Improving game experience using dynamic difficulty adjustment based on physiological signals

Du Nguyen

Kongens Lyngby 2014
DTU Compute-M.Sc-2014

Technical University of Denmark
DTU Compute
Building 303B, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253031
compute@compute.dtu.dk
www.compute.dtu.dk

Summary (English)

The goal of this thesis is to investigate whether dynamic difficulty adjustment (DDA) based on emotions is better than DDA based on a player's performance. DDA is a technique in computer games to change the difficulty of the game so that it meets a certain criteria. For the emotion-based DDA the goal is to keep the player in a state of Flow which was coined by Csikzentmihalyi. For the performance-based DDA the goal is to keep the player performing well without dying.

To accomplish this two DDA systems have to be built. Tetris was chosen as the game for the two DDA systems as difficulty could easily be controlled in Tetris. The emotion-based DDA system uses physiological signals to recognize emotions. Based on the literature study the physiological signals chosen are heart rate (HR), temperature and galvanic skin response (GSR). Due to equipment availability HR (based on blood volume pulse(BVP)), GSR and electromyography (EMG) of the facial muscles were used instead. These signals were measured while 5 test participants played Tetris and self-rated their valence and arousal which could be translated into emotions later. Another test was made where test participants had to look at images and self-rate their valence and arousal. This test, or validation study, was made to ensure that the Tetris results were correct. Using the data from these tests a classifier was trained using a total of 44 features. The results were not good as they ranged from 22% to 46% with a chance level of 33%. Therefore the emotion-based DDA was changed to only using heart rate as it correlated with arousal. The new DDA tried to keep the player's heart rate slightly elevated above the resting heart rate.

The performance-based DDA uses the game state to judge whether a change in difficulty is needed. The goal of the performance-based DDA is to keep the block height in Tetris in a certain level. If the block height is below the level

then difficulty is increased and vice versa.
9 participants played both DDA versions and self-rated their arousal, valence and sense of challenge. The results of this experiment showed no statistically difference between the two DDA versions although a manual inspection of the results indicated that some participants might have preferred the performance-based DDA.

Summary (Danish)

Målet for dette speciale er at undersøge om dynamisk sværhedsgradsjustering (DDA fra engelsk: dynamic difficulty adjustment) baseret på følelser er bedre, end DDA baseret på en spillers præstation i spillet. DDA er en teknik, som bruges i spil til at ændre sværhedsgraden, så spillet opfylder visse kriterier. For følelsesbaseret DDA er målet at holde spilleren inde i en Flow tilstand, som er et udtryk for en sindstilstand defineret af Csikzentmihalyi. For præstationsbaseret DDA er målet at sørge for at spilleren yder sit bedste uden at lade spilleren dø. For at kunne teste dette, skal de to DDA systemer først laves. Tetris er blevet valgt som spil for de to DDA systemer, da sværhedsgraden i Tetris er let at justere på.

Følelsesbaseret DDA benytter sig af fysiologiske signaler til at genkende følelser. Baseret på literaturstudiet er hjerte rytme (HR), temperatur og elektronisk hudrespons (GSR fra engelsk: galvanic skin response) blevet valgt. På grund af tilgængeligt udstyr er HR, GSR og elektromyografi (EMG) af ansigtsmuskler benyttet. Disse signaler er blevet målt på 5 forsøgsdeltagere der spillede Tetris og bedømte deres valens og ophidselse, hvilket så senere kan oversættes til følelser. En anden test blev udført hvor forsøgsdeltagerne skulle se på nogle billeder og igen bedømme deres valens og ophidselse. Denne test var udført for at tjekke om Tetris resultaterne var korrekte. Ved at bruge data fra disse tests kunne der trænes en klassificeringsalgoritme ud fra 44 særpræg. Resultaterne var ikke så gode da de lå i området 22%-46% hvor chance niveauet var på 33%. Den følelsesbaserede DDA blev hermed ændret til kun at bruge hjerterytmen da denne var korreleret med ophidselse. Den nye DDA skulle holde en spillers hjerterytme højere end hvilepulsen.

Den præstationsbaserede DDA bruger spiltilstanden til at bedømme om sværhedsgraden skal ændres. Målet for denne DDA er at holde bloknhøjden i Tetris

oppe i et vist niveau således at hvis blokhøjden er under dette niveau, så forøges sværhedsgraden og omvendt.
9 forsøgsdeltagere spillede begge versioner af DDA og bedømte deres egen ophidselse, valens og udfordring. Resultatet af dette forsøg viste ingen statistisk forskel mellem de to DDA versioner selvom en manuel inspektion af resultater indikerede at nogle forsøgsdeltagere måske foretrak præstationsbaseret DDA.

Preface

This thesis was prepared at the department of DTU Compute at the Technical University of Denmark in fulfilment of the requirements for acquiring an M.Sc. in Digital Media Engineering. The thesis supervisors are Tobias Andersen and Michael Rose. The thesis work was carried out in the period from September 2013 to February 2014 with a workload of 30 ECTS points.

This thesis is about improving game experiences. The idea for this thesis was inspired by a set of slides I saw on Valve's website[Amb11]. This was in late 2012 and inspired by this, I approached Tobias and Michael Rose about creating a system that could balance game difficulty using EEG signals as input. After speaking with Tobias we quickly settled on doing a 5 ECTS points course where I would do a literature study to examine whether the idea was feasible. The result of that course is included in Appendix A.

Game difficulty has always interested me from a young age where I used to modify games so that the difficulty would better suit me. Without knowing, I actually applied the concepts from some of the studies mentioned in this thesis. This thesis is a natural extension of my game modding and is hopefully the beginning of a career exploring more about game experiences.

The thesis consists of 8 chapters.

Chapter 1 is an introduction to the project and what the general plan is.

Chapter 2 is a description of an earlier work I had made which provides a high level overview over emotion-recognition and the relevance to games.

Chapter 3 is a short overview of related work to the thesis and what the state-of-the-art is.

Chapter 4 is a description of the two experiments and the technical work that I have done.

Chapter 5 is a description of the method I used for carrying out the experiments and which analysis methods I used.
Chapter 6 is the results of both experiments.
Chapter 7 is the conclusion of the thesis.
Chapter 8 is a discussion of the possible future directions for both emotion-recognition but also in terms of using emotions in games.

Lyngby, 02-February-2014

Du Nguyen

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisors Tobias Andersen and Michael Rose for their guidance and patience in helping me realise this project. Especially Tobias, who I also worked with earlier, do I owe a debt of gratitude for all of his support during the last year and for teaching me how to write a good report, how to research and much more!
I would also like to thank DTU and DTU Compute for the 5 years I have spent here and for allowing me to use the facilities for my project.
My thanks also goes to Kasper Eskelund for helping me with the Biosemi equipment and showing me how to use it properly.

I would also extend my gratitude to my test participants for their time and patience while I stuck electrodes on their faces and showed them gruesome pictures in the name of science.

Last but not least I want to thank my friends and family for keeping me sane and motivated during this time.

Contents

Chapter 1

Introduction

According to a report by the Entertainment Software Association (ESA)[ESA12] the total amount of money spent on computer and video games in 2011 was $24.75 billion in the US and $67 billion globally in 2012. Meanwhile games such as GTA V have development costs that rival Hollywood movies[McL13]. These numbers indicate that games must be entertaining but what makes games fun? Ralph Koster argues that games serve as a learning tool[Kos05]. He argues that learning to overcome challenges in computer games are what makes them fun. As a result boredom occurs when games stops providing new challenges. Other game designers have echoed this sentiment with games being challenges and choices for a person to overcome. Sid Meier once defined games as "a series of meaningful choices", Ernest Adams and Andrew Rollings defined games as "one of more causally linked series of challenges in a simulated environment"[Kos05]. The enjoyment of games is however highly personal. What is fun for one person is not necessarily fun for another person. Fun and enjoyment in games has often been linked with the concept of flow[Kos05] [Che07] [SW05] [SZ04]. The concept of flow was coined by Csikszentmihalyi and Csikszentmihalyi identified eight components of flow:

- A challenging activity requiring skill
- A merging of action and awareness
- Clear goals

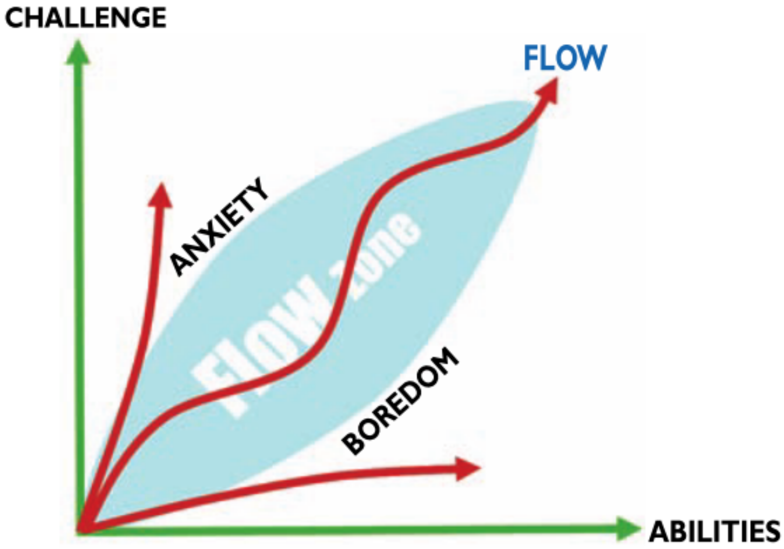


Figure 1.1: A graph with challenge on the y axis and ability on the x axis. If a person's ability is far greater than challenge it will result in boredom while a hard challenge will lead to anxiety. The figure is from [Che07]

- Direct and immediate feedback
- Concentration on the task at hand
- A sense of control
- A loss of self-consciousness
- An altered sense of time

When these eight components are fulfilled the player are in a state of flow according to Csikszentmihalyi. The concept of flow can be mapped onto a chart with difficulty of a task along one axis and skill level of a person along another axis[Che07]. This can be seen in figure 1.1. Flow occurs when the challenge matches the player's skill[Kos05]. If the player is too skilled then the player will feel the emotion boredom, and if the difficulty is too high then the player will feel the emotion anxiety. The problem is that as the player becomes more skilled while playing, the challenge has to be greater to properly match the skill of the player. Most games has so far been using preset difficulty levels such as easy, medium, hard but since the player can become more skilled while playing the game the risk is that the player will outgrow the selected difficulty. As mentioned earlier computer games are made to be fun and therefore if the difficulty is not matched with the player's individual skill the game quickly becomes frustrating or boring which decreases the fun of the game. To solve this a technique called Dynamic Difficulty Adjustment (DDA) has been invented.

A DDA system will monitor a player's progress and performance in a game and adjust the difficulty so that a player will constantly be challenged to their ability. DDA has been used in several commercial games such as Half-Life 2[Tol08], Max Payne 3[R* 12], Prey[Sie06], The Elder Scrolls IV: Oblivion[Hic11], Crash Bandicoot[SZ04], Jak and Daxter[SZ04], Left 4 Dead[Sal09] and more.

While DDA is a step towards personalizing games it still relies on a game designer's notion on how a player should be challenged. DDA systems are designed to alter difficulty based on the performance of the player but since this performance metric is designed by the game designer it does not always represent fun for everyone since fun is different for different people[Kos05]. In other words different players have different flow zones[Che07] and ideally DDA should guide each player into his own personal flow. A player can even have different flow zones based on their mood. To solve this the DDA system can take the player's emotions into account. An emotion-based DDA system would balance the game depending on the player's emotion and bring players into their own personal flow which should hopefully elicit positive emotions in the player.

The game academic Jesper Juul however argues that boredom or anxiety is not always due to a mismatch of skills and challenges[Juu10]. He argues that most people, who are not actively playing games, are not concerned by difficulty but rather time or specifically the lack of time. He writes that losing time by failing a game affects how players are entertained and that there is a balance between losing too much time (by failing too many times) and not losing enough time (not failing). Tailoring a DDA for this theory is much harder since the DDA would have to take the player's personality, game type, the time of the player into account. As such, this thesis will focus on DDA from a Flow-based point of view and investigate whether a game with emotion-based DDA will be more enjoyable than a performance-based DDA game.

To build the emotion-based DDA an appropriate method of extracting the emotions of a person has to be found. Studies have been made with classifying emotion from physiological signals. With the right physiological signals, such as heart rate, galvanic skin response and electromyography, emotions can be classified. Studies report classification rates of 84.1%[LN04] with 6 classes, 78.4%[KBK04] with 3 classes and up to 85%[RLSV06] with 5 classes. This thesis will investigate which physiological signals are most useful for emotion recognition, how emotion can be represented and how to classify emotion from physiological signals.

Two experiments are required for this thesis. The first experiment is made so that data can be gathered for how a person's physiological signal may look during different emotions. The data will be used to build an emotion classifier that will enable an emotion-based DDA game to be made. The first experiment will put the participants through different emotions so that the physiological signals can be measured for a specific emotion. The emotion will be elicited through playing games and looking at pictures. Using cross-validation the performance of the classifier can be judged. It is important to note that to ensure commercial viability the system should work out of the box. This means that the cross-validation has to separate test and training data by participants and not by samples as is customary. The cross-validation results is then representative on how the real-world performance can be.

The second experiment is to look at which game is better – one with emotion-based DDA or one with performance-based DDA. For this the emotion classifier built from the first experiment will be used. The second experiment will have participants play through both games and afterwards rate their experience. This can be used to analyze which game the participants enjoyed the most.

Chapter 2

Preliminary work

In a preliminary work I wrote a report of the possibility to create an emotion-based DDA solution. This report was written during the spring semester of 2013 at DTU under supervision of Tobias Andersen and nominated for a workload of 5 ECTS points. The report is included in Appendix A though it should not count in the review of this thesis.

In my report[Ngu] I argued that the dimensional theory of emotion should be used. The dimensional theory of emotion states that all emotions exist on a two dimensional map where the two dimensions are valence and arousal where valence is how positive/negative an event is perceived and arousal is how excited/calm a person is during an event. Another theory is the discrete emotion theory where each emotion is discrete however discrete emotions do not yield good enough classification rates to be considered.

For quantifying fun I looked at the theory of flow by Csikzentmihalyi. Csikzentmihalyi defines flow as a state where a person is in complete focus of a task and feels the loss of sense of time and consciousness. Csikzentmihalyi furthermore plotted flow in an ability/challenge map where low ability of a person and great challenge would result in anxiety, high ability and low challenge would result in boredom while the equal ability and challenge would result in the person being in a flow state. These three emotions, anxiety, boredom and flow/engagement can be plotted on a valence/arousal map which again confirms the dimensional theory of emotion as a better choice.

Deciding on which physiological measures to use, I settled on heart rate (HR), galvanic skin response (GSR) and temperature. From studies examined these physiological measures had classification rates that were 17-67 percentage points higher than chance level. This is markedly better that using electroencephalography (EEG) which had classification rates at 10-15 percentage points higher than chance level.

The choice of classifiers varied between the studies. Support Vector Machines (SVM) was the most popular as it was used in two studies. Other classifiers mentioned were K-Nearest Neighbor (KNN), Hidden Markov Models (HMM), Discriminant Function Analysis (DFA) and Marquardt Backpropagation (MBP).

For training the classifiers most studies used the Self-Assessment Manikin (SAM) which is a pictorial system and measures arousal and valence.

The method of validation is unclear in many studies. The validation of the model has to be done so that the system can be a plug n' play system because of commercial viability. That means that when validating the classifier a participant's data in the test set cannot occur in the training set. One of the studies specifically mentions this and validates using a third of the participants as the test set and the other participants as the training set which results in a classification rate of 78% and 62%, compared against a chance level of 33% and 25% respectively. Other studies does not specify how they validate and therefore the other results may not accurately predict the results that can be gained in this thesis.

The conclusion is that physiological signals such as HR, GSR, temperature worked better at emotion classification than using EEG signals. It is also important to cross-validate where data from a participant is not used in both the test data and training data. A survey of different classification algorithms reveals that SVM is a popular choice when using HR, GSR and temperature signals.

