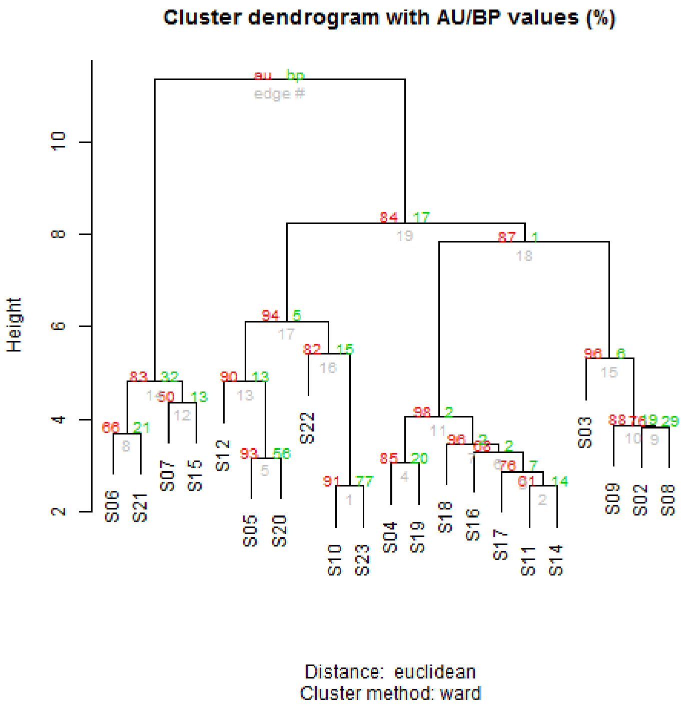

Figure 5: Final Clustering Result

As can be seen, the result presents high AU *p*-values throughout all the clusters found, leading us to the belief of a solid global solution. Moreover, the clusters seem well distributed, fact that can probably be attributed to the different demographics used in the extraction of the emotional responses.

As previously mentioned, the whole idea of this approach was to make the construction of the affective reaction models more congruent with the relations seen in human behavior, where several kind of people react similarly amongst themselves. As such, the choice of clustering relates to this fact. However, the final models work with a fixed set of clusters, whether they came from hierarchical or non-hierarchical clustering is irrelevant. The choice of hierarchical clustering relates to another fact. Because we have demographic information about the specific subjects in question, some analysis can be done relating the clusters and this information. Because of the way hierarchical clustering works, this can be done over all number of clusters. A clustering approach like for example x-means (k-means with automatic cluster number identification) would not preserve the clusters through the increase in the cluster numbers, disabling the possibility of studying this relationship. A manual observation was made regarding this issue, resulting in some encouraging results. Various clusters showed similar demographic information such as gender, type of gamer and the predisposition to horror games. Others presented correspondence in a combination of several features. Note however that, due to time issues, not all cluster numbers were properly analyzed.

3.3.1 Clustering Validation

Although AU *p*-values computed via multiscale bootstrap resampling provide a good measure of the clusters strength, one more question remains. Viewing that a number of clusters needs to be determined in order to construct the models, a way to evaluate the "goodness" of the resulting clusters is needed to evaluate the several possible number of clusters. As such, two internal indexes are used to measure this: cluster cohesion and cluster separation.

Cluster cohesion measures how closely related objects in the same cluster are. It is the sum of the weight of all links within a cluster, calculated via the within cluster sum of squares. An average of all the clusters values is used to present a global value of cohesion.

Cluster dispersion quantifies the level of distinctiveness between clusters. It is the sum of the weight of all links within a cluster, measured by the between cluster sum of squares. The same procedure of averaging all the values from a particular cluster number is applied for the discovery of a global value. Note however that in the case of separation each cluster has associated N-1 measures, being N the number of clusters.

$$Cohesion = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

$$Dispersion = \sum |C_i|(m - m_i)^2$$

$$C_i = Cluster\ i$$
$$|C_i| = Size\ of\ Cluster\ i$$
$$m_i = Centroid\ of\ Cluster\ i$$

A representation of both these measures along the number of clusters can be seen in Table 6.

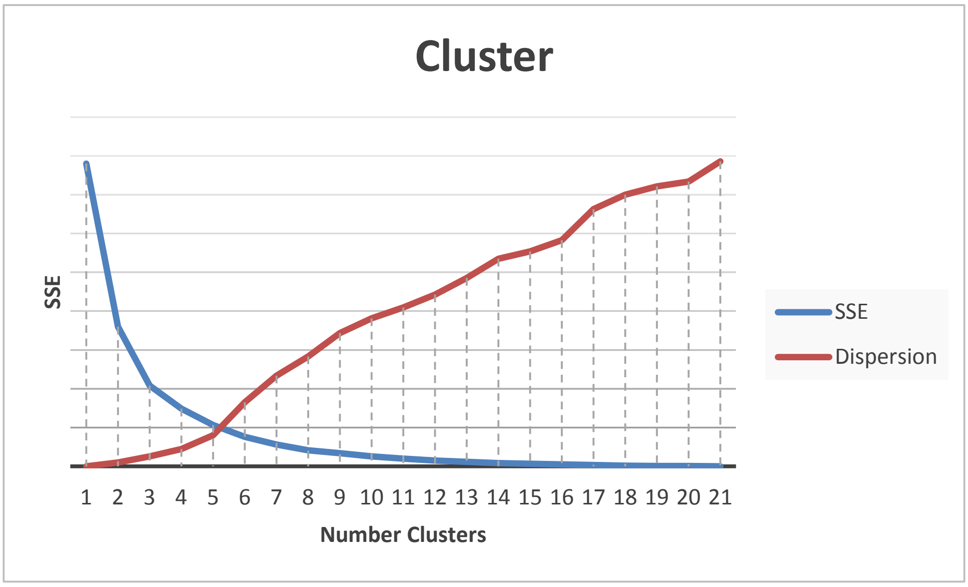

Figure 6: SSE and Dispersion over Number of Clusters

As seen in Figure 6, one can say the best number of clusters happens when the cohesion has low values and the separation high ones. However, as can be seen, these optimal values only happen at extremely high number of clusters, defeating the purpose of the clustering approach. As such, the best approach is to use the number of clusters that represent the best "gain" in both these dimensions comparatively to the previous number. This means we need to discover the number of clusters where the "acceleration" of the cohesion decreases and the "acceleration" of the separation increases. A simple way to tackle this is to integrate these values twice and inspect the local maximum and minimums. We are trying to find a minimum in cohesion, and a maximum in dispersion, representing a loss of efficiency when increasing the number of clusters. In the case of our clustering, the result obtained was 6 (six) as the best number of clusters.

After this number is calculated, the task of creating the various clusters' affective reaction models is possible. With these models, the players' individual ones can be easily represented as a composite of all the clusters models with different weights. These weights are related to the differences encountered between the original individual models and the clusters ones. The implementation finds the sum of all distances between one player and all the clusters and uses it to find all the weights.

$$Weight_{i,k} = \frac{\sum_x distance_{i,x}}{distance_{i,k}}$$

$$Weight_{i,k} = weight\ of\ cluster\ K\ in\ subject\ i$$
$$distance_{i,k} = distance\ of\ subject\ i\ to\ cluster\ K$$

Chapter 4

Demographic Study

To evaluate the relative impact of different types of Indirect Biofeedback (IBF) adaptation mechanics, the original extraction of the emotional reactions made the participants play three different versions of the game: two augmented using the biofeedback mechanics and one control condition. Each game version presented the same gameplay elements and mechanics. Both the game design and biofeedback adaptations were developed during an extended alpha-testing period using an iterative prototype over 3 months, gathering feedback from over 20 individuals not included in this study. After a brief description of the experiment and providing informed consent, players completed a demographics questionnaire. Participants also completed a game experience questionnaire (GEQ) (IJsselsteijn, Poels, & De Kort, 2008). Additionally, they were also asked to report their Fun ratings in a 10-point Likert scale. The full extracted features follow:

- *Demographic Data*
  - *GType - Reported Gamer Type: "Hardcore" or "Softcore"*
  - *Likes - Predisposition towards Horror Games: "Yes" or "No"*
  - *Gender - "Male" or "Female"*
- *Physiological Data*
  - *SeqDur - Game Session Duration (min)*
  - *{X} - Arousal or Valence dimension*
  - *{X}Mean - Mean of the signal*
  - *{X}SD – Standard Deviation of the signal*
  - *{X}P - Number of Absolute Peaks*
  - *{X}PMin - Number of Peaks Per Min*
  - *{X}PInt – Average Peak Intensity*
  - *{X}PMag - Average Peak Magnitude*
  - *{X}Max – Maximum signal value*
  - *{X}Min – Minimum signal value*
- *User Experience*

- *Challenge (Chall)*
- *Competence (Comp)*
- *Flow (Flow)*
- *Immersion (Imm)*
- *Fun (Fun)*
- *Tension (Ten)*

In this section we aim to create computational models of user experience through the usage of demographic and physiological features. An optimal feature subset selection procedure capable of capturing non-linear relationships is employed. Finally, we use the optimal feature subset for each user experience dimension identified by the best-performing feature selection algorithm (FSA) - measured in terms of its achieved root-mean square error (RMSE) - to create computational models of user experience. All features are explored in this phase (physiological and non-physiological).