Chapter 3

Related work

In recent years many studies have been conducted about the use of physiology as a method for recognizing emotion. The most typical physiological measures are Galvanic Skin Response (GSR), temperature and heart rate (HR). Studies using these measures have tried to recognize emotions such as sadness[LN04] [KBK04], anger[KBK04], boredom[CRBP08], surprise[KBK04], frustration[SFKP02] [RLSV06] and engagement[Ran05] [CRBP08]. Other measures such as Electromyography (EMG)[Ran05] and Electroencephalography (EEG)[Bos06] [CKGP06] [RMP13] have also been used although in my earlier work[Ngu] I established that EEG as the sole measure for emotion did not perform as well GSR, temperature, HR and EMG.
Emotions were elicited by various stimuli. Some studies used visual stimuli[CKGP06] [Bos06], others used a mixture or audio and visual stimuli[KBK04] [LN04] and lastly some studies asked participants to complete mental tasks such as math problems[LN04], anagrams[RLSV06] or play computer games[RLSV06].
For classification a multitude of algorithms have been used. In connection with EEG the algorithm most used is Fischer Discriminant Analysis[Bos06] [RMP13]. For GSR, temperature, HR and EMG the algorithms SVM, KNN, DFA, MBP have been used.
Rani et. al.[RLSV06] has conducted a comparative study where she takes a look at the machine learning algorithms in connection with classification of emotions. The study describes an experiment with 15 participants who are measured with ECG, bio-impedance, EMG, EDA, temperature, Blood Volume Pulse or BVP

and heart sound sensors while either performing an anagram task or playing a modified Pong game. This results in around 15 data sets with 46 features and 100 epochs for each data set. After training a regression tree (RT), Bayes network(BNT), KNN and SVM Rani concludes that the SVM is the best with an average classification rate of up to 85% with a chance level of 20% followed by RT, KNN and BNT at 84%, 75% and 74% respectively with chance levels of 20%. Rani's study suggests that SVM is the most promising classification algorithm for classifying emotions from physiological signals. Rani's study is especially relevant since a computer game was used to elicit emotion similar to the goals of this thesis.

For validating the results there are many differences in the literature. Most studies however do not specifically mention how they validate. Some studies[SFKP02] [KBK04] partition the data into a training set and a test set where each participant is either in the training set or the test set. This ensures that the results reflect the classifier's performance when presented with an unknown person. Other studies[RLSV06] [LN04] mention leave-one-out cross-validation but does not specify what is left out. Ideally a participant should be left out, again to ensure that the result reflect the performance of the classifier on an unknown person.

Dynamic Difficulty Adjustment (DDA) has been implemented in many games although few academic studies exist. One study by Hunicke[HC04] looks at a DDA implementation in a first-person shooter. The DDA implementation used performance indicators to determine whether to adjust the difficulty or not. Participants had to play the game with and without DDA and results showed that the performance of the participants increased and expert players showed a slight preference to the game with DDA. Another study by Liu[LASC09] uses physiological signals for a DDA implementation. The game is a modified Pong game and the experiment included 9 participants. The 9 participants had to play the game with performance-based DDA and affect-based DDA. The results showed that the performance of most participants was better with affect-based DDA. The majority also reported lower anxiety levels and higher enjoyment with affect-based DDA.

The Hunicke and Liu studies suggest that affect-based DDA can improve enjoyment of games.

Chapter 4

Experimental setup

The purpose of this thesis is to investigate how a game can be more fun. Specifically the thesis is about the usage of DDA as a method to keep a player in a flow state which is defined in the theory of flow, written by Csikszentmihalyi. As described in Rules of Play[SZ04] DDA systems are a feedback systems and can be thought of as having four parts according to LeBlanc[LeB99]:

- The game state that is the current state of the game
- The scoring function that measures some part of the game state
- The controller that looks at the scoring function and decides whether or not to apply feedback
- The game mechanical bias that is a game event or set of events that can be turned on or off depending on the controller.

An illustration of this system can be seen in figure 4.1 The game state and the game mechanical bias are both dependent on the game selected. Two different scoring functions and controllers is what this thesis will investigate – one where the scoring function depends on the emotion of the player and one where the scoring function depends on the game state and the player performance.

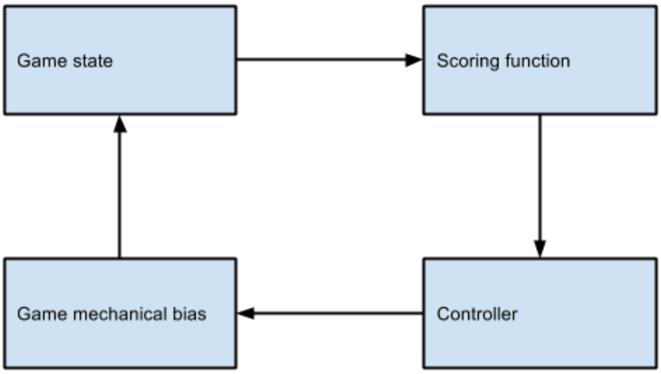


Figure 4.1: Games described as a feedback system. The game state corresponds to the current condition of the game. The scoring function measures some aspect of the game state. The controller looks at the scoring function and decides whether or not to apply feedback. The game mechanical bias is the set of game events that can be turned on or off.

As determined through my previous study and the related work the emotion-based DDA will be based on the physiological signals: EMG, HR and GSR. These signals can be used to train and classify emotions.

In this thesis, two experiments are required. The first experiment is to measure on participants while they are playing Tetris at various difficulty levels. The aim of the first experiment is to develop a model that can predict the emotion of a person. The first experiment also includes a validation study, which can be used to validate the results of the Tetris experiment, where the method of eliciting emotions is by displaying images to the participants instead of them playing Tetris.
The second experiment is to use the model generated from the first experiment and test whether a DDA solution using physiological measurements are better than a DDA solution using the player's in-game performance.

The following sections explain the game selection, DDA design and the technical work.

4.1 Game selection

A suitable game has to chosen that can be used with both emotion-based DDA and performance-based DDA. Some criteria for the game include:

- Simple control scheme – Since the measurements involves sensors on one hand the game has to be easily controlled with one hand.
- Open-source – The game has to be open-source so that the difficulty settings can easily be modified and DDA added later.
- Simple difficulty scaling – The difficulty of the game has to be very easy to change and it has to affect the game directly so side effects are avoided.

These criteria resulted in choosing Tetris for the experiments.
Tetris is a game released in 1984[Joh09] and sold more than 125 million copies[Ols09]. In Tetris the game environment consists of a 10x20 grid where blocks fall from the top to the bottom of this grid. Only one block will fall down at each given time though the blocks are one of seven different shapes. The objective of the game is to survive the constant onslaught of the falling blocks because when the blocks reach the top of the grid the game ends. To do this the player has a few options. The player can rotate the block and move it sideways or down. The player can also do a hard drop which forces the block to instantly drop down. To prevent the blocks reaching the top the player has to clear lines horizontally. Clearing lines is done by filling a line horizontally with blocks i.e. there cannot be any holes in the line. A screenshot of the Tetris game environment can be seen in figure 4.2
The specific Tetris version is written in Python by smartviking[Sma13] and was released in 2013 on the pygame.org website under GPL license.
Apart from the three features described above Tetris has other features such as gameplay that keeps being fun despite the repetitive nature, its use in many studies and its use in an emotion recognition study[CRBP08].

For my purpose Tetris provides a difficulty that can easily be changed simply by varying the speed that the blocks can fall. The standard Tetris controls are moving a block left/right, rotating a block, make a hard drop i.e. instantly drops the block to its would-be location and make a fast drop i.e. drops the block faster than the game speed. Since the last two controls, hard drop and fast drop, in some degree allows the player to control the difficulty these controls are removed.
Another consideration on the difficulty of Tetris is the increasing game speed. In a normal Tetris game the game speed increases when the player clears 10

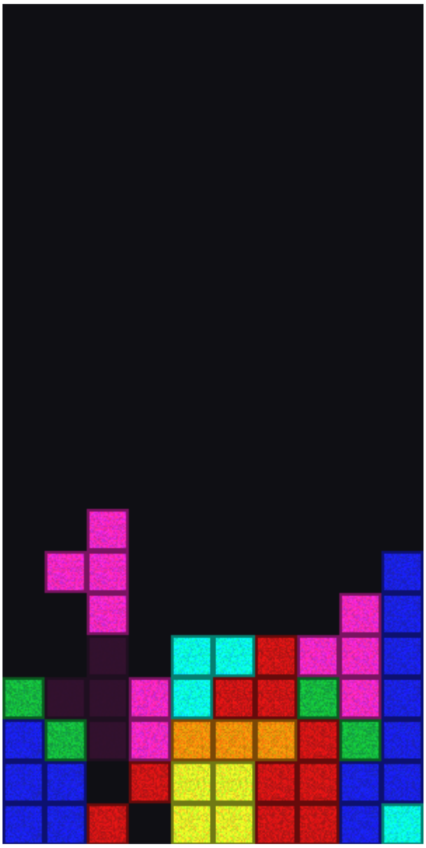


Figure 4.2: The Tetris game environment. Here the dropping pink block will clear two lines as evidenced by the shadow of the block.

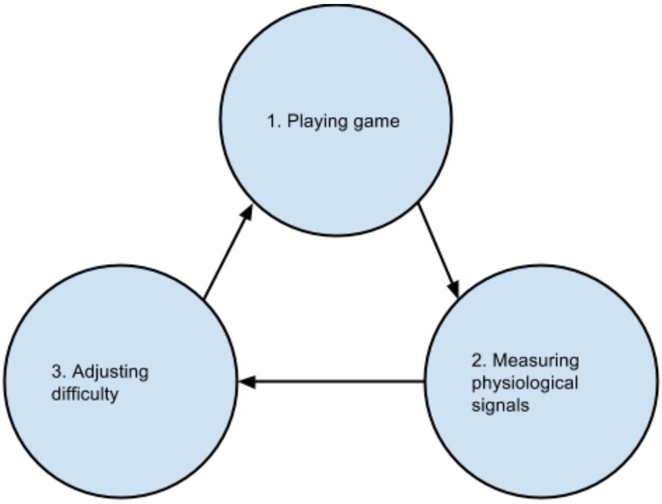


Figure 4.3: The feedback loop based on physiological signals. A player will play the game in 1. The player's physiological signals are measured in 2 and sent to the game computer in 3 where the signals are analyzed and the game difficulty is adjusted accordingly.

lines. This feature are also removed so that the difficulty is only controlled by the DDA.

4.2 DDA design

Two types of feedback will be investigated: one where feedback is based on the player's physiological signals and one where feedback is based on the player's performance in the game.
The first feedback loop, based on the physiological signals of the player, is illustrated in figure 4.3. The feedback loop begins with the player playing the game. While playing the physiological signals of the player is measured and sent to the game computer. The game computer will analyze the signals and adjust the difficulty accordingly.

Due to the results of Experiment 1 the only physiological signal used for this experiment is BVP for measuring the heart rate. The heart rate will be used as an indicator of arousal where the aim is for the game to elicit semi-high levels of arousal. The scheme can be seen in figure 4.4. The goal is to keep the player's heart rate at a certain level which is higher than the baseline heart rate. This

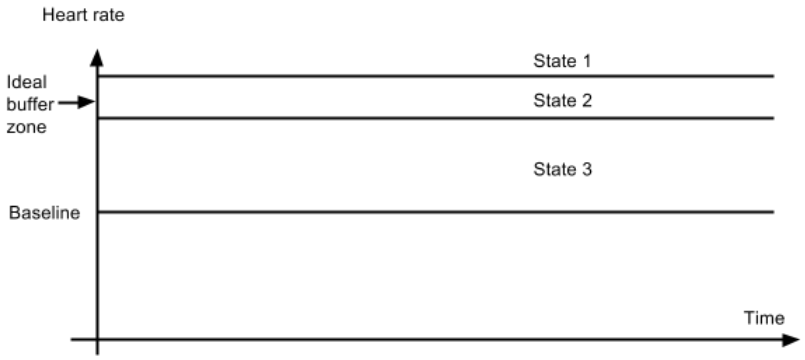


Figure 4.4: A heart rate – time graph. There are three states or three levels of heart rates. State 1 is having a heart rate above the ideal heart rate. State 2 is having a heart rate around the ideal heart rate and state 3 is having a heart rate below the ideal heart rate.

can be achieved by using the adaptive staircase method. The adaptive staircase method is a method often used in psychophysical experiments. The method starts with presenting a stimuli to the participant which the participant can either answer a correct answer or an incorrect answer. The stimuli is then doubled in intensity if an incorrect answer was given or if the correct answer was given the negative stimuli is presented.

In this experiment the adaptive staircase method is used as shown in figure 4.5. Here the player starts in one of the states: 1, above the ideal heart rate, 2, around the ideal heart rate and 3, below the ideal heart rate.

At state 2 the player has the ideal heart rate and the difficulty is maintained. If the player is below the ideal heart rate the difficulty is raised and if the player is above the ideal heart rate the difficulty is lowered.

To relate to the adaptive staircase method, the player would start at the baseline heart rate. A correct answer depends on the state that the player is in and will always raise the difficulty. An incorrect answer can only be given in state 1 where the heart rate is too high. This results in lowering the speed. At state 2 the difficulty is maintained until the player falls into another state where the difficulty is again modified using the adaptive staircase method.

The second feedback loop is based on the performance of the player.

The performance-based DDA is based on two criteria: the overall block height and the change of the overall height of the blocks.

The Tetris board is 10 blocks wide and 20 blocks high. The overall block height is the height of the highest stack of blocks while the change of the overall height of the blocks is when the first variable changes either negatively, which means that the player cleared one or more of the topmost lines, or positively which means that the player has stacked a block on the topmost line. From figure 4.2

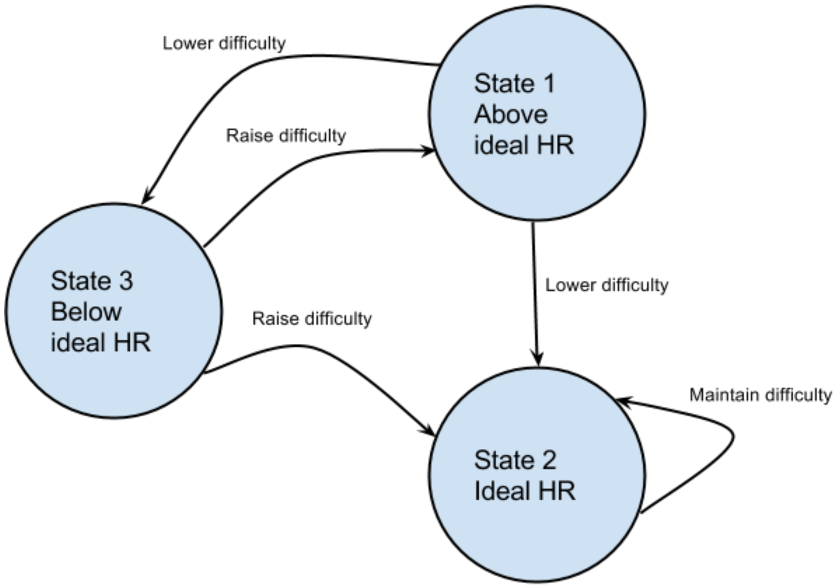


Figure 4.5: HR is heart rate. At state 3 the difficulty is raised which means that the resulting heart rate can either be too high or just right. At state 2 the difficulty is lowered which can effect in the heart rate being too low or just right. If the heart rate is ideal then the difficulty level is maintained.

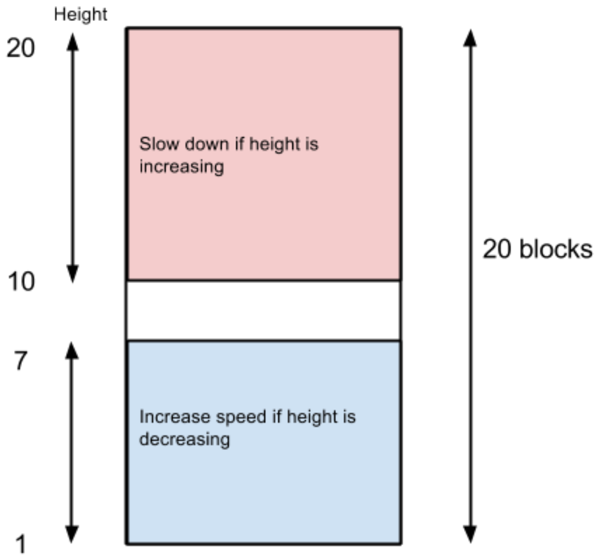

Height
20
10
7
1
Slow down if height is increasing
Increase speed if height is decreasing
20 blocks

Figure 4.6: The different zones of speed changes. The overall playing field in Tetris is 20 blocks tall. If the highest location of a block is below 7 then the speed is increased whenever the height is decreasing. This happens when the player clears a row. If the highest location of a block is above 10 the speed is decreased whenever the height increases so that the player can recover from mistakes.

the overall block height is 7 due to the rightmost column and the change of the overall block is -2 when the pink block falls down since it clears two rows.
The three conditions for a speed change are:

- Decrease speed very slowly if the highest block position is over 10 and the change in the height of the blocks is positive.
- Increase speed when the highest block position is below 7 and the change in the height of blocks is negative.
- Increase speed when the change in the height of blocks is negative and lower than -2.

The zones of speed changes can be seen in figure 4.6.
The first condition is to prevent players from being punished too hard if they make a mistake. Whenever a player increases the height then the game gets slower to allow the player to recover.

The second condition is to increase the speed so that the player is continually challenged. If the height is low and the player keeps clearing rows and keeping the height low then the speed is increased to challenge the player.
The third condition is to increase speed whenever the player clears 3 or 4 rows at a time. The reason is that for a player to clear that many rows at a time is an indicator that they are playing with a strategy of clearing many rows at a time to gain extra points. Therefore the speed is increased to challenge these kinds of players as the rise in block heigh is often planned.

There are however many methods to modify the difficulty. Below is a list and the reason that they are not considered for this project.

- Clearing speed. The difficulty could be based on how high the rate of cleared rows are. This is however not as effective since the removal of hard drops and fast drops limits and controls how high the rate of cleared rows can be.
- Zonal speed. The difficulty can changed according to the highest block position so that there are three zones where the lower zone has the highest speed, the middle zone has an optimal speed and the upper zone has the lowest speed. This could work but only if the optimal speed for each player was found first and therefore not considered since the focus was on a DDA system that would be plug n' play i.e a DDA system that can be used without training it first.
- Block spawn manipulation. This technique is to alter what kinds of blocks that can spawn. By spawning "easy" blocks the game can become significantly easier. Or the game could analyze the game state and spawn blocks that the player has been waiting for. This was not considered since it would require too much effort to design and the difficulty change is not entirely transparent.

4.3 Technical work

The system that I used consisted of the Biosemi ActiveTwo AD box, battery, usb receiver and two computers.
There were three technical challenges: The data analysis, modification of Tetris and programming the validation study.
The data analysis was carried out in Python. Data files from the first experiment generated in the Biosemi acquisition program ActiView was used. These data files were encoded in the .bdf format.

The libraries used in the data analysis program were: *pybdf* for reading the data file, *scipy* and *numpy* for data manipulation and preparation, *matplotlib* for plots and *sklearn* for classifying.

Another technical difficulty was to extract features from the data. Most notably was the extraction of heart rate from the BVP signal. Since no function existed in the libraries I had to program my own simple heart rate function that calculated the heart rate using R-R intervals.

The Tetris game was also written in Python. The modifications for Tetris include: removal of hard drops and fast drops, changing the game so that a continuous experiment cycle is possible, self-report menus so the players can self report their arousal and valence and adding sounds to enhance the immersion and in connection with the emotional responses of in-game events.

For the second experiment DDA has to be implemented for the Tetris game. This required the game to be able to receive data from the Biosemi acquisition program ActiView. In Python this was done by utilizing two threads. One thread was the game thread in which the game ran and controlled the difficulty. The other thread was the DDA thread which both received the game state and the physiological measurements and decided whether to increase, decrease or maintain the game speed. Compared to LeBlanc's model the game thread contains the game state and the game mechanical bias while the DDA thread contains the scoring function and the controller.

The validation study which was performed during the first experiment was an experiment where participants are asked to watch an image from the International Affective Picture System[BL94] or IAPS database. The images are varied from boring pictures of inanimate objects to pictures of gore. The program for this experiment was also programmed in Python using the *pygame* library. The program would go through 60 IAPS pictures and ask the participant to self rate their arousal and valence for each image.

Chapter 5

Method

5.1 Experiment 1

The first experiment is to gather data to build a classifier for the second experiment. The experiment will measure participants while they play a modified version of Tetris while wearing GSR, BVP and EMG sensors.

5.1.1 Equipment

The equipment consists of two computers and GSR, BVP, EMG and EEG sensors from Biosemi. The two computers each have their own role. One is used for data acquisition while the other is the game computer. For this experiment the data acquisition computer receives data from the Biosemi system and receives triggers and event codes from the game computer. The game computer will run the game and send triggers and event codes to the data acquisition computer. The game computer will also record the game state each time a trigger/event is sent to the data acquisition computer. The system is illustrated in figure 5.1

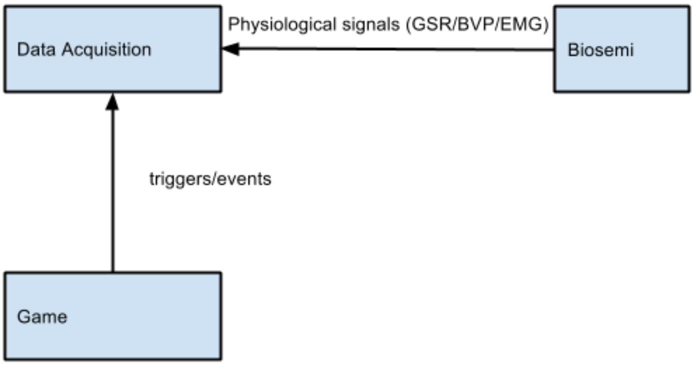


Figure 5.1: The data acquisition computer gets physiological data from the biosemi system and triggers/events from the game computer

5.1.2 Participants

5 participants were chosen for testing. All were in the age range 18-25, all male and all were engineering students.

5.1.3 Experiment protocol

The experiment lasted approximately one and a half hours with small breaks. The participants had sensors put on them and calibrated before being told about the nature of the experiments. The participants also had the SAM explained.
The first part was for the participant to look at 60 IAPS pictures. The picture was preceded by a 3-second relaxation screen to reset their emotional state, then 5 seconds to look at the picture before self-assessing their emotion.
After the first part, the participants had a chance to take a break before starting the second part.
The second part was to play Tetris. The participants had to play a Tetris game where they had to complete a SAM after each 30 seconds. After each 30 seconds the block falling speed increased or decreased so that they were either playing a Tetris game that became increasingly faster or slower. When the valence ratings stopped climbing the game started from the opposite of the speed spectrum and the player had to play the 30-second Tetris games again. When they finished their personal optimal speed was calculated by taking the average of the speeds when the valence rating stopped climbing.
With the personal optimal speed, two new speed ratings were calculated. One speed rating was easy which was 200 ms slower than the optimal speed and

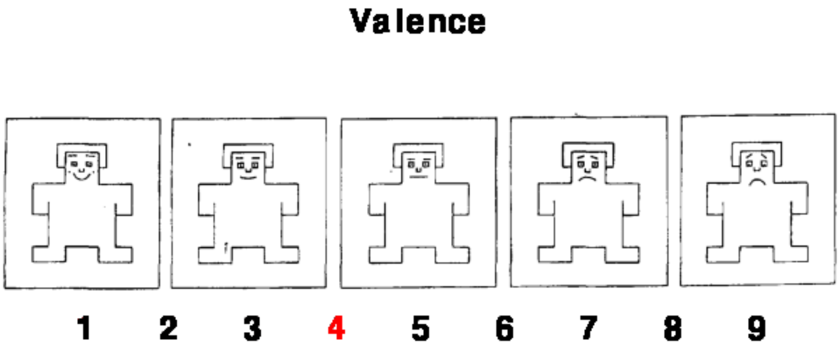


Figure 5.2: The SAM for valence used in both the validation study and the Tetris experiment. The red number is the marker for the selected rating.

one speed rating was hard which was 200 ms faster than the optimal speed. The participants had to play three 5-minute sessions of each difficulty level in randomized order.
The self-assessments were asking the participants to rate their valence and arousal on a 1-9 scale where 1 was high valence i.e. positive or high arousal i.e. excited while 9 was low valence i.e. negative or low arousal i.e. calm. The self-assessment also showed the figures from the original study by Bradley and Lang[BL94]. They can be seen in figures 5.2 and 5.3.

5.1.4 Physiological measures

Based on my earlier work and the literature the physiological measures chosen are GSR, temperature, HR and EMG. Due to availability of equipment GSR, BVP and EMG are measured.
GSR is measured using two passive electrodes. The electrodes are fixed to the middle finger and the ring finger on the participant's left hand. Since the game is only played with one hand it does not interfere with the measurements.
The BVP is measured using a (photo)plethysmograph which is fixed to the index finger on the participant's left hand.
EMG is measured using Biosemi ActiveTwo Sensors. Two sensors, called EMG1 and EMG2 in the rest of the report, are fixed right above the eyebrow and two sensors, called EMG3 and EMG, are fixed on the cheek. The sensors above the eyebrow, EMG1 and EMG2, measure electrical activity from the Corrugator

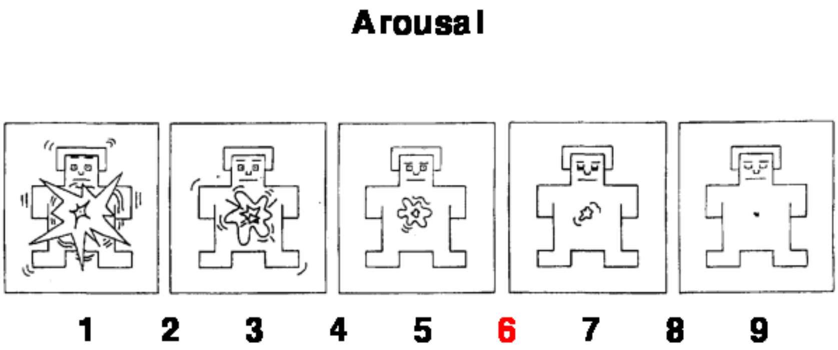


Figure 5.3: The SAM for arousal used in both the validation study and the Tetris experiment. The red number is the marker for the selected rating.

Supercilii muscle while the cheek sensors, EMG3 and EMG4, measure the Zygomaticus Major muscle activity. The placements of EMG sensors can be seen in 5.4.
All sensors are sampled at 128 Hz.

5.1.5 Features

The features are extracted from the six signals. Four EMG signals, one BVP signal and one GSR signal.
Table 5.1 shows all examined features:

The features examined are gathered from the Chanel 2008[CRBP08] and Rani 2006[RLSV06] studies. Some features such as the EMG average, maximum, minimum, standard deviation, slope are standard but the high quartile and low quartile was added since the signal could be noisy and produce extremely high maximum and minimum values. The EMG mean power was added after a power spectral analysis revealed that there was a difference between the mean power of EMG signals and arousal states.
The GSR signal features include the average to describe the average level of skin conductivity. The mean and mean negative derivatives and derivatives are included to describe the changes in skin conductivity where the mean negative derivatives comes from the Chanel 2008 study and is used to describe the average decrease rate during decay time. The range and maximum were included after inspecting some of the GSR signals that showed that range and maximum could

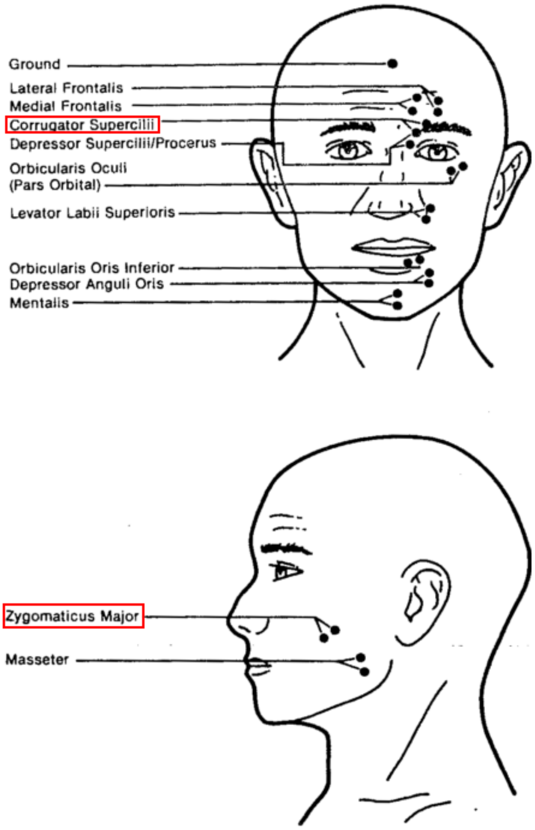


Figure 5.4: Two sensors are placed on the location of the Corrugator Supercilii muscle and two sensors on the location of the Zygomaticus Major muscle. The figure is from Fridlund 1986[FC86]

Signal	Feature	Description
EMG 1-4	Average	Average activity of muscle
	Maximum	Maximum activity of muscle
	Minimum	Minimum activity of muscle
	Standard deviation	The variance of muscle activation
	Slope	Measure of muscle activity change
	High quartile	High quartile of muscle activation
	Low quartile	Low quartile of muscle activation
	Mean power	Mean power of EMG signal
GSR	Average	Average skin conductivity
	Mean derivatives	Measure of skin conductivity change
	Mean negative derivatives	Measure of negative skin conductivity change
	Derivatives	The change of skin conductivity
	Range	Maximum-minimum skin conductivity
	Maximum	Maximum skin conductivity
BVP	Average	Average blood pressure
	Standard deviation	Measure of stability of blood pressure
	Average heart rate	Average heart rate
	Average derivation of heart rate	Measure of heart rate change
	Standard deviation of heart rate	Measure of heart rate stability
	Range of heart rate	Maximum-minimum heart rate

Table 5.1: A table of all 44 examined features

be used to distinguish between arousal states.
For the BVP signal the average blood pressure and the standard deviation of the blood pressure are used to describe the general level of blood pressure and the (in)stability of the blood pressure. The heart rate is calculated from the BVP signal and all heart rate measures except the range of the heart rate are featured in the Chanel 2008 study. The range of the heart rate was added after inspecting the heart rates of some trials that indicated that the range of the heart rate perhaps could be used to differentiate between low and high arousal states.

5.1.6 Analysis

Analyzing the data requires the usage of several steps using different methods in signal processing, data mining and machine learning.
The first step is to separate the data into trials and label these trials. Each trial had a valence and arousal rating and these ratings were split into three categories: low, medium and high. Low valence/arousal was the ratings from 7-9, medium was the ratings from 4-6 and high was the ratings from 1-3. The first experiment generated two sets of data: the validation study and the tetris data.
The next step is to filter the signals and examine the signals.
After filtering the signals the features are extracted and principal component analysis is used to find the most significant features.
Lastly a SVM classifier is trained and the data cross-validated. Since the data is cross-validated with the leave-one-out method or in this case leave-one-participant-out the classification rate is comparable with classifying a person outside the training set.

5.1.7 Signal processing

For both sets of data the signals from EMG, GSR and BVP have to be filtered because of drift. For this purpose a 4th order Butterworth filter was used for applying a 0.1 Hz high-pass filter. Because of Danish power mains running on a frequency of 50 Hz a 49 Hz low-pass 4th order Butterworth filter was applied. This means that everything over 49 Hz was lost but a look at the power spectrum reveals that nothing of significance was lost.
Power spectral density is also useful for analyzing the signals. Power spectral density can reveal the frequencies and their power.
Power spectral density is done for the four EMG signals and the GSR signal where each signal is averaged across trials. Power spectral density is not done

for the BVP signal since an averaged BVP signal could remove the waves that is used to generate heart waves. Instead the power spectral density is applied for averaged heart rate signals generated from the BVP signals.

5.1.8 Feature selection

The total number of features are 44 (8 features each from 4 emg signals/32 features from emg signals, 6 from gsr signal and 6 from bvp signal). Not all of these features are equally good though and therefore Principal Component Analysis or PCA is very useful to show which features are the most important.

5.1.9 Cross-validation

For machine learning SVM with a radial basis kernel is used for classification since it is a popular choice and because it outperforms other machine learning techniques which is seen in Rani's study[RLSV06]. A classifier is trained for both the validation study and the Tetris experiment. Cross-validation is then done with leave-one-out where one label is used i.e. one person's data as the test set and the rest as training set. The reason for leaving a label out is to measure how well the classifier will perform on people not used for training.

5.2 Experiment 2

The second experiment is to test the difference between Tetris with emotion-based DDA or performance-based DDA. Because of the poor results of the first experiment (see section 6.1.2), the emotion-based DDA used for this experiment will be based on heart rate where the aim is to keep the heart rate of a participant higher than the baseline heart rate. This is due to finding a slight connection between heart rate and arousal. Using this the emotion-based DDA can be optimized so that the player will always have a heart rate higher than the baseline and therefore an arousal higher than normal.