4.1 Single Model

The first step for the analysis of this data, is the creation of a single predicting model. Because the features include physiological and non- physiological data, three different feature subsets are used in order to further compare these two sources of information. One subset only contains biofeedback features, another only demographic ones, and the last one can contains both. This allows us to differentiate between demographic and physiological data for the classification of user experience. Additionally, the usage of all features can be seen as a baseline.

A global overview regarding the first phase where a single model was constructed follows:

	Imm	Ten	Comp	Chall	Flow	Fun	Average
All	1,682971	1,83367	2,54955	2,153161	2,025733	1,229042	1,912354
Bio	1,491728	1,859111	2,148719	1,949156	2,17748	1,273107	1,81655
Non-Bio	1,423103	1,52644	2,687004	1,905734	1,976445	0,989578	1,751384
Average	1,532601	1,73974	2,461758	2,002684	2,059886	1,163909	1,826763

Table 12: RMSE over Feature Segmentation

	Likes	SeqDur	Sex	Cond	VPMin
BestFirst	12	10	10	8	8
GeneticSearch	12	10	10	8	8
LinearForwardSelection	11	12	10	8	8
Average Feature Usage	64,81%	59,26%	55,56%	44,44%	44,44%

Demographic Study

Table 13: Features Usage over FSA

As show in Tables 12 and 13, the construction of the single model stems some large errors in certain related user experiences. However, “Fun” and “Immersion” reported low error values. Another deduction that can be made relates to the features. Nearly all classes presented smaller errors when the features used where only Demographic ones. Additionally, the FSAs’ vastly recognized this type of features as very valuable. This is probably the same situation seen in the Machine Learning approach in the construction of the affective reaction models, human nature leads to the existence of similar experiences between some groups of people. To tackle this problem, viewing that demographic data is available and presents better results, the whole population was segmented via these features. This means that instead of one single model for the prediction of a class, several ones are created, each one containing a demographic segmentation of the population, for example male players.

4.2 Feature Selection Algorithm

Some data retrieved regarding the several Feature Selection Algorithms follows:

	BestFirst	GeneticSearch	LinearForwardSelection
All	1,772586	1,803699	1,772586
Bio	1,762442	1,794179	1,762442
Non-Bio	1,694701	1,691236	1,694701

Table 14: RMSE over FSA and Feature Segmentation

		BestFirst	GeneticSearch	LinearForwardSelection
Cond	NB	2,178031	2,22728	2,178031
	NV	1,723597	1,753616	1,723597
	V	1,855355	1,949551	1,855355
GType	Hard	1,451051	1,514739	1,451051
	NA	1,461829	1,442075	1,461829
	Soft	2,148233	2,079193	2,148233
Likes	N	1,981011	2,085534	1,981011
	Na	1,461829	1,442075	1,461829
	Y	2,051743	2,01593	2,051743
Sex	F	2,204128	2,299506	2,204128
	M	1,572497	1,632421	1,572497

Table 15: FSA over Demographic Segmentation

Chall	ASD	VSD	VPInt	SeqDur	Sex
	51,52%	42,42%	42,42%	36,36%	33,33%
Comp	ASD	VMin	Cond	VPMin	Likes
	54,55%	42,42%	36,36%	33,33%	30,30%
Flow	SeqDur	VPMag	Sex	VPInt	Likes
	57,58%	42,42%	36,36%	36,36%	33,33%
Imm	Cond	VMax	SeqDur	Likes	APMag
	42,42%	36,36%	33,33%	27,27%	27,27%
Fun	Cond	VPMag	Sex	Likes	ASD
	45,45%	45,45%	42,42%	39,39%	36,36%
Ten	Cond	APMin	SeqDur	GType	VMean
	42,42%	33,33%	33,33%	24,24%	24,24%

Table 16: Top 5 Selected Features

	SeqDur	Cond	Sex	Likes	ASD
Best First	72	67	60	58	55
Genetic Search	83	67	59	58	58
Linear Forward Selection	72	67	60	58	55
Average Feature Usage	38,22%	33,84%	30,13%	29,29%	28,28%

Table 17: Top 5 Globally Selected Features

Table 14 presents the global error rates for every FSA. Note the average smaller values presented in this approach comparatively to the single model one. Additionally, one can compare RMSE values between FSA. In this case, both BF and LFS present nearly identical and better results than GS. As such, BF was used in posterior phases.

Regarding the demographic segmentation, some interesting results arose. Namelly, the large error values reported in the "softcore" type of players. The low error present on people where no "Likes" status was extracted is also observed. One of the most striking differences comes from sex information, male players reported vastly smaller errors comparatively to female gamers.

Tables 16 and 17 present the selection of features by FSA, both through classes and FSA. The selection of mainly demographic results is still very noticeable. However, some features such as Arousal Standard Deviation and some Valence-related measures appear frequently. Only looking at biofeedback features, for each class predicted the most frequent feature can give some insight into the relationship between physiological data and the reported user experience. These are:

- *Challenge: Arousal Standard Deviation*
- *Competence: Arousal Standard Deviation*
- *Flow: Average Valence Peak Magnitude*
- *Immersion: Valence Max*
- *Fun: Average Valence Peak Magnitude*
- *Tension: Arousal Number Peaks Per Minute*

With everything mentioned in mind, the chosen FSA is the Best First. It presents both a low error rate and a low number of features selected. The subsequent construction of the models will focus on comparing different classifiers.

4.3 Model Creation

Some data portraying the error values of the several Classifiers used are presented in Tables 18 and 19.

	LinearRegression	M5P	MLP	Average
Imm	1,22896	1,19948	1,741577	1,390006
Ten	1,376419	1,427436	2,177174	1,660343
Comp	2,270985	2,186488	3,290406	2,582626
Chall	1,732424	1,67693	2,300696	1,90335
Flow	2,063282	1,914901	3,14331	2,373831
Fun	0,889803	0,94905	1,304085	1,047646
Average	1,593646	1,559048	2,326208	1,8263

Table 18: RMSE over Classifiers

		Imm	Ten	Comp	Chall	Flow	Fun
GType	Hard	1,623915	0,917246	2,431235	0,925988	1,605507	1,202415
	NA	0,663444	2,277615	2,38055	0,85707	2,045486	0,546811
	Soft	1,584162	1,879106	2,820788	2,920685	2,594046	1,090611
Likes	N	1,708644	1,685748	2,575817	2,032455	2,4538	1,429605
	NA	0,663444	2,277615	2,38055	0,85707	2,045486	0,546811
	Y	1,76493	1,256505	3,233416	2,944377	2,083295	1,027937
Sex	F	1,84039	1,80568	2,825053	2,668803	2,728756	1,356089
	M	1,294685	1,610251	2,162014	1,650839	1,677035	1,040158
Cond	NB	1,774829	1,711753	3,059622	2,078954	2,839993	1,603033
	NV	1,314722	1,899903	2,325132	1,7849	2,231661	0,785262
	V	1,056901	0,942352	2,214715	2,21571	3,807077	0,895372

Demographic Study

Table 19: Classes RMSE over Demographic Segmentations

The most prominent conclusion can be made with the analysis of Table 18. Regarding the classifiers, MP presents a larger error rate in all classes. In general, M5P presents better results throughout except in the "Tension" and "Fun" classes. Concerning the overall classifiers results for the demographic segmentations, to note some results that can give some insight into this relationship:

- *Tension reports very small error rates for "Hardcore" Gamers and for the "Visible Biofeedback" game session.*
- *Challenge varies vastly between Gamer Types, having "Hardcore" player better results. Additionally, the "Non-Visible Biofeedback" variation of the game presents smaller errors than their counterparts.*
- *Flow presents two extreme error values. A minimum for "Hardcore" players and a maximum for the "Visible Biofeedback" game variation.*
- *Fun shows great differences for the gameplay conditions. "Non-Biofeedback" presents the larger error value.*

These relationships between demographic/physiological data and reported user experience can potentially lead to a better understanding of how the emotional states of a player can affect its gaming experience. This insight can be used to better apply the players' affective reaction models, in an attempt to provide a better user experience. However, these relationships were not very consistent with the addition of not being easily translated to better gaming experiences. As such, these models were not directly used in posterior phases.