5.2.1 Equipment

The equipment used is the same as in Experiment 1 except that an ethernet crossover cable was added between the acquisition computer and the game com-

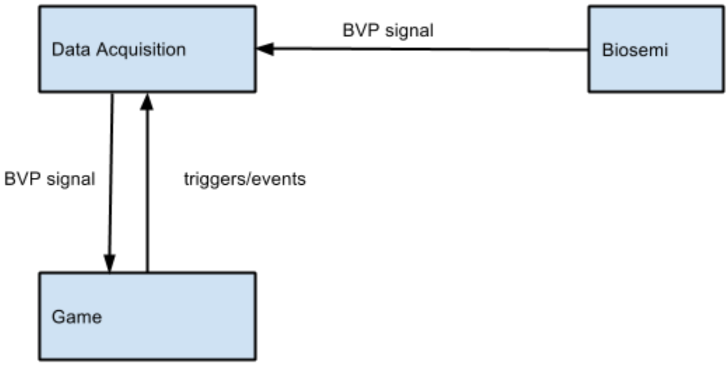


Figure 5.5: The second experiment has the data acquisition computer send the BVP signal on to the game computer.

puter. This was added so that the measurements of the biosemi system can be picked up in realtime and used for the feedback loop. The setup can be seen in figure 5.5.

5.2.2 Participants

9 participants were chosen for this experiment. All the participants were male engineering students in the age group 18-25.

5.2.3 Experiment protocol

The participants were asked to play 6 5-minutes sessions of Tetris. Half of the sessions was Tetris with emotion-based DDA while the other half was with performance-based DDA. The sessions were randomized so that the participants did not know which DDA solution they were playing.
After each session each participant was asked to self-rate their valence and arousal and the level of challenge they felt. Valence and arousal were the same as from last time but the challenge was a scale from 1-9 where 1 was too difficult, 5 was the sufficiently challenged and 9 was too easy. The number of deaths were also recorded as a measure of performance.

5.2.4 Analysis

With these self-ratings the results can be analyzed to show which type of DDA will lead greater enjoyment. The valence ratings are useful in that the higher the rating is, the more positive the DDA is perceived to be by the participants. The same however is not the case with arousal where higher arousal is not always preferable. This has been shown in Chanel 2008[CRBP08] where the highest valence ratings does not correlate with the highest arousal ratings. The challenge rating is used to asses how good each DDA solution is at balancing the game. The number of deaths can be used to asses which DDA solution will cause the participant to perform better.

For analyzing the results a paired t-test is used to see whether there are statistically significant difference between each DDA solution. A paired t-test is suitable because of the low number of participants in this experiment.

Chapter 6

Results

6.1 Experiment 1

The two data sets generated were for the validation study and Tetris experiment.

The validation study has data from 5 participants where each participant had 60 sessions where each session was 5 seconds long. This resulted in 300 trials for the validation study.

The Tetris experiment has data from 5 participants although due to an error for one participant, data from 4 participants are used. The participants were to play 5 minutes of Tetris 9 times. This leaves 36 trials.

The power spectral density was applied for each participant on the averaged signal over each arousal/valence class on the Tetris data set. E.g. for a participant who has 3 high arousal trials the signal are averaged into a new averaged signal and then the power spectral density was found for each of the arousal classes. This was only done on EMG signals and the GSR signal. Here an interesting property was found. As seen in figure 6.1 one class has lower power than the rest. This is consistent across the other three participants. The signal is the EMG4 signal which is the signal from the sensor on the cheek. The three classes are high arousal, medium arousal and low arousal where the low arousal class is the one that shows lower power than the others.

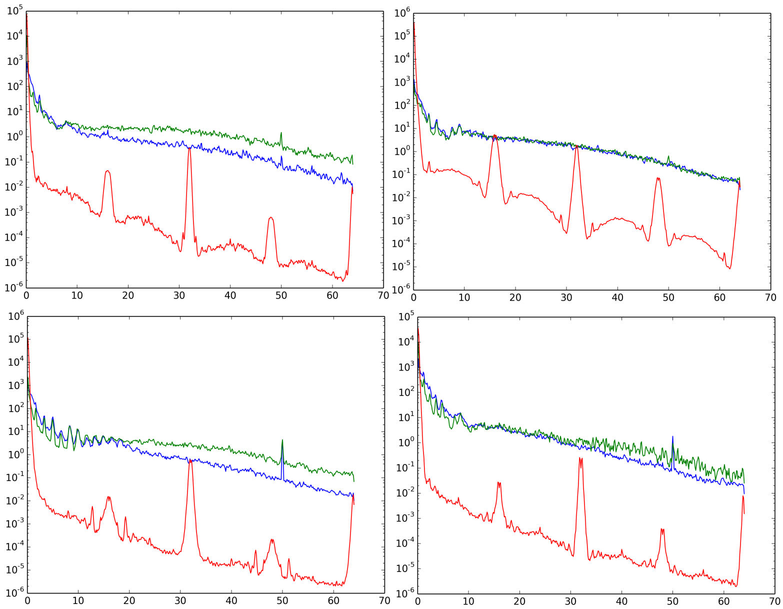


Figure 6.1: The power spectrum for four participants with three averaged signals – low, medium and high arousal. Blue is high arousal, green is medium arousal and red is low arousal. The three averaged signals are calculated from 9 Tetris trials. The three peaks that appear across all power spectra seem to be an artefact from the measuring equipment since they all appear at 16, 32 and 48 Hz.

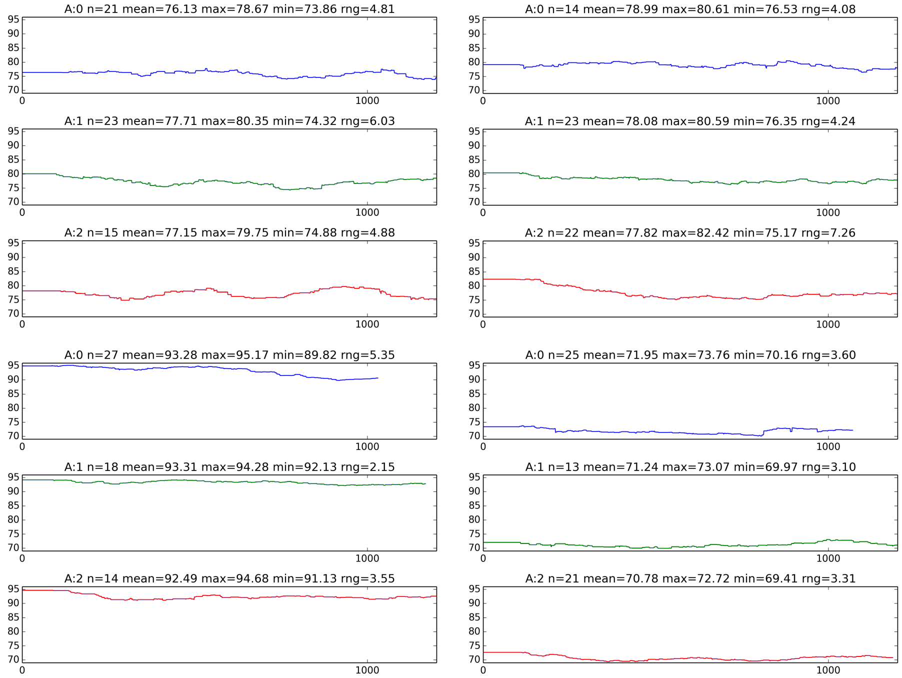


Figure 6.2: Four plots where each plot belong to a participant in the validation experiment. There are three graphs for each participant where the top graph is the averaged heart rate signal of all of the high arousal trials of that participant, middle graph is the averaged heart rate signal of the medium arousal trials and bottom graph is the averaged heart rate signal of the low arousal trials.

The signals otherwise do not seem to differ much. Using averaged signals there might however be differences between the signals.

Figure 6.2 shows the averaged heart rate signal over trials rated as having high arousal, trials rated as having medium arousal and trials rated as having low arousal for four participants. From the figure only the top left plot shows that in high arousal trials the participant had a lower heart rate than medium and low arousal trials. For the rest of the participants high arousal trials resulted in a higher heart rate than low arousal trials.

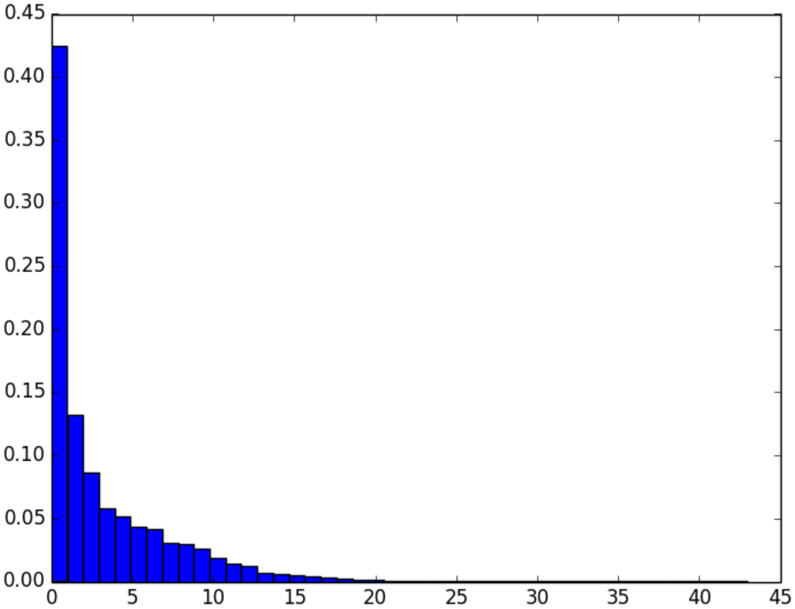


Figure 6.3: The most contributing component accounts for about 42% of the variance followed by 14% and 9% for the next two components.

6.1.1 Feature selection

PCA was run on both data sets.
For the validation study the result was that component that contributed most of the variance, contributed 49% of the variance. The next two components contributed 11% and 9% of the variance. All the components can be seen in figure 6.3.

The first component i.e. the component that describes most of the variance accounts for 42% of the variance. The most contributing features for this component are the maximum, high quartile, average, low quartile and minimum of the EMG3 signal. This might indicate that features belonging to smiling account for a lot of the variability.
The second component, which accounts for 14% of the variance, has the maximum, high quartile, average, low quartile and minimum of the EMG1 signal as the most contributing features. This is quite interesting how the second compo-

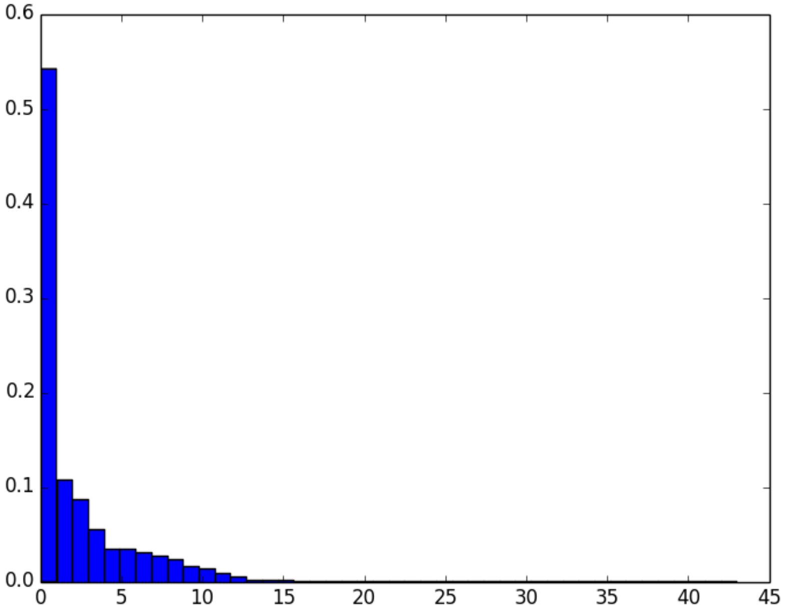


Figure 6.4: The most contributing component accounts for about 54% of the variance and the next two components account for 11% and 9% of the variance.

nent is dominated by the activation of the Corrugator Supercilii muscle. This muscle is under the eyebrow and is used when you frown. As of such the second component could describe frowning but it could also be blinking because frowning would also activated the EMG2 sensor whose features are not present as the most contributing features.
The third component, which accounts for 9% of the variance, has the slope and mean power of the EMG1 and EMG2 signals and slope of EMG3 and EMG4 signals as the most contributing features. This component could describe the changes in facial expressions rather than the actual facial expression.

The result of PCA on the Tetris experiment data is similar to the PCA of the validation study. The most contributing component accounts for 54% of the variance and the next two components account for 11% and 9% of the variance. All the components can be seen in figure 6.4.

The Tetris components have almost the same contributing features as the vali-

dation study. The first component which accounts for 54% of the variance, has the maximum, high quartile, average, low quartile and minimum of the EMG2 signal as significant features. This component could be very dependent on the action of frowning.
The second component, which accounts for 11% of the variance, has the slope of EMG3, EMG4, EMG2 and EMG1, mean derivation of GSR i.e. the average slope, the range of GSR and mean power of EMG4 as significant features. These features can describe the differences from one state to another, particularly the slopes of the EMG signals which of course is highest when the muscles are being activated or deactivated.
For the third component, which accounts for 9% of the variance, has the minimum, low quartile, average, high quartile and maximum of EMG1 signal as significant features. This is similar to the second component of the validation study and could indicate that the act of blinking shows up as variance in the data.

For both experiments the samples were plotted with component 1 and 2 along the axes. The plots can be seen in figure 6.5 and 6.6. Here the real problem shows clearly. The scatter plots are colored after participants and therefore it is clear that the data is separated due to differences between participants and not differences between arousal/valence levels.

The PCA scatter plot shows that the variance in the data is more dependent on participants than arousal/valence levels. In effect the classification will probably not work as well.
The two PCA scatter plots in figure 6.5 and 6.6 show how the variance in participants influence the components. In figure 6.7 the data from the 5 participants from the validation study has been used to create scatter plots with component 1 and 2 along the x and y axis respectively. The markers are colored after the arousal level where red is low arousal, green is medium arousal and blue is high arousal. This is to ensure the plots do not show the variance between the participants but rather the variance between arousal or valence states. Figure 6.7 does not show any differences between any arousal states for any participants. The same is done for data from the Tetris experiment which is data from 4 participants. The scatter plots can be seen in figure 6.8. Here the results mimics the results from the validation study in that there are no specific variance between the arousal states of any participants.
Because the results are not convincing the two features examined previously, mean power of EMG4 and average of heart rate, are extracted and used for PCA. Another feature - range of the GSR - is chosen as well to represent the stability of the skin response. The PCA for these three features show that for all 5 participants of the validation study, the first component account for 70%-91% of the variance and all have the GSR range as the most significant

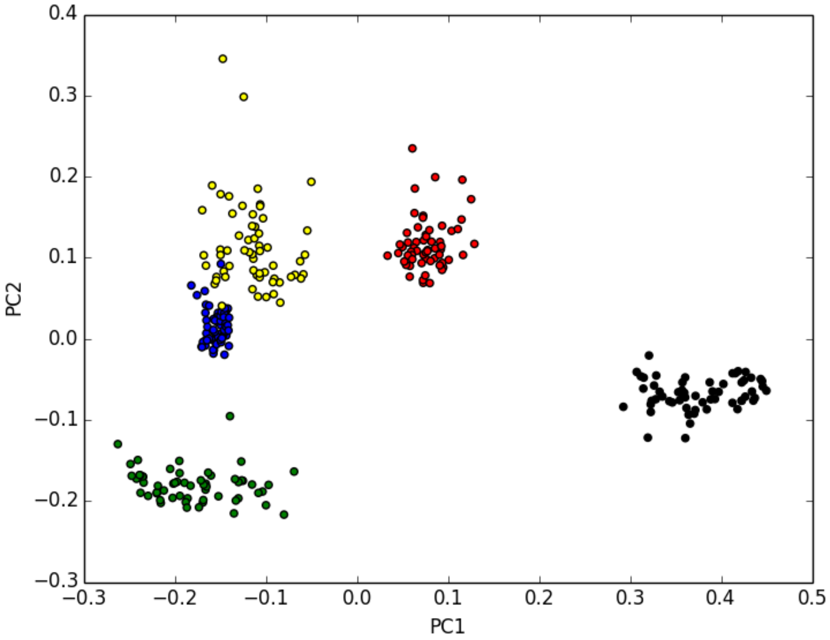


Figure 6.5: Scatter plot for validation study with principal component 1 and 2 on the axes. Each color represents a participant and it is clear that the data is separated according to participants.

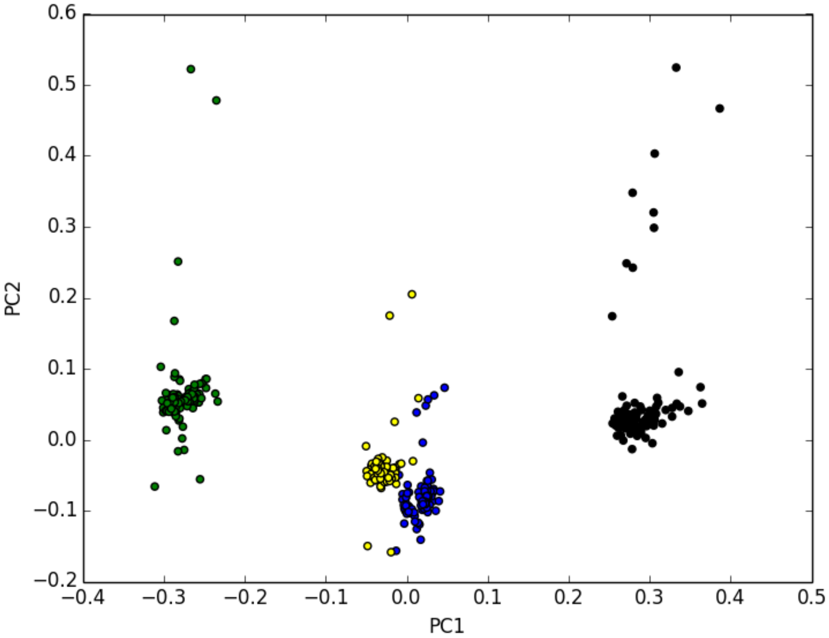


Figure 6.6: Scatter plot for the Tetris experiment with principal component 1 and 2 on the axes. Each color represents a participant and it is clear that the data is separated according to participants.

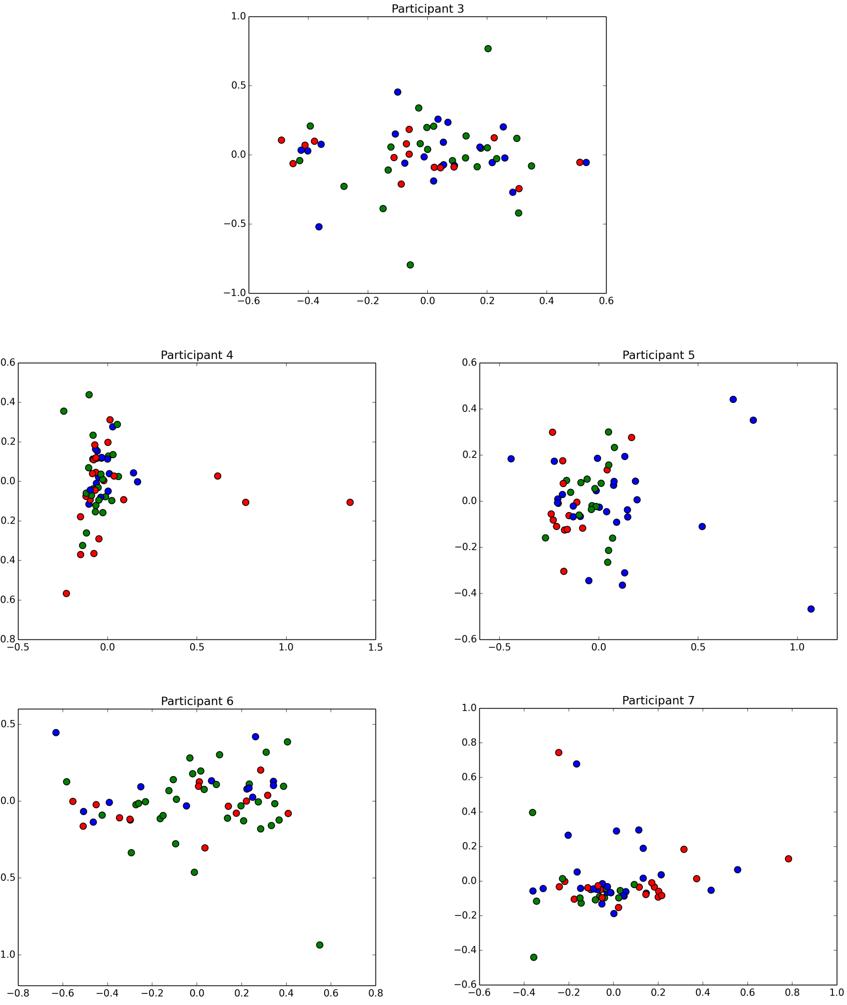


Figure 6.7: Scatter plots for all participants in the validation study with component 1 and 2 along the axes. The markers are colored after the level of arousal where red markers are low arousal, green is for medium arousal and blue is for high arousal.

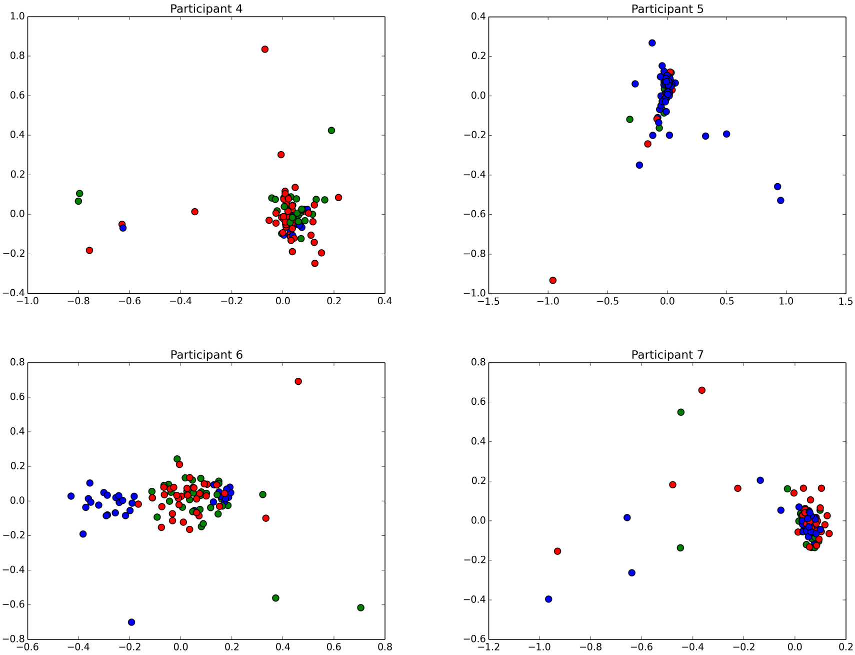


Figure 6.8: Scatter plots for all participants in the Tetris experiment with component 1 and 2 along the axes. The markers are colored after the level of arousal where red markers are low arousal, green is for medium arousal and blue is for high arousal.

Description	Classification rate	Chance level	Percentage points improvement on chance level
Validation, valence	46%	33%	13
Validation, arousal	23%	33%	-10
Tetris, valence	33%	33%	0
Tetris, arousal	22%	33%	-11

Table 6.1: A table over the cross-validation results. The two first classification rates uses data from the validation study which is 5 participants with 60 samples each. The next two classification rates uses data from the Tetris experiment which is 4 participants with 99 samples each. The cross-validation is leave-one-participant-out.

feature. The two other components which are the same for all the participants are mean power of EMG4 and average heart rate respectively. The scatter plots can be seen in figure 6.9. In the figure red markers are low arousal, green is medium arousal and blue is high arousal. The figure shows that there are no clear differences between component 1 and component 2.

6.1.2 Cross-validation

The results of the cross-validation using SVM classifier with all 44 features can be seen in table 6.1. Two cross-validations are made for the validation study and the Tetris experiments - one for valence and one for arousal. The validation study uses 5 participants with 60 samples each and the Tetris experiment uses 4 participants with 99 samples each. The cross-validation is leave-one-participant-out. From the table it is clear that the results is not very good. The valence ratings are better than arousal with 46% and 33% against the arousal classification rates of 23% and 22% for validation and Tetris respectively.
For the validation study the classification rate is lower than chance level.

6.1.3 Conclusion

Looking at the signals there seems to be some connection between the power of the EMG signal on the Zygomaticus major muscle and arousal. Similarly there seems to be a slight connection between the heart rate and arousal where a higher heart rate generally indicated that the participant was more aroused.
Moving on to the PCA showed that most of the variance was between participants and not arousal/valence states. The study of the components showed

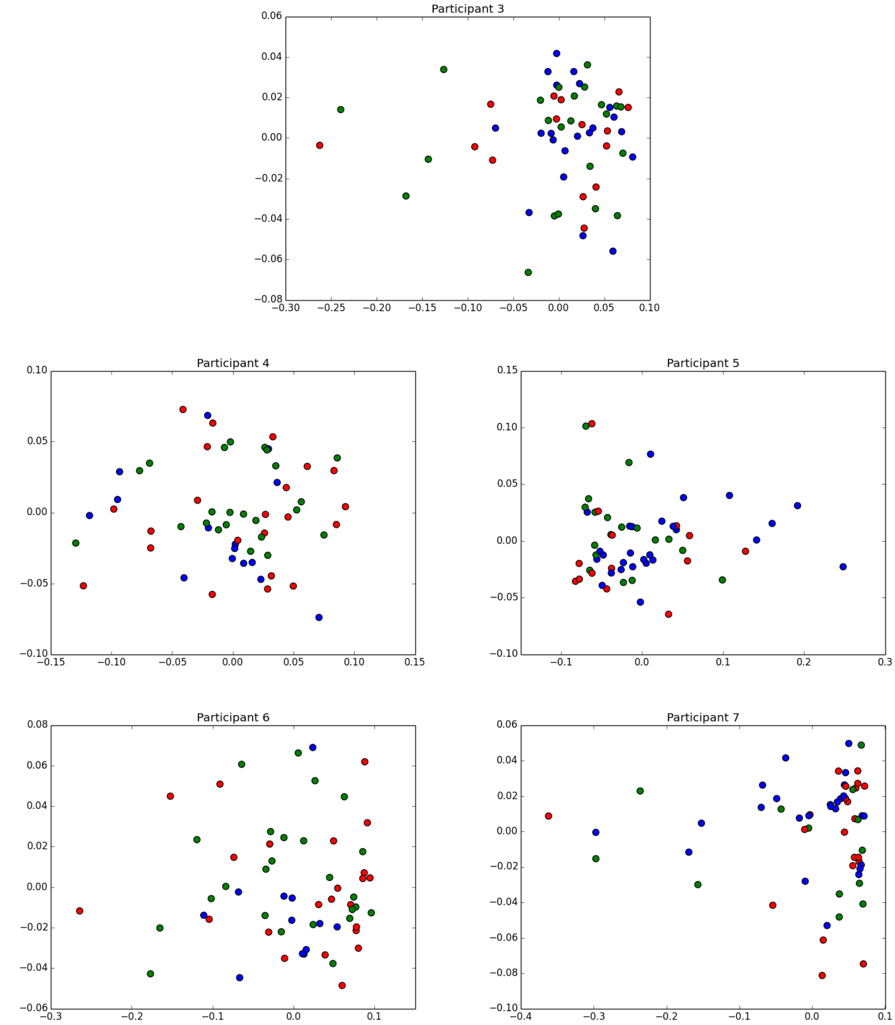


Figure 6.9: Scatter plots for all participants in the validation study with three with component 1 and 2 along the axes. Component 1 is primarily composed of the range of GSR feature while component 2 is primarily composed of the mean power of EMG4 feature. The markers are colored after the level of arousal where red markers are low arousal, green is for medium arousal and blue is for high arousal.

that features of the EMG1 signal was quite prominent. Other important features were the slope of the EMG3 and EMG4 signals that were important both in the validation study and Tetris experiment. The average value of the BVP signal also featured but oddly enough neither heart rate features or most of the GSR features were deemed important.
The results of the cross-validation is not ideal since the classification rates are too low to be of any use. For arousal the classification rates are actually much worse than chance level with classification rate of 23% and 22% for the validation study and Tetris experiment respectively. This means that since chance level was 33% then the validation study was 10 percentage points worse than chance level and the Tetris experiment was 11 percentage points worse. For valence the Tetris experiment was equally good as chance level at 33% while the validation study showed results of 13 percentage points better than chance level at 33% with a classification rate of 46%.
The cross-validation however confirms the findings from the signal processing and PCA. From the signal processing no significant difference was found except an increase of power on an EMG signal between low arousal and medium/high arousal. There was also a slight connection between arousal and heart rate although it was very minor. The PCA also showed that the variance in the data was between participants and not between arousal/valence states. The cross-validation results very much reflects this.
After reducing the data set to three features - mean power of EMG4, range of GSR and average heart rate, PCA was run again on the data. The PCA did not show any variance between these three features, even if they looked promising when examining the raw signals, the difference in the raw signals did not carry over to the PCA. As such the emotion-based DDA for experiment 2 was redesigned to use heart rate since some studies[LASC09] [CRBP08] show a link between arousal and heart rate and using this could be enough for a DDA solution.

6.2 Experiment 2

The participants had to self-assess their valence, arousal and challenge. These three ratings and the number of times they died were written into a table.
The average of each participant's heart rate based DDA trials and performance-based DDA trials were calculated and then compared.
The comparison of valence which can be seen in figure 6.10. Here participant 3 and 9 rate the performance-based DDA higher but for both of them, no statistical significance was found with p=0.057 for participant 3 and p=0.199 for participant 9. A paired t-test was also conducted and the results are a t-value of 1.305 and a p-value of 0.228. This suggests that there is no difference between

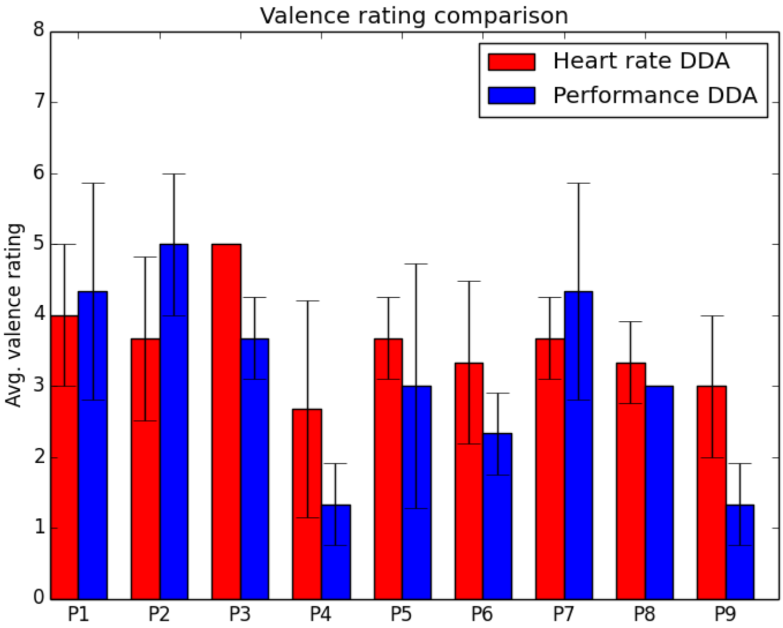


Figure 6.10: Comparison of average valence ratings between the heart rate based DDA and performance-based DDA. Lower values indicate higher valence. Participant 3 and 9 seemed to generally prefer the performance-based DDA but a t-test reveals that there are no statistical significance with p=0.057 and p=0.199 for participant 3 and 9 respectively.

the two DDA solutions with ($p \leq 0.05$).

The comparison of arousal can be seen in figure 6.11. From the figure participant 4 rates the heart rate based DDA higher in arousal but participant 8 and 9 rate the performance-based higher in arousal. There are however not statistical significance with p=0.184, p=0.057 and p=0.184 for participant 4, 8 and 9 respectively. The results from the paired t-test reveal that with the t-value at 0.392 and p-value 0.705 there no statistical significance for any difference between the two DDA solutions ($p \leq 0.05$).