Chapter 5

Simulator

The construction of the previously discussed models serves as a means to an end. Our goal is to use these models to accurately describe players' emotional responses to individual game events. Their posterior usage will be done by a new software tool, designed to allow game designers to construct target experiences and subsequently discover the optimal way in which to elicit them. Because this process is bound to happen during the game testing phase, it works as both a discovering and debugging emotional response tool. The tool created was nicknamed GODx (Game Optimal Design eXperience). It works in conjunction with a symbolical game simulator in order to find the best way to elicit the desired emotional states.

5.1 GODx

The main goal of this tool is to give game developers a simple yet powerful way to describe a players' emotional state along time. This brings a change in the way the game is developed. The creation of these emotional states must be explicitly stated, and the posterior discovery of the necessary flow of events is calculated automatically. Gaming as a form of transmitting emotions shifted from intuition/experience that derived from the game developers, to a more precise way.

The tool itself has been divided in three big modules: the emotional state representation, the debug/live module and the options area. Although GODx has been conceptualized as a general purpose tool, allowing it to be adapted to all games, due to time constraints, the developed GODx does not have all the functionalities implemented. A description of the three modules will follow, depicting both the guidelines and principles desired and the actual work made.

5.1.1 Emotional State Representation

The Emotional State Representation model provides a way to accurately and easily represent and visualize any desired emotional states along time. Due to it being based on (Posner et al., 2005) circumplex model of affect, that presents two dimensions, and the need to express this another dimension (time), the representation used is of a dual two-dimensional graphic (Figure 7). Each graphic holds the information of one of the circumplex dimensions along time. Because the desired emotional state is commonly cyclic, a time is provided, representing the total time of the experience drawn. This means that the first and the last time periods must be automatically equivalent to allow a flow in emotion. Additionally, if the total time of the experience created is lower than the total simulation time, the simulator must take this into consideration and create a cyclical behavior.

The way devised to create this emotional state over time employs Quadratic Bezier curves. These curves have some necessary properties: they are injective (there is a unique emotional state for any time), they are continuous (emotional state changes pass through all intermediary states) and easy to understand and work. A way to easily add, remove and edit the first and last control points and to edit the intermediate control point is all that is necessary to provide a way to easily create the desired emotional states. Because of the way it was formulated, a desired emotional state can be straightforwardly added to a desired timestep. To complement on this idea, and viewing that the circumplex model of affect is not known by all game developers, information relating a series of emotional keywords were added, giving the possibility to add a desired emotional state at any desired time by simply choosing its emotional keyword. The emotional keywords and respective valence and arousal values were taken from (Hepach, Kliemann, Grüneisen, Heekeren, & Dziobek, 2011). This work, which gathered data from 100 participants, collected the reported arousal and valence values of 62 emotional keywords. By providing these emotional keywords to the game developers, a more precise and easy way to create the desired emotional states is possible.

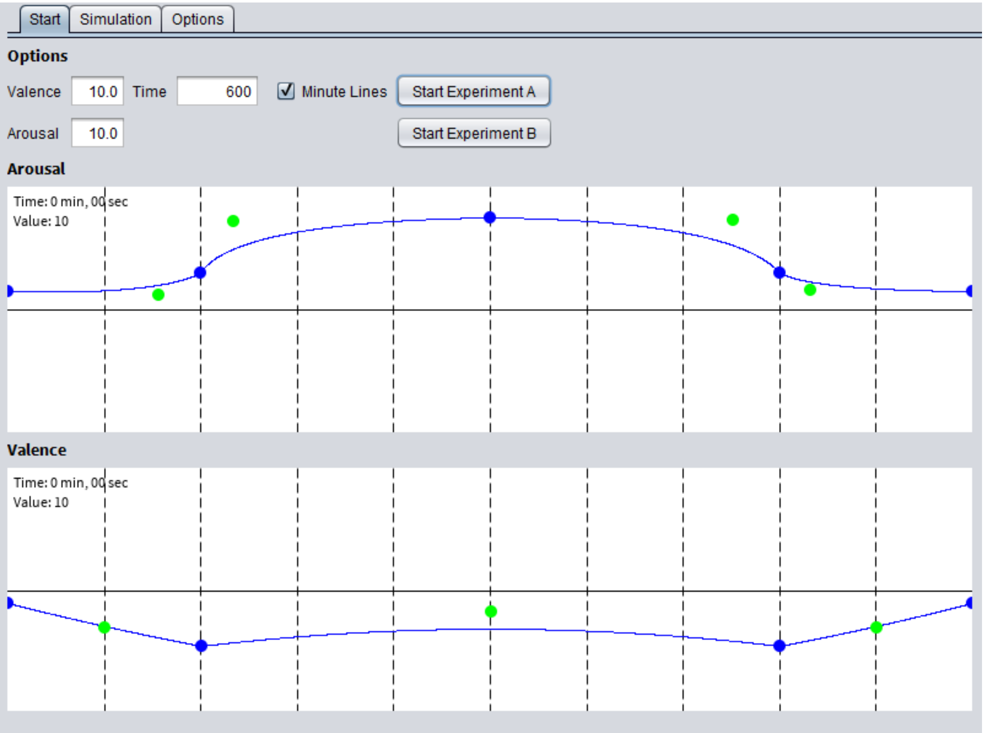

Figure 7: Sample of GODx emotional experience representation

As it is visible on Figure 7, complex emotional states over time can be easily represented and understood through this tool. Additionally, a range for all dimensions is provided (arousal, valence and time) and the possibility to include dashed lines to mark the minutes is present. This gives an overall better view of the global scope of the experience.

5.1.2 Debug/Live Module

The Debug/Live module acts as an interface between the simulations and the game developer. Due to the lack of human interaction and the nature of the data created in the simulation phase, a visual and easy way to understand these simulations was desirable. Moreover, in future applications this visualization can happen in real-time, making a game that adapts itself from direct psychophysiological input easier to understand. For this to be possible, and to be abstract enough to be present in multiple game genres, we concluded that a generic GUI should be present. As such, the formulated way was to design this interface via a XML file. Graphic primitives such as squares, circles and images and their respective positions are present, allowing a large amount of flexibility and power. For a real-time application, the game itself could be constantly updating this file.

Furthermore, a graphical representation of the circumplex model should be present, displaying both the current emotional state and the desired emotional state. Transitions that occur from each event are also highly advised to be represented.

Due to time constraints, this module was not fully implemented. The generic XML approach is not present, being replaced by a simpler version that functions solely for our case study game (Vanish). Moreover it is not connected to the data generated by the simulator.

5.1.3 Options area

Arising from a need to present to the simulator various parameters, this module is a straightforward necessity. Although it does not require any graphical representation, this module must present some form of versatility. All the options available should be game specific, and be customizable. As such, a similar formula as the previous module should be considered, a XML file that describes all the necessary options to the game being worked on. Due to the investigative nature of this work, this module is only currently functional for Vanish.

5.2 Simulator

The simulator is the backbone of GODx. It takes the models produced in early phases and uses them to discover the best ways to elicit the desired emotional response. With this automatic game sessions can be produced, being the emotional responses obtained via the clustering methods discussed. This allows to create simulations of gameplay sessions for each subject, with the addition of containing the emotional state along the experience and the correspondent succession of events. The study of these simulations is of utmost importance in order to lead the subject to the desired experience.

At its core, two distinct and different approaches are used, denominated Non Biofeedback (NBF) Experiment and Emotional Regulated Indirect Biofeedback (ER-IBF) Experiment. The first approach, views these simulations as a passive agent, external to the game engine and only visualizing the gameplay session. The simulations are exactly as the original game engine was designed, and the models are used to determine how the player would react to each game event. Thus, the predicted emotional state of the player is recorded.

ER-IBF Experiment on the other hand works as an active agent. It dynamically changes the game flow to provide emotional experiences closer to the intended target.

5.2.1 NBF Experiment

This experiment can be thought of as a simulated gameplay session. Instead of emotional responses from real players, the models calculated previously are used to determine these emotional responses. Since this phase's purpose is to determine the best arrangement of game parameters that lead to the desired experience, a large set of these combinations must be tested. In the case of our case study game, the parameters represent the probability of a certain event happening at each (discrete) point in time. The choice of the parameter values (PV's) to perform simulations on is a tough task. However, viewing that the game previously used some classical studies on how to improve players' experience and taking into consideration it is now in an alpha release, the original/vanilla PV can offer us both a robust baseline and a good starting point. The creation of new PV's is therefore handled as small deviations over the vanilla one. An increment and an interval are chosen for each parameter, taking into consideration the vanilla PV, and all possible combinations are computed based on these values. With this solution, only relatively small changes on the game are done, therefore preserving its' overall identity. Regarding the simulation of each PV, taking into consideration the games' probabilistic nature of the procedural content generation, the same PV can generate different game experiences. To reduce the variance of the obtained result, each one is reutilized several times (N = 100) in independent simulations. The final results is the average of all these simulations.