The challenge ratings can be seen in figure 6.12. Here participants 5 and 9

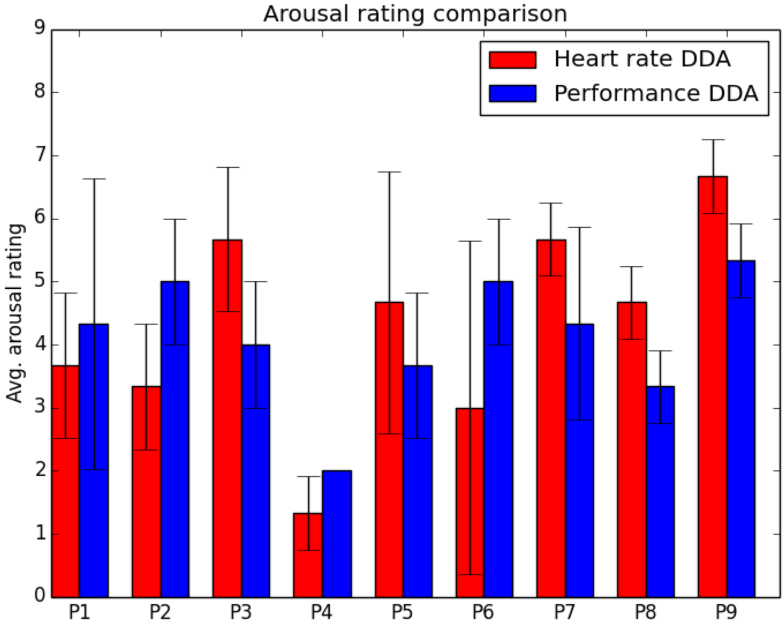


Figure 6.11: Comparison of arousal ratings between heart rate based DDA and performance-based DDA. Lower values indicate higher arousal. Participant 4 rates the heart rate based DDA higher while participant 8 and 9 rate the performance-based DDA higher but as in the valence ratings there are no statistical significance. For participant 4 a t-test shows p=0.184 and for participants 8 and 9, p=0.057 and p=0.184.

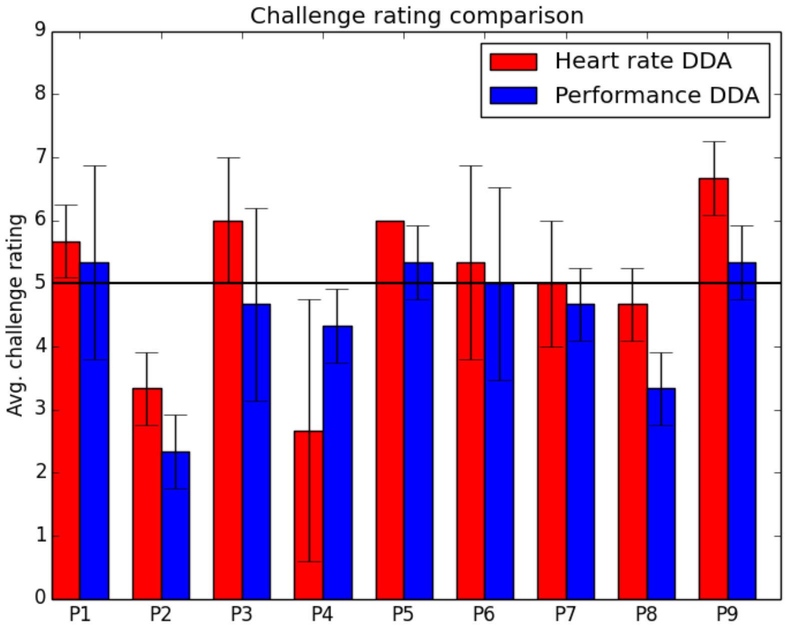


Figure 6.12: Comparison of challenge ratings where values closer to 5 are better. Participants 5 and 9 seem to prefer the performance-based DDA while participant 8 preferred heart rate based DDA. There are however no statistical significance for the difference for participants 5, 8 and 9 with p=0.184, p=0.225 and p=0.057 respectively.

seemed to prefer the performance-based DDA while participant 8 seemed to have preferred heart rate based DDA although no statistical significance was found for any of them with p=0.184, p=0.225 and p=0.057 for participants 5, 8 and 9 respectively. The results of the paired t-test is t=1.776 with p=0.114. There is no difference between the challenge ratings ($p \leq 0.05$).

As a measure of the participants' performance the number of times they died were also counted. The comparison can be seen in figure 6.13. The performance of the participants are really similar. The paired t-test supports that with t=-0.604 and p=0.562. The results of the performance also show no difference ($p \leq 0.05$).

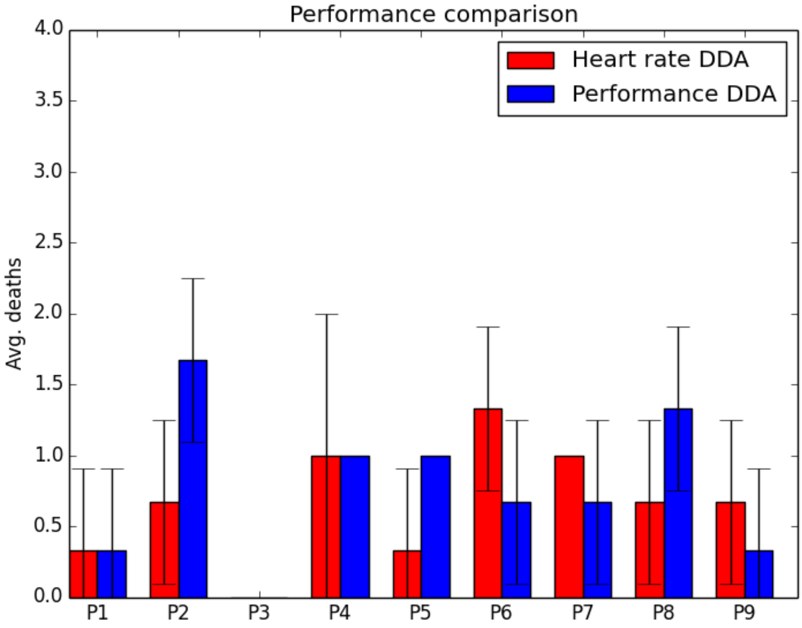


Figure 6.13: Comparison of the performance measured in average deaths across the each type of trial where lower is better. There is no clear preference with any participant and for participant 1, 3 and 4 the number of deaths in both DDA solutions are exactly the same.

6.2.1 Conclusion

The results of the second experiment show that there is statistically no difference between the two DDA solutions but a slight preference towards performance-based DDA which participant 3 and 9 rated more positive, or higher valence, in comparison to heart rate based DDA. Arousal was also in slight favor of performance-based DDA with participants 8 and 9 rating the performance-based DDA higher. It is however important to note that a DDA solution do not have to be designed with maximum arousal in mind as the level of enjoyment do not correlate linearly with the level of arousal as seen in Chanel 2008[CRBP08].
The perceived level of challenge however is important as that will affect the enjoyment that the participants will derive from the game. In the second experiment participants 5 and 9 felt the performance-based DDA suited them better. It is interesting that the performance level of the participants, measured in the number of times they died, does not always show the same pattern as challenge or valence. An example is participant 3 who did not die in any of the trials but still rated the performance-based DDA higher in valence.
None of the four measures did show any statistical significance however and no participants showed any clear preference so it is hard to conclude that there is any real difference between these two DDA solutions. Subjectively though some participants did seem to prefer the performance-based DDA and there might be reasons for that. One is that the heart rate based DDA works a bit slower in which it only changes speed every 10 seconds, compared to the performance-based DDA which can change speed every second if needed. This difference ensures that the performance-based DDA can often save a participant who is about to die by quickly slowing the game. This can affect the valence score in favor of performance-based DDA. Another reason could be that the heart rate changes in Tetris do not change that much. From the first experiment the difference in heart rate is around 3 bpm from low arousal to high arousal for some participants. This makes it hard for the heart rate based DDA algorithm to keep changing the game speed to challenge the player. Furthermore the heart rate algorithm is not consistently precise i.e. there is a large variation on heart rate readings, and therefore this affects the heart rate based DDA negatively which can explain why the performance-based DDA is preferred by some participants.
The silver lining in this is, however, that several participants had the opinion that any kind of DDA was more fun than the standard Tetris difficulty which increases speed until the player dies. They specifically noted that the start of the game can be quite boring and they would prefer that the speed was increased in the beginning and later when the block level has increased they would like to have a chance to turn the game around.
One participant also mentioned that doing DDA for Tetris did not seem like a good idea as the game gets boring after a while even with DDA. While it might have been a personal preference there is a valid point that Tetris is too small

a game to fully utilize DDA as the core gameplay mechanic did bore some of the participants in the end. DDA in Tetris also eliminates any sense of progress as one metric of progress in Tetris is the highscore. When DDA is active the highscore cannot be compared to others' highscores because the a high highscore does not necessarily mean that the player is good. And beating a highscore does not necessarily mean that the skill of the player has increased since the DDA corrects for this. Perhaps other games or games genres would be more suitable for DDA.

Chapter 7

Conclusion

The project was about trying to improve game experience. A model of game experience can be found by looking at the theory of flow by Csikszentmihalyi. Flow is a mental state where a person loses track of time and is completely absorbed by a task which Chen[Che07] notes is very similar to what game designers try and design their games for. Csikzenmihalyi has mapped flow in a map with skill level on one axis and difficulty of a task on the other where flow is when a person's skill level correctly corresponds to the difficulty of the task. If the task is too difficult compared to the skill level of the person then the person would feel anxiety or if the task is too easy then the person would feel boredom. Using the theory of flow a game could provide a better experience for a player if the difficulty were adjusted to suit the player's skill level. This technique is called dynamic difficulty adjustment or DDA.

DDA algorithms has to take an input to know how to change the difficulty. The input used most in commercial games is the performance of the player. I propose the use of emotions of a player to adjust the difficulty. This thesis looks at two types of DDA, emotion-based and performance based, and tries to determine which type is the best.

The first type of DDA is one that uses emotion to determine the difficulty level. For designing this a literature study was carried out. From this study it was determined that using the dimensional theory of emotions was suitable since the two dimensions arousal and valence were easy to detect. Following this an experiment was set up to build a classifier that could classify emotions. The experiment had participants play Tetris in various difficulties and look at pictures where they rated their own arousal and valence. This happened while they were wearing a photoplethysmograph for measuring the blood volume, galvanic skin response sensors for measuring the skin conductivity and electromyography sensors for measuring muscle activation in facial muscles. These sensors were chosen on the basis of the previous literature study which showed good classification results with these particular sensors. Another alternative was considered with electroencephalography but the literature study showed that the results were not as good as with the other sensors.

The results of the first experiment did not produce any usable classifier. With results ranging from 22% to 46% and a chance level of 33% for classifying on 3 classes the classifier was not good enough to use. Some encouraging sign showed that there was a slight correlation between the heart rate and arousal. I decided to design the emotion-based DDA to only use heart rate.

The design of the emotion-based DDA only used heart rate. The aim was to elevate the player's heart rate above the baseline heart rate which is the player's resting heart rate. By adjusting difficulty for that aim and using the adaptive staircase method this was achieved. The design of the performance-based DDA uses the game state to try and keep the Tetris interesting by increasing speed when the player was far from dying and decreasing speed when the player was close to dying.

Having the two DDA solutions a second experiment was conducted where each participant played both versions of DDA and self rating their arousal, valence and the challenge of the game.

Analyzing the results of the second experiment showed a minor preference for the performance-based both in valence and challenge where the performance-based was rated more positive and having the right level of challenge. However no statistical difference was found. One reason for the results of the emotion-based DDA could be the fact that the emotion-based DDA only changed speed every 10 seconds compared to the variable speed changing of the performance-based DDA.

Another point is the selection of Tetris for DDA. Some participants got bored of playing Tetris after a while no matter which DDA version was used.

As of such the emotion-based DDA cannot yet outperform performance-based DDA because the emotion recognition is too imprecise. Perhaps when emotion recognition becomes more precise, an emotion-based DDA can guide players into their own personal flow zones.

Chapter 8

Future work

For future studies it might be interesting to use DDA on other games and games genres. Even though there was a preference for the performance-based DDA then the emotion-based DDA was not far behind. With an emotion classifier that is more reliable than the one in this thesis it might be possible to bridge the gap found between emotion-based DDA and performance-based DDA.

Another interesting possibility is to make DDA systems that takes both emotions and performance in account. Such systems have even greater potential to bring players into the flow state by both recognizing the emotions of the player but also taking the game state into account.

For future studies it could be interesting to study performance-based DDA, emotion-based DDA and behavior-based DDA and whether some difficulty measures are better suited for one kind of DDA than another. Behavior-based DDA could take point in the Scheirer 2002 study[SFKP02] where she studied how behavoir (key-press patterns) could be an indicator for emotion. Based on this a kind of behavior-based DDA could be created using key-press patterns as input and changing difficulty to achieve the right kind of behavior. In Tetris this could manifest itself as number of rotations where rotating too much or fast could indicate too much stress.

Another kind of DDA could be inspired by Jesper Juul's notion of time[Juu10]. By looking at time and lost time in connection with failures the DDA could adjust the difficulty based on how much time it could cost the player if the player were to fail. This is however a very advanced and requires a lot of information about the player, type of game and the context of which the player is in.

Future studies about emotion recognition should perhaps focus on events instead of long stretches of emotions. The data in the thesis clearly showed very little difference from low arousal to high arousal which and this could be due to the participants having to maintain the same emotion in 5 minutes and failing at that. This is evidenced from the classification rate of 33% versus 46% for a chance level of 33% for Tetris and the validation study respectively. Here the classifier from the validation study is better than the one from the Tetris study which could perhaps be from participants experiencing many emotions during the 5 minutes. This leads to sketchy self-ratings where the final self-rating could be averaged over the entire 5 minute session. By switching the focus to good events and bad events the data might produce a better classifier that can consistently recognize emotion. This direction could be taken further where a DDA system can try and emulate patterns of good events and bad events instead of trying to aim for the player to experience one constant state of emotion.

In commercial settings the difference in cross-validation with leave-one-participant-out versus leave-one-sample-out is important. In the first case the system would be plug n' play and directly usable without any training which is desirable in a commercial product. The second case would necessitate training the system before it is ready to use which can limit commercial adoption. However both methods could be fused into an adaptive system that is already trained but would continually adapt itself to the player when it is used.

With the current level of sensors in current technology it is possible to make very complex DDA systems and with the increasing number of sensors in an all-connected future it seems that DDA systems will only be utilized more to provide a completely personal experience in games.

Appendix A

Feasability study

This report was written in the period from February 2013 to May 2013 by me. The report was written for a special course supervised by Tobias Andersen at DTU Compute and was nominated at 5 ECTS points. The full title of the report is: The use of physiological measurements to assess user involvement in computer games. The full report follows beneath.

The use of physiological measurements to assess user involvement in computer games

Du Nguyen, s082968

Kongens Lyngby 2013
DTU COMPUTE-2013

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253031
compute@compute.dtu.dk
http://www.compute.dtu.dk/ DTU COMPUTE-2013

Contents

Chapter 1

Introduction

The computer game industry is a massive industry. According to a report [ESA12] by the ESA (Entertainment Software Association) the total amount of money spent on computer and video games in 2011 was $24.75 billion in the US while it was $67 billion [Gau12] globally in 2012. It is clear that computer game industry is no longer a niche market but a large global market that will only continue to grow.

What is the secret behind the meteoric rise of the computer game industry? Often when speaking of entertainment and the lure of entertainment the term escapism is central. Merriam-Webster defines escapism as "habitual diversion of the mind to purely imaginative activity or entertainment as an escape from reality or routine" [MW13]. It can be argued that computer games provide a higher degree of escapism than movies or books because of their inherent nature of interactivity.

The quality of computer games are more or less defined of how fun they are to play. If a game is fun to play and provides that sense of escapism it often tends to score higher in reviews and sell more. The challenge in game development is how to create an experience that is fun and especially an experience that is fun for everyone.

The report proposes a system that monitors a player's physiological and mental state which can be used by games to dynamically adapt to the player and improve the experience for the player. This system can be used both in the development phase of the game and the "living room" phase. There are appealing reasons to use such a system for both the development phase and living room phase. For development this system can be used for play testing the game since this system returns objective metrics of the mental state of a player. For the living room phase the system can personalize the game experience and provide a smoother and individual game experience. As the living room system is more complicated it requires more thought in how such a system is designed. The rest of the report will deal with the issues of constructing a living room system such as: How can emotions be conceptualized? How can you measure fun and how can fun be created? How do you measure emotions and how do emotions relate to the feeling of fun?

Chapter 2

Gaming Theory

The ultimate goal of the current project is to increase game experience for computer game players. To achieve this, a theory of fun and game experience has to be defined and how these relate to emotions and a player's mental state. Knowing how game experience relate to emotions and how game experience is created will make it possible for the current project to alter the game experience for the player.

In an article Chen [Che07] describes how flow and games can be combined. Flow is a concept introduced by Mihaly Csikszentmihalyi which represents a feeling of complete and energized focus in an activity with a high level of enjoyment and fulfillment. Csikszentmihalyi identified eight major components of Flow:

- A challenging activity requiring skill
- A merging of action and awareness
- Clear goals
- Direct, immediate feedback
- Concentration on the task at hand
- A sense of control

- A loss of self-consciousness
- An altered sense of time

Flow can be induced by any kind of activity and Chen argues that most computer games include and leverage the eight components of Flow. Chen further argues that since all users have different skills and expectations an experience has to offer many different choices that adapts to the different user's Flow Zones. Chen concludes that to keep a user within the user's Flow Zone the game should offer choices that allow the user to enjoy Flow in their own way.

In Rules of Play [SZ04], Salen and Zimmerman also describe flow as one approach to describe pleasure in games. Like Chen, Salen and Zimmerman use the eight major components of Flow. They split it up in effects of flow and prerequisites for flow. The effects are:

- A merging of action and awareness
- Concentration on the task at hand
- A loss of self-consciousness
- An altered sense of time

And the prerequisites are:

- A challenging activity requiring skill
- Clear goals
- Direct, immediate feedback
- A sense of control

These four prerequisites are the ones that need to be maximized. Csikzentmihaly has furthermore distilled these four prerequisites into two dimensions: Challenge and ability. Figure 2.1 shows the two dimensions and the flow zone.

Csikzentmihaly shows that if a person has a high ability but the challenge is low then the person feels boredom. If a person is low in ability but the challenge is too great then that person will have a negative and intimidating experience leaving the player to feel anxiety. If the person is challenged to his/her ability then the person is in the flow zone.

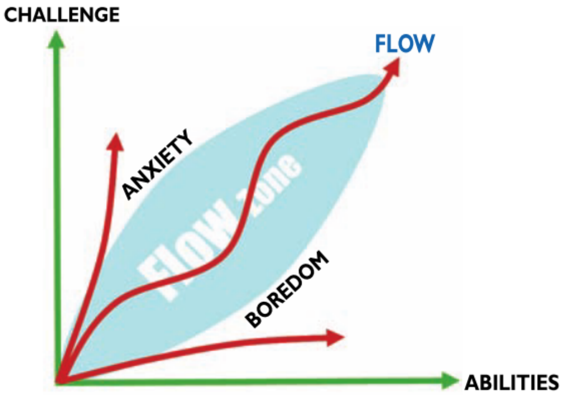


Figure 2.1: Flow zone in a challenge/ability chart.(From Flow in games [Che07])

Looking at the four prerequisites and their application in games Salen and Zimmerman discuss the importance of fulfilling these prerequisites. They emphasize the relationship between action and outcome. This relationship is presented by the clear goals and direct, immediate feedback from the prerequisites. The goals of a game must always be clear and the game must communicate where the goal is, how it might be achieved and whether the player is making progress toward it. The game must also make it clear for players what the outcome is to any certain action.
The two other prerequisites deal with challenge. Salen and Zimmerman argue that an anxiety state will result in the game feeling arbitrary. Therefore the challenge level must match the ability of the player. Salen and Zimmerman mention Dynamic Difficulty Adjustment or DDA as an approach to the management of challenge in games.
DDA is a technique that uses feedback loops to adjust the difficulty of play. An example is Crash Bandicoot where the player must maneuver through a series of jumping and dodging obstacles, overcome damaging hazards and reach objectives to finish a level. DDA is implemented in Crash Bandicoot by evaluating the number of times a player has died in a particular location in a level and adjusting the difficulty based on that criteria. If the player is having trouble the game automatically puts in a few more helpful objects or removes some enemies.

Hunicke implements a DDA system in the article "The Case for Dynamic Difficulty Adjustment in Games" [Hun05]. Hunicke describes several methods of implementing DDA in games. DDA can be manipulating the game economy, the narrative structure, the artificial intelligence or even the physical layout of maps or levels. Hunicke also notes that it is important when implementing DDA that

it must be unobtrusive. An example of badly implemented DDA is the Rubber Band effect which is often featured in racing games. The rubber band effect makes sure that the player does not get too far ahead of the computer which in racing games mean that no matter how far ahead a player is, the computer can always catch up by suddenly driving faster.
For the DDA implementation Hunicke is manipulating the game economy in an FPS. The game economy in an FPS is the amount of health of the player and enemies, amount of ammunition in the level, amount of health in the level and so on. By manipulating these currencies the game will become more or less challenging to the player which can be used to keep the player in the flow zone. The DDA is implemented by calculating the probability of player death. Based on this metric the game can adjust the difficulty by means of the game economy. The results show that the mean of player deaths decreased from 6.4 to 4.0 and that expert players report slightly elevated levels of enjoyment.
Hunicke concludes that even the small changes made in the game showed improvements in player performance without removing agency from the player.

For this project to work a clear method to improve game experience has to be implemented. Using the theory of flow by Csikzentmihaly a clear method for maximizing game experience is found by maximizing the time a player spends in the flow zone. Salen and Zimmerman identified four of Csikzentmihaly's eight major components of flow as being prerequisites for a flow experience. Csikzentmihaly has identified two dimensions on which the four prerequisites can be described: challenge and ability. The ability relies on the player's skill while challenge both relies on the player's ability but also on the level of challenge posed by the game. By having the game adjust the challenge according to the player ability the player can get into the flow zone. This is what Dynamic Difficulty Adjustment (or DDA) can do.
By having the difficulty adjusted dynamically, the challenge posed to the player can be changed dynamically. Hunicke presents a system for DDA in an FPS game. This system uses the game economy as a dynamic system in which the difficulty can be adjusted. The results showed better player performance and a slight improvement in enjoyment among expert players.
DDA has also been used several times in commercial games. Examples include: Half-Life 2 [Tol08], Max Payne [R* 12], Prey [Sie06], The Elder Scrolls IV: Oblivion [Hic11], Crash Bandicoot [SZ04], Jax and Daxter [SZ04], Left 4 Dead [Sal09] and more.
Hunicke's results and the use of DDA in commercial games show that DDA is a viable tactic for increasing game experience.

Chapter 3

Emotion Mapping

The current project requires a player's mental state to be identified before anything else can happen. For entertainment purposes the most important element is to tap into the emotions of a player. This necessitates the need for a terminology for emotions and the exploration of different emotion theories.

In a review by Hamann [Ham12] theories of emotion are described and the controversy between the theories are discussed.

There are two main theories of emotion. One where emotion can be conceptualized as discrete categories and one where emotion can be conceptualized as dimensions such as arousal and valence.

Discrete emotion theories usually propose a small number of emotions that are the most basic and universal emotions in humans. These emotions have specific characteristics and unique physiological and neural profiles. Emotions such as happiness, sadness, anger, disgust, fear and surprise are often mentioned as being the most basic emotions.

Dimensional emotion theories propose that emotions can be mapped onto a few dimensions where the most often dimensions mentioned are arousal and valence. Arousal is a measure of how alert a person is and ranges from calm to excited. Valence is a measure for how pleasurable an event is to a person and ranges from highly positive to highly negative.

Hamann describes studies which study the discrete emotion theory. The results point towards limited evidence for consistent associations between brain regions and basic emotions. Rather the results indicated that one brain area could be activated by different emotions instead of one brain area per emotion.

Hamann writes that meta-analyses of the dimensional emotion theory of arousal and valence show that representation of arousal and valence is quite complex. The studies show that arousal and valence both involve multiple brain regions and that these dimensions are not independent.

Hamann states that a one-to-one mapping of emotions to brain regions is not possible and that a network model where each emotion is a network of multiple brain regions is more correct.

The Hamann report suggests that a network view of emotions is a future direction for studies in emotion representation. Currently the choice for the current project is between basic emotions or a dimensional emotion representation since it is too early to be using a network view of emotions. The network view furthermore requires fMRI which requires a person to lie still inside a large machine and therefore limits it to clinical studies. Since a network view of emotions requires a high spatial resolution, it suggests that EEG would not useful since spatial resolution is one of the weaknesses of EEG. This leads to the conclusion that the valence/arousal model is clearly a better choice. This is due to the simplicity of having two parameters, valence and arousal, that can be tweaked. Furthermore it is widely accepted that arousal correlates with skin conductance response and valence with facial muscle activity. This simplifies the process of reading the emotion of a person and combined with the fact that the signals from these sensors are stronger than EEG signals this makes the valence/arousal model more suited for the current project.

Chapter 4

Physiological Measures

Since the current project is about measuring a player's mental state and emotions a clear approach to measuring the emotions of a player has to be found. Two approaches are described below. The first is measuring electrical activity on the scalp also known as EEG. The second is physiological measures that does not include EEG such as skin conductance, heart rate and temperature. It is important to correctly assess the emotions of the player and for this a clear and precise method of measuring emotion has to be found. In the following sections both approaches are described and results of the approaches are presented and compared.

4.1 EEG

Electroencephalograhy or EEG is the recording of electric activity on the scalp. The electrical activity corresponds to neuronal activity in the brain. The use of EEG for detecting emotions in games has not been a large research topic. For this report four studies are presented. A study by Chanel [CKGP06] and a study by Bos [Bos06] both look at recognizing emotions using images as stimuli. Both of these studies use images from the International Affective Picture System [Uni11] or IAPS which is an image database compiled by Bradley and Lang where all images are tagged with valence and arousal. The two other studies uses games as stimuli – Chanel 08 [CRBP08] and Reuderink 13 [RMP13]. Interestingly all studies uses the valence/arousal model of emotion with the Reuderink study using an additional dominance dimension.

The Chanel 06 [CKGP06] study details an experiment where 4 participants were asked to look at images from the IAPS database. After each image the participants had to self-assess their valence and arousal using the self-assessment manikin [BL94] or SAM which is a non-verbal pictorial technique to measure valence, arousal and dominance. The participants wore an EEG device, a galvanic skin response (GSR) sensor for skin conductance, plethysmograph for blood pressure measurements, respiration belt for abdominal and thoracic movements and a temperature sensor. The participants were asked to look at 100 images each – 50 images that were tagged as low arousal in the IAPS database and 50 images that were tagged as high arousal.
The data was processed using both a Bayes classifier and a classifier based on Fischer Discriminant Analysis (FDA). The classifiers were trained for each participant with leave one out cross validation where one pair of classifiers (Bayes and FDA) used two ground-truth classes and another pair used three ground-truth classes. The ground-truth classes were constructed from the SAM which contained 5 steps of arousal from calm to exciting.
The results show a mean over all subjects as 60% for the Bayes classifier using only EEG features and two ground-truth classes while it is 55% for the FDA.
Using three ground-truth classes the mean for the Bayes classifier using only EEG features is 45% while for FDA it is 40%.
Figure 4.1 shows a plot of the reults.
Chanel concludes that EEG signal seem to perform better than physiological signals and that the concatenated signals (both EEG and physiological) with FDA shows the best performance. Another conclusion from the study is that using the self-assessments as ground-truth classes were better than using the IAPS ratings.

The results from the study show unconvincing results. For all of the classification problems the mean over all participants are barely above chance level and for some participants the mean is below chance level.
Apart from unconvincing results the results also show large variance with participants below chance level and other participants well above chance level.
Another problem is the fact that the concatenated signals are not equal to or better than the best non-concatenated signal. This is seen in Figure 4.1 in the top left plot. Here the EEG signal is better than the concatenated signal which is odd since the concatenated signal also contains EEG signals and could therefore choose to disregard the physiological signals to get the same performance of the EEG signals.
The study also builds a classifier for each participant which makes it less useful

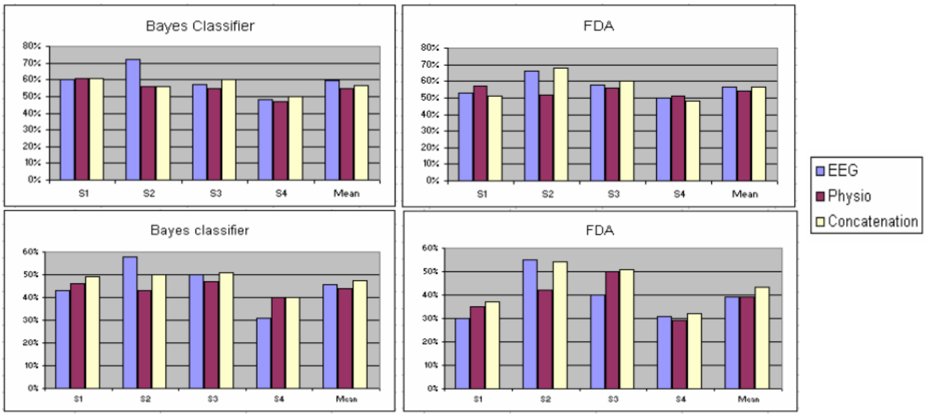


Figure 4.1: The plots in the top row are the results for classifying into two classes. Bottom row plots are for the result for classifying into three classes. Left column is the plot for Bayes and right column is for FDA. Along the axes are subjects and mean on the x-axis and how correct the classifier is in percentages along the y-axis. The blue bar is classification from EEG signals, the red bar is classification from physiological signals without EEG signals and the beige bar is the fusion of EEG signals and physiological signals.

for practical applications.

Reuderink [RMP13] describes an experiment where the goal is to confirm or reject hypotheses about correlations between EEG signals and arousal, valence and dominance. The experiment was conducted with 12 participants who played a rigged pacman game. The participants played 30 minutes of the pacman game in 2-minute blocks wherein one third of the 2-minute blocks would ignore 15% of key presses and therefore induce frustration. After each block the participants were asked to fill out a SAM.
The participants were wearing an EEG advice, EOG for measuring eye position and an EMG device for measuring muscular activity. Other sensors were also worn but the data was not used. The EOG and EMG were used for finding ocular and muscular artifacts.
The results of the SAM which can be seen in Figure 4.2 indicate that frustration leads to low valence and low dominance.
The study also looks at the relation between time, experimental condition, valence, arousal and dominance which is gathered from the SAM. A table of correlations can be found in Figure 4.3. The study shows that there are no correlations with time and anything else which suggests that the emotional ratings do not drift over time. There is also a correlation between valence and dominance which according to the study should not be since valence and dominance are supposed to be orthogonal.

The focus of the study is however to study several hypotheses regarding to measuring valence, arousal and dominance. Most notably Reuderink studies the hypothesis that left-right brain alpha asymmetry as a measure of valence. That is the asymmetry in the alpha frequency band between the left and right brain hemispheres is related to valence. A full list of hypotheses can be seen in Figure 4.4.

The results of testing the alpha asymmetry hypothesis can be seen in Figure 4.5 Reuderink concludes that they found correlates for valence and arousal in theta, delta and alpha bands. Reuderink also confirms that alpha asymmetry can be used as a measure for valence.

The results in both Figure 4.2 and 4.3 show some interesting observations in which gameplay dominance seems to correlate with valence while valence and arousal does not correlate.
This study shows that the dominance dimension is not needed for the current project because the dominance and valence dimensions correlate. This is even more relevant when taking into account that the environment of the study was a computer game because it shows that the valence/arousal model is good enough for computer games and that the dominance dimension is unneeded since it correlates with the valence dimension.
The correlates found in the study are not very strong. The correlates can be found in Figure 4.6. Here the correlations typically range from 0.2 to -0.2 and

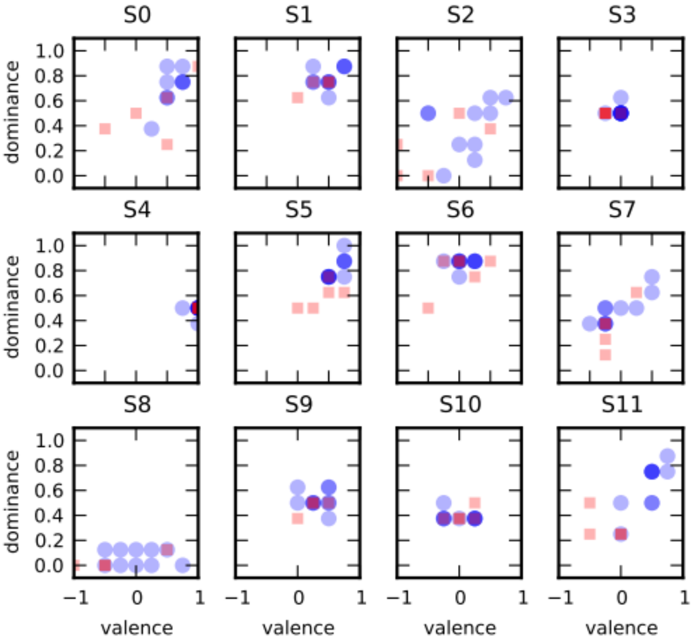


Figure 4.2: Valence and dominance ratings plotted for each participant. The normal play sessions are denoted by blue circles, frustration sessions are denoted by red squares.