Because of the nature of the emotional response of a subject to external stimuli over time, a decay rate is necessary. This decay rate represents the speed at which a subject progresses to his baseline emotional state. However, due to the nature of the game, this baseline point is not necessarily a neutral one. This is due to the full ambience that the subject is exposed, at all times. Hence, this emotional state was determined by averaging all subjects' emotional states along all the experiences. This allows us to determine a more realistic base emotional state.

A full simulation of a single gameplay session consists of three steps: the decaying of the emotional state to the predetermined base one, the generation of Pseudo-Random Events (based on the PV) and the emotional reaction to these events. This steps are repeated for the whole experience duration.

Considering everything mentioned before, the global simulator approach is as follow: all PV's are generated and subsequently, for each one, several simulations are performed. Data relative to the type of events and all the emotional states of the subject is then recorded. However, further analysis must be done to this data in order to compare parameters values. Objective emotional states must be determined along with a way to rate an experience.

First of all, the definition of objective emotional states arises from the need to grade a specific simulation. Viewing this as an exploratory work, the intent was to have a large variety of emotional states that contrasted each other, with a vast range among the circumplex spectrum. Besides this property, static and dynamic emotional states are present. This means that, for objective emotional states along time, some static keywords such as "Confident" or "Anxious" are present, establishing the same objective emotional state throughout the simulations. On the other hand, cyclical and changing emotional states were created. These dynamic emotional states vary only on its period of repetition while presenting a dynamic target.

The creation of these dynamics keywords and how these can be used to enhance users' experience is a challenging task. These were created intuitively to try and mimic a good experience in a game of this genre. The formulated hypothesis was that cycles of arousal between the base emotional state and higher values along with cycles of valence between the base and lower values would create a pleasurable experience (Figure 7).

Posterior to the election of the desired target emotional states, a way to determine each simulation's fitness is necessary. Due to the vast emotional range present in the designated emotional keywords, and the less fluctuating emotional states obtained through the simulations (seeing the nature of the game this is to be expected), a linear function was not probably the better solution. Moreover, the main purpose is to highly differentiate the good from the bad emotional experiences. As such, a sigmoid function is used to better tackle this problem. The formalization is as follows:

Let p be an emotional state in a two-dimensional Euclidean space (circumplex model), such that $p = (arousal, valence)$. Furthermore, let d be the weighted Euclidean distance between two points $p1$ and $p2$, such that:

$$d = \sqrt{\alpha(p1_{arousal} - p2_{arousal})^2 + \beta(p1_{valence} - p2_{valence})^2}$$

Where α and β are weighting parameters for each of the Euclidean dimensions, such that $\alpha, \beta \in \mathbb{R} \wedge \alpha, \beta \geq 0 \wedge \alpha + \beta = 2$. These weighting parameters are meant to favor or penalize each dimension, according to the perceived difficulty in adjusting it. For example, valence might have a lower weight since it is harder to elicit, or is not as relevant for the desired affective experience. Since in this study we want to perform an unbiased analysis, $\alpha = \beta = 1$

The fitness of a certain point p_c to a target point p_t is given by f and is inversely correlated to its distance from p_t. Since we aimed at penalizing values further from p_t, it can be trivially concluded that a linear correlation function between distance and fitness would not be adequate. Thus, $f$is given by a sigmoid function of the form

$$f_{pc,pt} = \begin{cases} 1, & \sigma + d_{pc,pt} = 0 \\ \sigma + 1 - \dfrac{d_{pc,pt}}{\sqrt{\varphi + d_{pc,pt}^{\ 2}}}, & otherwise \end{cases}$$

With $\varphi \in \mathbb{R}$ being its exponential tuning parameter and $\sigma \in \mathbb{R}$ a threshold value.

5.2.2 ER-IBF Experiment

The previous experience could be described as purely observational. The simulations are performed with no changes to the original game engine, and the emotional states of the players are recorded and afterwards compared. This experience tries to work as an active agent, performing changes to the original engine in order to create an adaptive dynamic experience. Whereas earlier the events were randomly generated according to the parameters values currently being calculated, here events are calculated by taking in consideration both these values and the fitness value. This means that a target emotional state is required to actually run this experience, viewing that the selection of events that unfold has its basis on the fitness value. Note however that this new way of generating events is not performed at all timesteps. To avoid over fitting and to maintain the character of the parameters values, dynamically generated events are only allowed at a fixed minimum number of timesteps - with this number being parameterisable. On the remaining timesteps where dynamic events are not allowed (non-dynamic timesteps), the simulator works identically as in NBF Experience. A high-level description of the process of determining the best course of events to happen is presented:

1. *All possible combination of events are generated, sorted in ascending order by number of events*
2. *For each of these events combinations, a distance to the target emotional state is calculated*
    a. *If lower than the minimum distance, record this as the minimum distance*
    b. *If lower than a designated threshold save its reference*
3. *For all combination of events that are within the threshold, maintaining the previous order*
    a. *Determine if it will be selected, being its probability based on the current PV.*
4. *If no combination of events was nominated, return the combination of events that presented the minimum distance*

Because of the way the events are determined, a small number of events are favored and the possibility of them occurring is correlated to the current parameters values. This conserves part of the games' original individuality, while still trying to provide better experiences. To note, however, that if no set of events meets this condition, the one that maximizes the fitness is used, disregarding the probability of it happening.

Regarding the overall objective of the experiment, because its goal is to verify the validity of a dynamically generated game that tries to maximize the users' experience, not all previous sets of game parameters are used. The used ones are obtained from the previous experiment, being a total of three: the vanilla PV, the globally best PV, and the individually best PV. The first one is constant, representing the original set of parameters values. The second one is constant for each emotional state desired, and represents the PV that had the best average fitness for all subjects. Lastly, the third one is the best PV for the subject being simulated regarding the desired target emotional state.

5.2.3 Results

As stated above, a large range of results were obtained from the simulations made. Fitness, Arousal and Valence levels over time were all recorded. This data can later be used to produce summaries of these values throughout the whole experiences and PV's whether for individual players, clusters or for the global population. Additionally, events triggered and respective emotional reactions are also recorded.

To better understand these results, several plots and tables depicting the fitness, arousal, valence and other related values will be presented.

To note that all plots that represent data of a particular set of parameters values over time, are the average of all the simulations made for that set.

Regarding NBF Experiment, a study of its fitness along all the parameters values provides valuable information about the increase in effectiveness.

	Anxious	Bored	Concerned	Confident	Confused	Desperate	Enthusiastic	Frantic
Best	0,6548	0,6004	0,9394	0,3019	0,8537	0,4739	0,4435	0,4045
25%	0,6437	0,5932	0,9353	0,2976	0,8456	0,4656	0,4354	0,3955
Half	0,6405	0,5888	0,9336	0,2960	0,8424	0,4631	0,4323	0,3918
75%	0,6374	0,5815	0,9313	0,2934	0,8380	0,4606	0,4297	0,3894
Worst	0,6270	0,5592	0,9261	0,2865	0,8269	0,4529	0,4220	0,3830
Vanilla	0,6412	0,5899	0,9334	0,2950	0,8447	0,4640	0,4300	0,3929

	Frustrated	Jumpy	Proud	Shocked	Surprised	Triumphant	2min	3min	5min
Best	0,5295	0,6562	0,3985	0,5455	0,7112	0,5730	0,6957	0,7063	0,6801
25%	0,5224	0,6444	0,3910	0,5349	0,6955	0,5606	0,6899	0,7009	0,6725
Half	0,5202	0,6404	0,3883	0,5311	0,6901	0,5562	0,6881	0,6991	0,6698
75%	0,5172	0,6373	0,3849	0,5282	0,6842	0,5516	0,6860	0,6970	0,6675
Worst	0,5086	0,6278	0,3744	0,5197	0,6666	0,5378	0,6791	0,6903	0,6600
Vanilla	0,5215	0,6406	0,3866	0,5318	0,6857	0,5529	0,6881	0,6990	0,6700

Table 20: NBF Experience Improvements along sorted PV's

Due to the large number of set of parameters used in the simulations and since the goal is to perceive discernible differences in fitness values, only a few PV's are shown. As such, all PV's are ordered by their mean fitness value and presented here are the quintiles of this list. The vanilla parameters values are shown for comparison purpose. The emotional keywords "2min", "3min" and "5min" are the dynamic keywords, being that the number represents the period of the cyclical behavior.

As can be seen in Table 20, the vanilla parameter vector shows a similar performance to the median parameter set. This can be explained due to the fact that the set of all parameters values were generated as deviations from the vanilla PV. Moreover, the range of fitness values for each of these emotional keywords varies more largely when this fitness value is neither in the lower or upper ranges of the whole spectrum. Keywords such as "Enthusiastic" and "Shocked" reveal a larger range of fitness than, for example “Concerned”. This is as a direct product of the fitness function being of the sigmoid type, expanding small deviations in the median ranges and compressing large deviations in both the lower and upper ranges.

Regarding emotional keywords, some observations can be made. First, the low levels of fitness demonstrated by “Confident”, “Proud”, “Triumphant” and other keywords provide incentive results. These are emotional states not usually present in games of this genre, and their low fitness levels are both to be expected and a good indication. On the other hand, keywords such as “Concerned”, “Confused” and “Surprised” that are common in the atmosphere of horror games, showcase a high degree of fitness.

A more in-depth analysis can relate the whole AV emotional space to the fitness levels. Emotional keywords present in the second quadrant (High Arousal and Low Valence) of the circumplex model, the quadrant where the base emotional state is contained, expose significantly higher fitness values. This relates to the nature of the game. Being a horror game, it is very hard to elicit on a player high levels of valence (“happy” feelings) and low levels of arousal (“tranquility” mood). Both the events triggered and the base emotional state lead the players to this area of the AV emotional space, consequently revealing higher fitness values.

This means no PV can, over several simulations, display dynamic behavior.

5.2.4 Experiences Comparison

One simple way to compare the two experiments is to visualize the fitness of both throughout the game sessions’. Some samples follow:

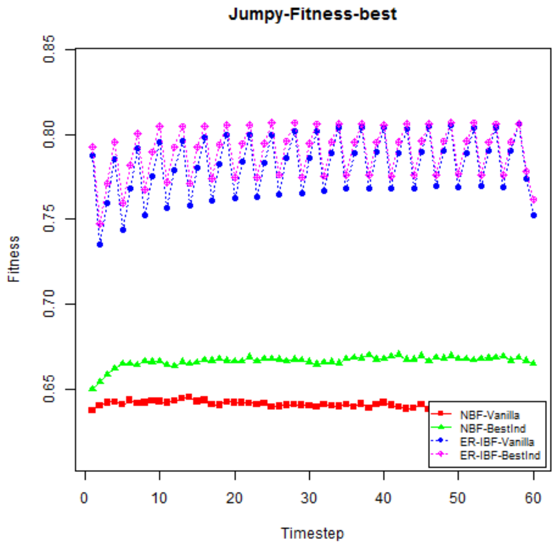

Figure 8: Fitness over time Comparison

The above plots show the average fitness values of all subjects for each desired emotional state. For the dynamic keywords, the name represents the period of the cyclical emotional states. Each plot portrays the vanilla and the individually best parameter vectors for each experiment. This allows us to compare the improvement of fitness of each experience over the vanilla PV. In addition, both experiences can be compared in fitness values. As can be seen, ER-IBF Experiment displays overall better results. This is to be expected viewing that it intermittently choses a better series of events. Similarly, the difference between the Vanilla and the individually best parameters values is evident on both experiments, a notorious increase in fitness is visible.

With a closer look, some perceivable jaggy behavior can be seen on Experiment B. This is not an experimental error. This relates to the way these experiences work. As stated above, ER-IBF Experiment works similarly to NBF Experiment, being the only difference the less random way of choosing events that occurs at certain timesteps. These timesteps, presented in the above plots, occur at a distance of three. This is linked to the period of the jaggy behavior in the plots above. When on these timesteps, fitness rises considerably due to a better choice of events. Any other time, the fitness decays to that of the shown in NBF Experiment. This result can be seen in both Arousal and Valence dimensions as well. A sample of Arousal values over time follows.

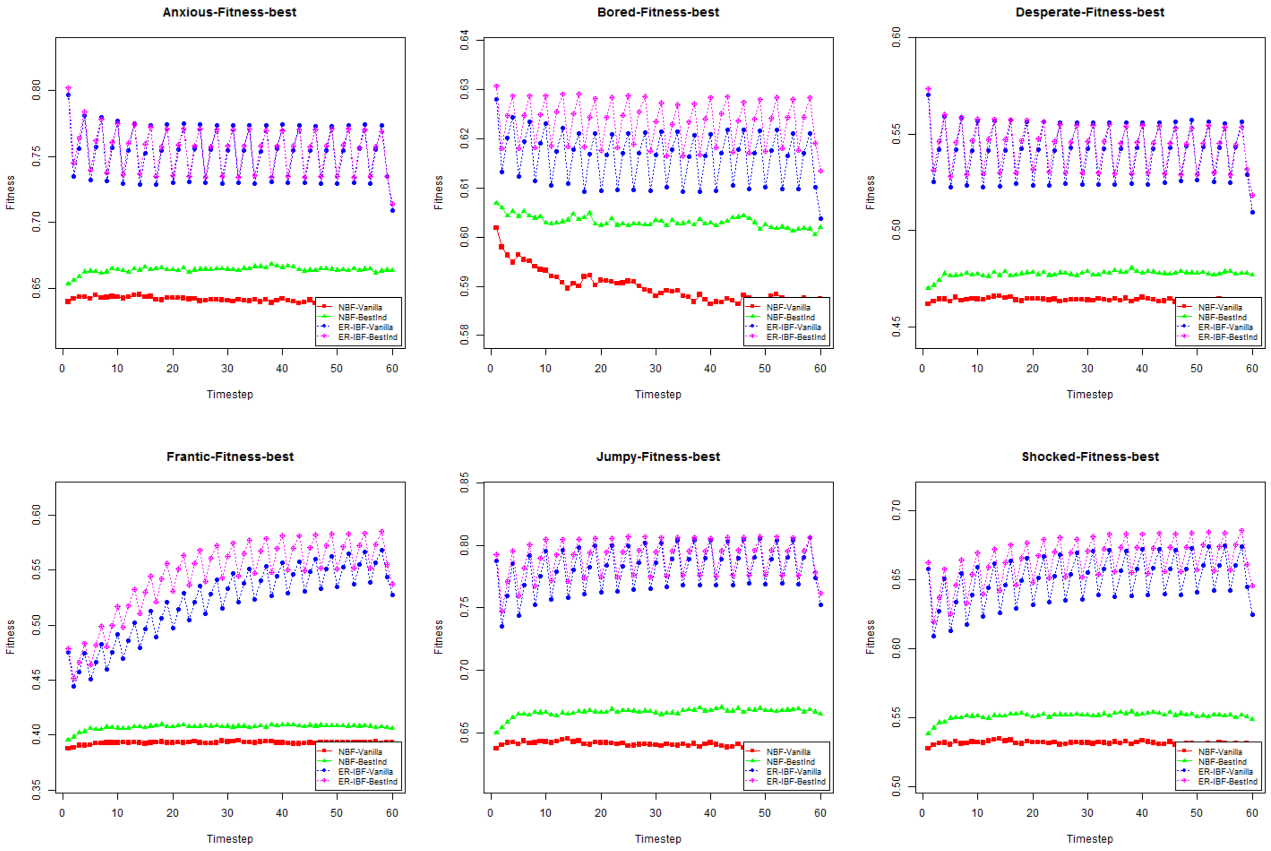

Figure 9: Arousal over time Comparison

As can be seen, the same behavior occurs and is to be expected. It is intrinsic to the technique employed. It is because of this sharp increase in fitness that the experiment displays an overall better performance.

Concerning experiments comparison, one of the best ways to analyze this duality is to link their mean fitness's.

Simulator

	Anxious				Bored				Concerned			
	ERIBF		NBF		ERIBF		NBF		ERIBF		NBF	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Vanilla	0,7534	0,0829	0,6412	0,0473	0,6165	0,0336	0,5899	0,0318	0,9462	0,0409	0,9334	0,0418
Best	0,7597	0,0762	0,6548	0,0551	0,6237	0,0282	0,6004	0,0265	0,9511	0,0367	0,9394	0,0375
Best Ind	0,7558	0,0764	0,6639	0,0576	0,6235	0,0281	0,6031	0,0270	0,9412	0,0429	0,9536	0,0399

	Confident				Confused				Desperate			
	ERIBF		NBF		ERIBF		NBF		ERIBF		NBF	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Vanilla	0,3163	0,0234	0,2950	0,0141	0,8853	0,0770	0,8447	0,0574	0,5413	0,0579	0,4640	0,0311
Best	0,3229	0,0229	0,3019	0,0148	0,8926	0,0739	0,8537	0,0573	0,5457	0,0520	0,4739	0,0358
Best Ind	0,3227	0,0227	0,3023	0,0149	0,8847	0,0774	0,8667	0,0584	0,5441	0,0528	0,4774	0,0379

	Enthusiastic				Frantic				Frustrated			
	ERIBF		NBF		ERIBF		NBF		ERIBF		NBF	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Vanilla	0,4707	0,0467	0,4300	0,0239	0,5208	0,0647	0,3929	0,0262	0,5931	0,0715	0,5215	0,0357
Best	0,4800	0,0444	0,4435	0,0277	0,5442	0,0702	0,4045	0,0323	0,5929	0,0658	0,5295	0,0398
Best Ind	0,4797	0,0447	0,4447	0,0267	0,5441	0,0699	0,4072	0,0337	0,5884	0,0666	0,5349	0,0403

	Jumpy				Proud				Shocked			
	ERIBF		NBF		ERIBF		NBF		ERIBF		NBF	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Vanilla	0,7820	0,0755	0,6406	0,0468	0,4226	0,0407	0,3866	0,0231	0,6504	0,0668	0,5318	0,0380
Best	0,7927	0,0707	0,6562	0,0563	0,4310	0,0383	0,3985	0,0241	0,6628	0,0622	0,5455	0,0452
Best Ind	0,7900	0,0719	0,6663	0,0602	0,4315	0,0380	0,3994	0,0251	0,6642	0,0638	0,5513	0,0492

	Surprised				Triumphant			
	ERIBF		NBF		ERIBF		NBF	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Vanilla	0,7560	0,0816	0,6857	0,0459	0,6118	0,0677	0,5529	0,0365
Best	0,7703	0,0723	0,7112	0,0455	0,6233	0,0610	0,5730	0,0370
Best Ind	0,7699	0,0735	0,7132	0,0490	0,6234	0,0619	0,5746	0,0397

Simulator

	2min				3min				5min			
	ERIBF		NBF		ERIBF		NBF		ERIBF		NBF	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Vanilla	0,7524	0,1733	0,6881	0,2122	0,7724	0,1686	0,6990	0,2120	0,7797	0,1534	0,6700	0,2010
Best	0,7580	0,1602	0,6957	0,2057	0,7798	0,1554	0,7063	0,2059	0,7958	0,1397	0,6801	0,1962
Best Ind	0,7554	0,1550	0,7022	0,2029	0,7770	0,1503	0,7141	0,2037	0,7942	0,1357	0,6875	0,1937

Table 21: Fitness Comparison

Two important facts can be inferred from Table 21. First, ER-IBF Experiment shows significantly better fitness results across all emotional keywords. Second, the standard deviation presented is smaller on ER-IBF Experiment, further increasing the improvement shown in the fitness department. This means that the experience is closer to the intended one and is largely more stable.

To note the behavior ER-IBF Experiment fitness values exhibit. Their range vary from keyword to keyword, similar to NBF Experiment. As previously stated, this is due to the sigmoid nature of the fitness function.

To get a global comparison between experiences, a more general overview can be calculated by averaging the fitness values throughout the emotional keywords. However, due to the nature of the emotional keywords present, the distinction between dynamic and static keywords was maintained. As has been noted, static keywords refer to emotional states that remain constant throughout the whole simulations. Dynamic keywords present a varying and cyclical experience.

	Static Keywords			
	ERIBF		NBF	
	Mean	SD	Mean	SD
Vanilla	0,6333	0,0594	0,5650	0,0357
Best	0,6423	0,0553	0,5776	0,0382
Best Ind	0,6402	0,0565	0,5828	0,0400

	Dynamic Keywords			
	ERIBF		ERIBF	
	Mean	SD	Mean	SD
Vanilla	0,7682	0,1651	0,6857	0,2084
Best	0,7779	0,1518	0,6940	0,2026
Best Ind	0,7755	0,1470	0,7013	0,2001

Table 22: Fitness Global Comparison

Table 22 allows us to draw several conclusions. First, the increase in fitness is larger for dynamic keywords. However this increase comes at a cost. Due to the fluctuating nature of these keywords, the standard deviation presented is vastly larger. The bigger increase in fitness values for dynamic keywords may happen due to the less ample emotional range needed. Some static keywords showed low levels of fitness because the desired emotional states were very hard to elicit for this specific game (which is understandable as we did not expect any combination of game events to elicit some emotional states).

Besides this dynamic versus static emotional keyword comparison, it can be noted that ER-IBF Experiment has overall higher fitness values, as well as a lower associated standard deviation, independently of the type of emotional experience desired. This means that, not only the Experiment presents better overall results, this increase in fitness comes from "tighter" results, less erratic over time. This decrease in standard deviation is even more important on the dynamic keywords, viewing that a decrease in this value indicates the emotional experience had the same dynamic behavior.

Since fitness is an abstract concept that does not distinguish between the involved emotional dimensions, its analysis is not sufficient to understand the obtained results. As such, Figures 10 and 11 represents the observed differences in arousal and valence to the target emotional states over time for each emotional keyword on both static and dynamic emotional regulation experiments.

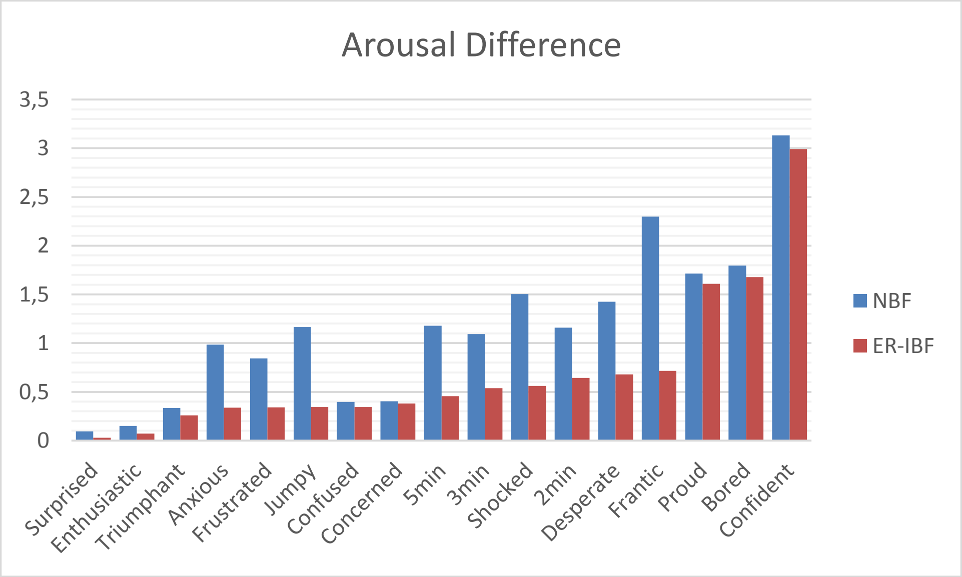

Figure 10: Arousal Difference

Simulator

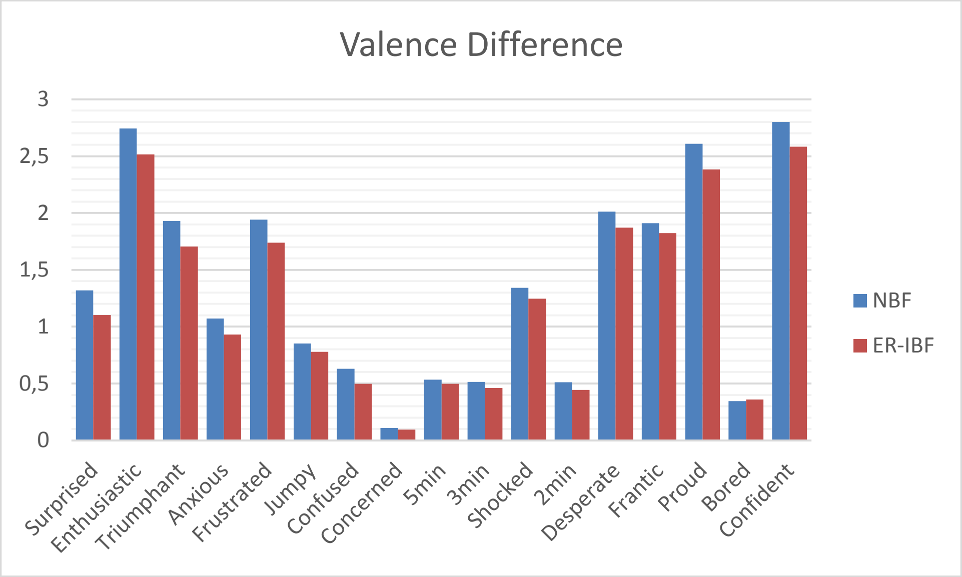

Figure 11: Valence Difference

The bar charts in Figures 10 and 11 portray the difference between the simulated and the desired emotional state (i.e. lower values mean better results). With this in mind, it seems evident that ER-IBF Experiment produces affective experiences closer to the desired one. However, on closer inspection, more insightful conclusions can be drawn. Firstly, arousal presents larger variances between experiences. In fact, the valence dimension expresses almost identical results between experiences. This happens due to the difficulty in eliciting notorious valence responses, originated by the type of game and the players' lack of valence expression. On the other hand, arousal does not seem to suffer from this problem. An interesting observation was that arousal displays the behavior of presenting considerably similar values when the desired emotions have globally low levels. The most predominant cases are the “Proud”, “Bored” and “Confident” keywords. These keywords have low arousal values and as such, are hard to elicit, resulting in lower fitness values.

In order to validate and more closely examine the aforementioned results, plots that show mean arousal/valence values and respective distances to the desired emotion throughout the simulations were created. These can provide a visual aid to the players' emotional reactions and thus constitute a valuable asset in both online and offline future applications of this technology. Some illustrative examples are presented in the following sub-section.

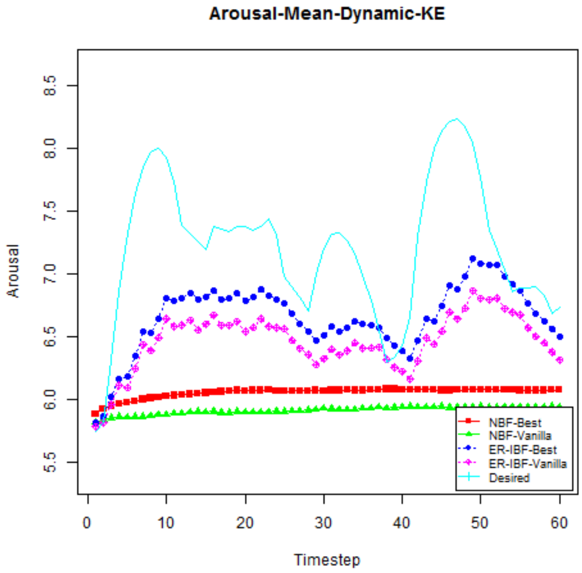

Figure 12: Arousal Mean over time

The above plot shows the mean arousal values for all dynamic keywords combined. One can easily notice the better results of the ER-IBF Experiment and perceive its adaptive behavior. This adaptive behavior is of extreme importance for dynamic experiences.

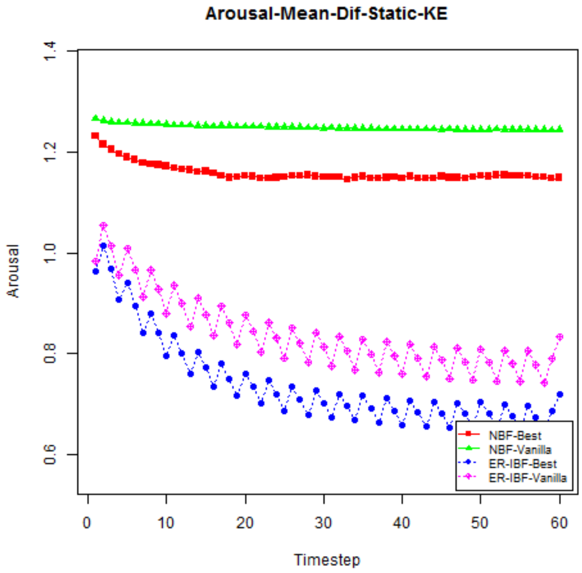

Figure 13: Arousal Mean Difference over time

The distance to the desired emotional state for all static emotional keywords is presented above in Figure 13. Besides the expected better performance for ER-IBF Experiment, note the progression of the time series. Because the desired emotion is a static value, in both experiments a state of equilibrium is reached after some time. An initial phase can be perceived that serves as a bridge to the point where this equilibrium is reached, meaning that the decay rate and the emotional responses balance themselves.

Given the improvements observed in ER-IBF experiment, we also became interested in how much its intrusiveness value could influence the overall obtained fitness values. Recall that intrusiveness is the period at which events are dynamically generated through the player's emotional reaction model instead of through the game's parameter vector.

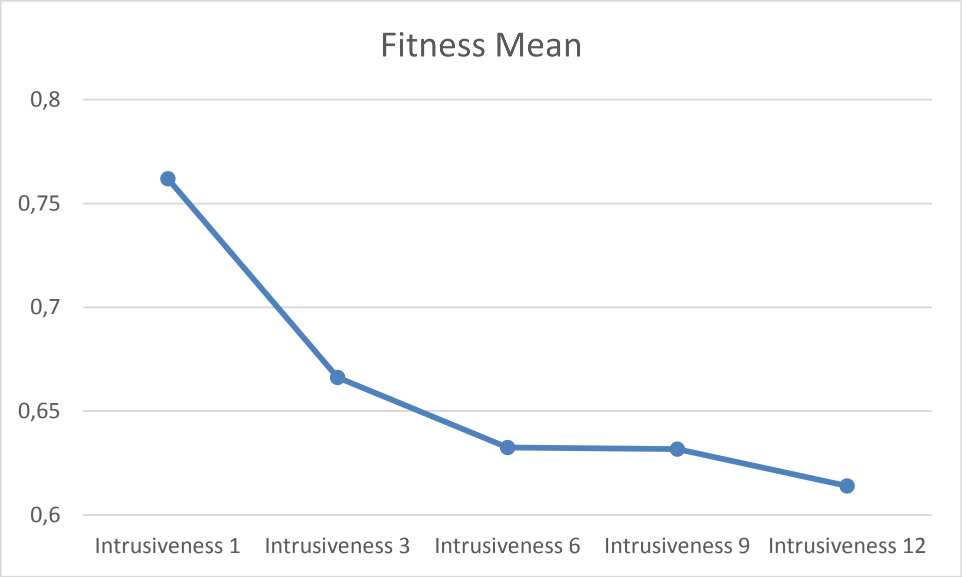

Figure 14: Fitness Mean Over Intrusiveness

As expected, the lower the intrusiveness the better are the results. This indicates that dynamically generating game events provides better results than random ones. A quick analysis of Figure 14 reveals an inverse exponential correlation function between overall fitness and intrusiveness. Despite this, it would be advisable to exercise caution in interpreting these results on a practical application, as maximal intrusiveness (adjusting the game every 10 seconds) can quickly make the gameplay too erratic or hectic, which would not be necessarily good. In our case study this could result in overfitting, thus biasing our results. As such, this was a major factor in choosing a lower intrusiveness value (30 seconds). In a real-world application we would advise a mix of intrusiveness levels 3 and 6 (3 for more phasic game events such as enemy and item spawns and 6 for more tonic level parameters such as level geometry and atmosphere).

Chapter 6

Conclusions and Future Work

The main goals of this work were two-fold: the creation of accurate affective reaction models that could satisfactorily predict players' emotional responses to in-game events and the posterior use of these models for parameterisable and adaptive affective gaming.

The models initially created via Machine Learning showed error rates higher than desired. The fact that gathering emotional reactions through psychophysiological data has some limitations and the singular way these models were created, probably lead to these not sought after results. By simulating the human world, where groups of people show similar emotional responses, some encouraging results were extracted via a clustering approach.

The subsequent use of these models for both the discovery of the best set of parameter values and the creation of dynamic affective gaming experiences provided some good results. It was shown that small improvements can be done to the original game parameters for a few selected objective emotional states. These improvements are largely increased if a dynamic system is implemented, where the game itself progresses having into account both the current players emotional state and the desired one. The implementation shown provided a proof of concept that can be used in several other domains. As such, the main goals of the work were achieved.

6.1 Future Work

Being that this thesis presented a proof of concept and due to its exploratory nature, not many developments can be made over the developed work at the moment in order to make this process commercially viable. However, future research on this subject that implements a full automatic extraction of biofeedback data, allowing for the extraction of emotional data several magnitudes larger is of great value. Unfortunately, due to the nature of psychophysiological data, this is currently unfeasible. The analysis of a greater dataset could allow the reach of new conclusions regarding the construction of the affective reaction models. The same conclusion can be made regarding the study of the relationship between related User experience and physiological and non-physiological features.

Another improvement over this work would be to fully abstract all the implemented features, allowing for the use in a wider range of games and even in other domains.

References

Ambinder, M. (2011). Biofeedback in gameplay: How Valve measures physiology to enhance gaming experience. *Game Developers Conference*. Retrieved from http://www.modexpo.com/publications/2011/ValveBiofeedback-Ambinder.pdf

Bersak, D., McDarby, G., & Augenblick, N. (2001). Intelligent biofeedback using an immersive competitive environment. *Proceedings of the UbiComp Workshop on Designing Ubiquitous Computing Games*. Retrieved from http://medialabeurope.org/mindgames/publications/publicationsAtlanta2001rev3.pdf

Bidarra, M. (2013). A Generic Method for Classification of Player Behavior, 2–8. Retrieved from http://photon.marlonetheredge.name/paper.pdf

Conati, C. (2002). Probabilistic assessment of user's emotions in educational games. *Applied Artificial Intelligence*, 1–20. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/08839510290030390

Dekker, A., & Champion, E. (2007). Please biofeed the zombies: enhancing the gameplay and display of a horror game using biofeedback. *Proc. of DiGRA*, 550–558. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.190.1500&rep=rep1&type=pdf

Drachen, A., & Nacke, L. (2010). Correlation between heart rate, electrodermal activity and player experience in first-person shooter games. *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games*. Retrieved from http://dl.acm.org/citation.cfm?id=1836143

Ermi, L., & Mäyrä, F. (2005). Fundamental components of the gameplay experience: Analysing immersion. *Worlds in Play: International Perspectives on …*.

Figueiredo, R., & Paiva, A. (2010). "I Want to Slay That Dragon!"-Influencing Choice in Interactive Storytelling. *Interactive Storytelling*, 26–37. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-16638-9_6

Gilleade, K., Dix, A., & Allanson, J. (2005). Affective videogames and modes of affective gaming: assist me, challenge me, emote me. *Proceedings of DIGRA'2005*, 1–7. Retrieved from http://comp.eprints.lancs.ac.uk/1057/

Gratch, J., & Marsella, S. (2005). Evaluating a computational model of emotion. *Autonomous Agents and Multi-Agent Systems*. Retrieved from http://link.springer.com/article/10.1007/s10458-005-1081-1

Haag, A., Goronzy, S., Schaich, P., & Williams, J. (2004). Emotion recognition using bio-sensors: First steps towards an automatic system. *Affective Dialogue Systems*. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-24842-2_4

Hall, M. (1999). Correlation-based feature selection for machine learning, (April). Retrieved from https://www.lri.fr/~pierres/donn%E9es/save/these/articles/lpr-queue/hall99correlationbased.pdf

Hepach, R., Kliemann, D., Grüneisen, S., Heekeren, H. R., & Dziobek, I. (2011). Conceptualizing emotions along the dimensions of valence, arousal, and communicative frequency - implications for social-cognitive tests and training tools. *Frontiers in Psychology*, *2*(October), 266. doi:10.3389/fpsyg.2011.00266

Hudlicka, E. (2008). What Are We Modeling When We Model Emotion? *AAAI Spring Symposium: Emotion, Personality, and …*. Retrieved from http://www.aaai.org/Papers/Symposia/Spring/2008/SS-08-04/SS08-04-010.pdf

Hudlicka, E. (2009). Affective game engines: motivation and requirements. *Proceedings of the 4th International Conference on …*. Retrieved from http://dl.acm.org/citation.cfm?id=1536565

Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., & Walton, A. (2008). Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies*, *66*(9), 641–661. doi:10.1016/j.ijhcs.2008.04.004

Joe H . Ward, J. . (1963). Hierarchical Grouping to Optimize an Objective Function. *American Statistical Association Stable*, *58*(301), 236–244.

Kalyn, M., Mandryk, R. L., & Nacke, L. E. (2011). Biofeedback Game Design : Using Direct and Indirect Physiological Control to Enhance Game Interaction, 103–112.

Leite, I., & Pereira, A. (2010). Closing the loop: from affect recognition to empathic interaction. *3rd Internation Workshop in Affective Interaction in Natural Environments*. Retrieved from http://dl.acm.org/citation.cfm?id=1877839

Leon, E., Clarke, G., Callaghan, V., & Sepulveda, F. (2007). A user-independent real-time emotion recognition system for software agents in domestic environments. *Engineering Applications of …*, *20*(3), 337–345. Retrieved from http://www.sciencedirect.com/science/article/pii/S0952197606001011

Mandryk, R. L., & Atkins, M. S. (2007). A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International Journal of Human-Computer Studies*, *65*(4), 329–347. doi:10.1016/j.ijhcs.2006.11.011

Martínez, H., Garbarino, M., & Yannakakis, G. (2011). Generic physiological features as predictors of player experience. *Affective Computing and …*. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-24600-5_30

Nijholt, A., & Tan, D. (2007). Playing with your brain: brain-computer interfaces and games. *Proceedings of the International Conference on …*. Retrieved from http://dl.acm.org/citation.cfm?id=1255140

Nogueira, P. (2013). Towards Dynamically Crafted Affective Experiences Through Emotional Response Player Modelling. *Proceedings of the Ninth Annual AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *2013*(2009), 17–20.

Nogueira, P., Aguiar, R., Rodrigues, R., & Oliveira, E. (2014a). Computational Models of Players' Physiological-based Emotional Reactions: A Digital Games Case Study. *Proceedings of the 2014 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*.

Nogueira, P., Aguiar, R., Rodrigues, R., & Oliveira, E. (2014b). Designing Players ' Emotional Reaction Models : A Generic Method Towards Adaptive Affective Gaming. *Information Systems and Technologies (CISTI)*.

Nogueira, P., & Rodrigues, R. (2013). Guided emotional state regulation: Understanding and shaping players' affective experiences in digital games. *... and Interactive Digital ...*, (Fairclough 2009), 51–57. Retrieved from http://www.aaai.org/ocs/index.php/AIIDE/AIIDE13/paper/viewFile/7367/7588

Nogueira, P., Rodrigues, R., Oliveira, E., & Nacke, L. E. (2013a). A Hybrid Approach at Emotional State Detection: Merging Theoretical Models of Emotion with Data-Driven Statistical Classifiers. *Proceedings of the 2013 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (pp. 253 – 260)*. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6690797

Nogueira, P., Rodrigues, R., Oliveira, E., & Nacke, L. E. (2013b). A Regression-based Method for Lightweight Emotional State Detection in Interactive Environments. *XVI Portuguese Conference on Artificial Intelligence (EPIA). Angra Do Heroísmo, Açores, Portugal.* Retrieved from http://paginas.fe.up.pt/~niadr/PUBLICATIONS/2013/2013_EPIA.pdf

Nogueira, P., Torres, V., & Rodrigues, R. (2013). Automatic Emotional Reactions Identification : A Software Tool for Offline User Experience Research. *In Entertainment Computing--ICEC 2013 (pp. 164–167). Springer Berlin Heidelberg.*

Pedersen, C. (2010). Modeling player experience for content creation. *Computational Intelligence and AI in Games, IEEE Transactions*, *2*(1), 54–67. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5420018

Posner, J., Russell, J. a, & Peterson, B. S. (2005). The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, *17*(3), 715–34. doi:10.1017/S0954579405050340

Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The Motivational Pull of Video Games: A Self-Determination Theory Approach. *Motivation and Emotion*, *30*(4), 344–360. doi:10.1007/s11031-006-9051-8

Shaker, N., Yannakakis, G., & Togelius, J. (2009). Towards Automatic Personalized Content Generation for Platform Games. *AIIDE*, (Hudlicka 2008), 63–68. Retrieved from http://www.aaai.org/ocs/index.php/AIIDE/AIIDE10/paper/viewPDFInterstitial/2135/2546

Stern, R., Ray, W., & Quigley, K. (2000). Psychophysiological Recording. *Oxford University Press, USA, 2nd Ed.*

Tognetti, S., Garbarino, M., Bonarini, A., & Matteucci, M. (2010). Modeling enjoyment preference from physiological responses in a car racing game. *Proceedings of the 2010*

IEEE Conference on Computational Intelligence and Games, 321–328. doi:10.1109/ITW.2010.5593337

Yannakakis, G., & Hallam, J. (2008). Entertainment modeling through physiology in physical play. *International Journal of Human-Computer …*, (March). Retrieved from http://www.sciencedirect.com/science/article/pii/S107158190800075X

Yannakakis, G. N. (2009). Preference learning for affective modeling. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 1–6. doi:10.1109/ACII.2009.5349491

Yannakakis, G. N., & Togelius, J. (2011). Experience-Driven Procedural Content Generation. *IEEE Transactions on Affective Computing*, *2*(3), 147–161. doi:10.1109/T-AFFC.2011.6