	Time	Cond.	Val.	Ar.	Dom.
Time	1.00	0.01	−0.04	0.07	0.01
Condition	0.01	1.00	−0.32	0.08	−0.32
Valence	−0.04	−0.32	1.00	0.10	0.43
Arousal	0.07	0.08	0.10	1.00	−0.15
Dominance	0.01	−0.32	0.43	−0.15	1.00

Figure 4.3: Table of correlations.

Dimension	Delta	Theta	Alpha	Beta	Gamma
Valence ↑	$H_{v\delta}$: fron.-med. ↑	$H_{v\theta1}$: l-hemi. ↑ $H_{v\theta2}$: r-hemi. ↓ $H_{v\theta3}$: fron.-med. ↑	$H_{v\alpha1}$ l-hemi. ↓ $H_{v\alpha2}$ r-hemi. ↑	-	$H_{v\gamma1}$: l-temp. ↓ $H_{v\gamma2}$: r-temp. ↑
Arousal ↑	$H_{a\delta}$: posterior ↑	$H_{a\theta}$: posterior ↑	$H_{a\alpha1}$: frontal ↑ $H_{a\alpha2}$: global ↓	$H_{a\beta}$: parietal ↑	$H_{a\gamma}$: gamma ↑
Dominance ↑	-	-	-	-	-

Figure 4.4: Table of hypotheses for valence, arousal and dominance. Note that no hypotheses for dominance are listed.

	Fp1-Fp2	AF3-AF4	F3-F4	FC1-FC2	C3-C4	F7-F8	P3-P4
Condition	−0.15∗	−0.04	−0.07	−0.00	−0.09	−0.05	0.05
Valence	0.19	0.00	0.12	0.16∗	0.11	−0.05	0.06
Arousal	0.00	−0.13	−0.12	−0.16∗	0.08	−0.04	0.06
Dominance	0.07	0.01	0.18∗	0.18 ∗ ∗	−0.06	0.03	−0.01
PC0	0.19∗	0.04	0.08	0.05	0.11	0.02	−0.01
PC1	0.03	0.04	0.03	−0.07	0.03	0.10	−0.12
PC2	−0.01	0.13	0.17∗	0.19 ∗ ∗	−0.13	0.05	−0.04
PC3	−0.05	−0.10	−0.08	−0.14	−0.04	−0.00	0.02

Figure 4.5: Table of correlations for alpha asymmetry for different sensor pairs. PC0-3 indicate principal components where PC0 more or less corresponds to positive emotions, PC1 corresponds to negative emotions, PC2 to relaxed and dominance, PC3 to aroused and dominance. One asterisk indicates $p<0.05$, two indicates $p<0.01$

this is and seeing as the EEG can show both positive and negative correlations right next to each other (e.g. valence at 3 Hz) it shows that EEG does not provide good enough measurements and that the spatial resolution on EEG devices cannot provide enough detail as mentioned in the Hamann [Ham12] study.

Reuderink confirms that alpha asymmetry can be used as a measure for valence although looking at the valence row in Figure 4.6 it looks to be a weak correlation. Figure 4.5 also shows weak correlates. Furthermore Reuderink concedes that for most of the hypotheses no significant correlations were found and this greatly reduces the usage of EEG as an emotion recognizer.

The Bos 06 [Bos06] study is similar to the Chanel 06 study. The experiment in this study was one where 5 participants had to look at 12 images from the IAPS database, listen to 12 sounds from the International Affective Digital Sounds (IADS) database which are sounds rated for valence and arousal, and lastly look at and listen to 12 audiovisual stimuli which are IAPS images and IADS sounds combined.

Each participant was wearing an EEG device and the resulting data was used for classification with FDA. The classification is for two classes – high and low arousal or valence and a classifier was trained for each participant using 3-fold cross validation.

The results show the average classification rate for arousal to be around 60-70% with the best classification rate around 90%. For valence the results are 60-70% average classification rate and around 90% for the best classification rate.

The results for arousal classification are seen in Figure 4.7 and for valence in Figure 4.8.

Bos concludes that their good results would not be equally impressive if the classified arousal and valence has to be used to classify emotions as this could

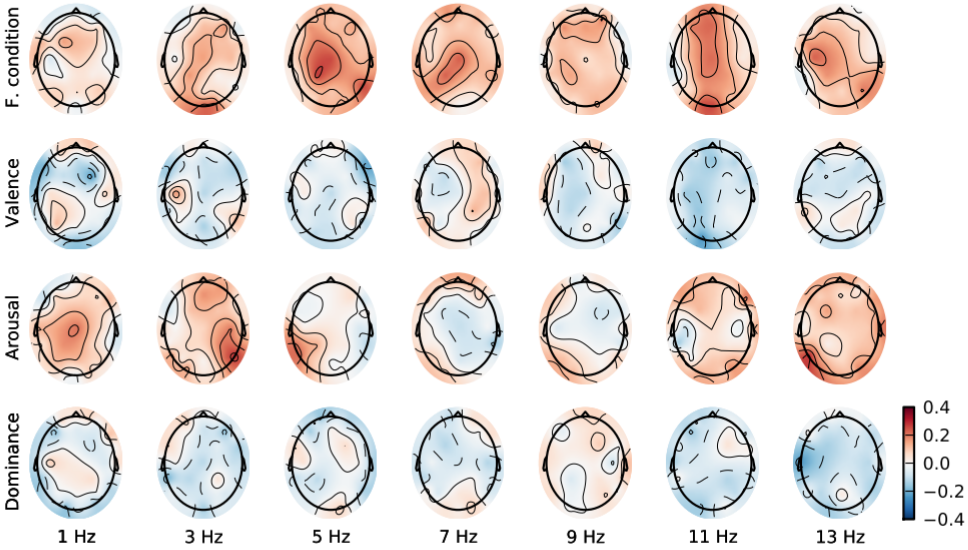


Figure 4.6: Mean subject correlations where the rows indicate frustration, valence, arousal and dominance while the columns is the frequencies. Each circle represent a spatial view of a head where colors indicate correlations. For example does arousal at a frequency of 3 Hz show a correlation at about 0.3 at the right back of the head.

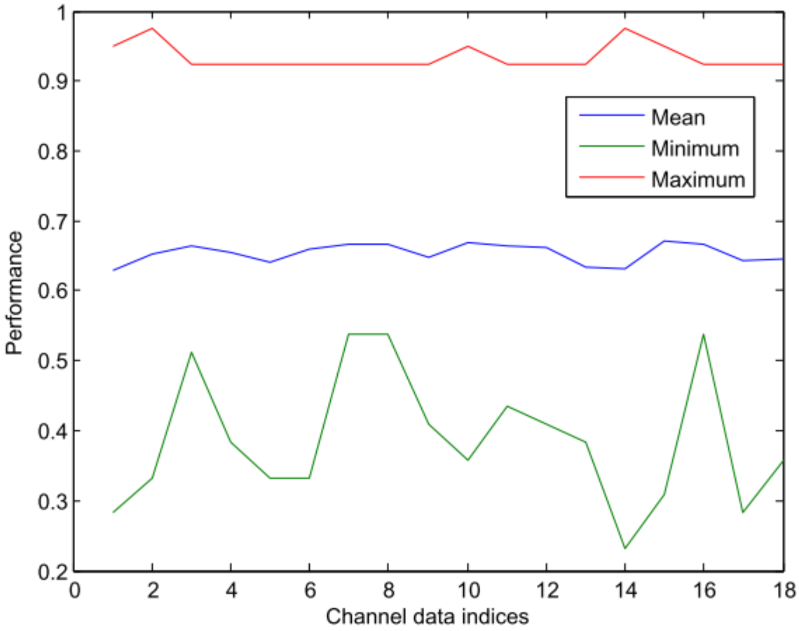


Figure 4.7: Performance of each channel feature for arousal classification. The performance is percentages correct for classification in two classes – high arousal and low arousal. The channel data indices refer to features which are combinations of frequency bands over the EEG channels. That means channel 1 is a feature which is a certain frequency band measured on a specific EEG channel.

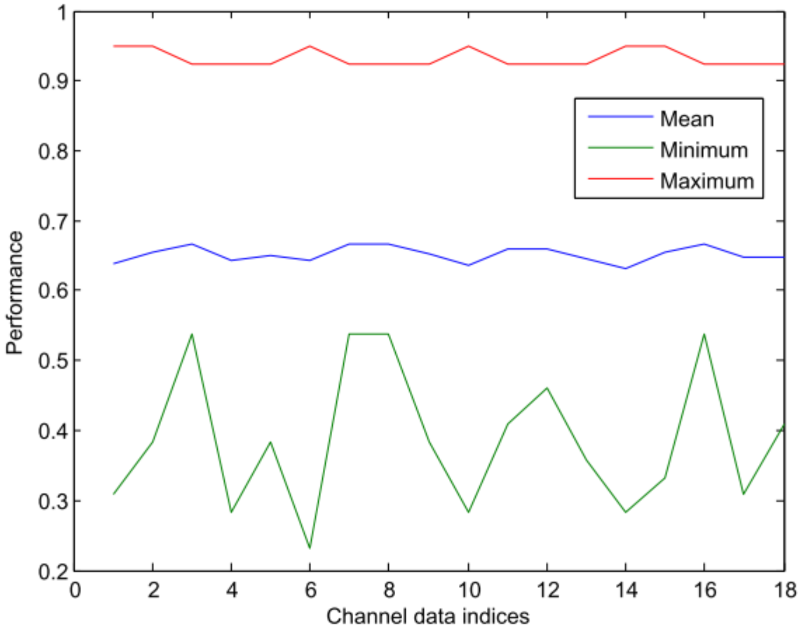


Figure 4.8: Performance of each channel feature for valence classification. The performance is percentages correct for classification in two classes – high arousal and low arousal. The channel data indices refer to features which are combinations of frequency bands over the EEG channels. That means channel 1 is a feature which is a certain frequency band measured on a specific EEG channel.

introduce more errors.

The study by Bos is on close inspection not good enough. One problem is that the channel data indices are not stated anywhere although the conclusion does provide the frequency band and location of the electrodes for the best performing channel.
The main problem is however that the worst performing classifier is just above chance level for some few frequency band and location of electrodes combinations. This is a worry since the mean is around 60-70% and the maximum is at 90%. That indicates large variance in the data or at least some outliers.
Another smaller problem is the fact that the ground-truth classes for the classifiers are based on the IAPS/IADS ratings which is a problem because the images and sounds will not have the same effect on people. This is something that Chanel found out in the Chanel 06 study where he decided to bypass the IAPS ratings and solely use the SAMs.
As seen in the Chanel 06 study a classifier was trained for each participant which again makes the usage of EEG or at least the method from this report less practical.

Another study from Chanel[CRBP08] describes an experiment where emotions are measured using physiological measures and a computer game as stimuli.
Using the psychology of Flow Chanel sets out to measure three states: anxiety, flow and boredom. Since direct measurement of these three states cannot be done Chanel has mapped the three states to the valence/arousal theory of emotion. Anxiety is therefore negative valence, high arousal, flow as positive valence and high arousal and boredom as negative valence and low arousal.
The experiment was conducted with 20 participants where each participant wore an EEG sensor, plethysmograph, respiration belt, GSR sensor and a temperature sensor. The game was Tetris where each participant would play for a while to determine their skill level. The skill level would equal a medium difficulty and easy and hard difficulty was calculated from the medium difficulty. After every 5 minutes of play the participants filled out a SAM and a 30-question questionnaire. The participants would play two 5 minute sessions of each difficulty level.
The data was classified into three classes of bored, flow and anxiety using support vector machines and using signals from all sensors except EEG. The classifier for each participant was trained using features from the other participants.

The results of the SAMs show no surprises in that the medium difficulty shows the highest valence and medium level of arousal. The arousal increases as the difficulty increases. These results can be seen in Figure 4.9.

The results of the classification are presented in Figure 4.10. The overall classification rate is 53.33% for 3 classes though there are quite significant differences

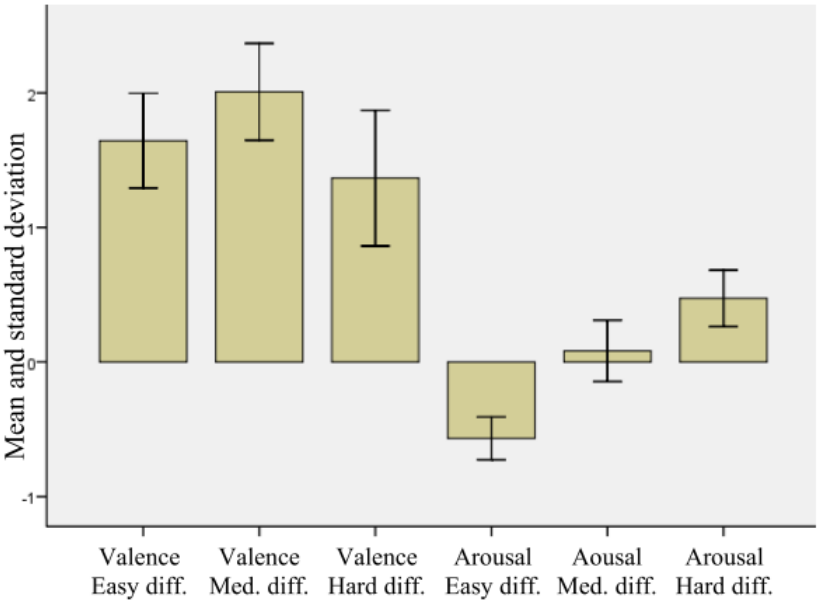
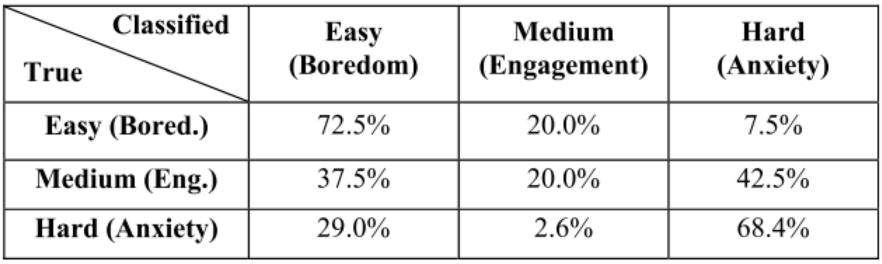


Figure 4.9: Mean and standard deviation for valence and arousal in the three difficulties. Valence and arousal values are from questionnaires.



Classified / True	Easy (Boredom)	Medium (Engagement)	Hard (Anxiety)
Easy (Bored.)	72.5%	20.0%	7.5%
Medium (Eng.)	37.5%	20.0%	42.5%
Hard (Anxiety)	29.0%	2.6%	68.4%

Figure 4.10: Confusion matrix for the classification of the three classes.

between each category. For boredom the classification rate is 72.5%, flow is 20% and anxiety is 68.4%.

Chanel concludes that the classification rates for two of the emotions, boredom and anxiety, are good. Chanel also concludes that the classification of flow is unnecessary since it tended to be classified as boredom or anxiety instead and proposes only measuring for boredom. Chanel argues that while boredom is unwanted in games, anxiety is not entirely unwanted as a greater challenge than the player's ability can stimulate the player to learn and perform better at the game.

The study shows good results although EEG was not used. The classification rates for boredom and anxiety are great although there is a worry with the classification rate for the flow state. The classification rate for the flow state is 20% which means that the classifier is more likely to classify the flow state as either boredom at 37.5% or anxiety at 42.5%. This is a worry but as Chanel argues it might be a better solution to only detect boredom and perhaps extreme frustration or anxiety.
Furthermore classification was done using other participants as training set which means that the classifier and results are not user specific. This is very useful for the current project since it is cumbersome to train a classifier each time a user wants to use the system.

The two studies, Chanel 06 and Bos 06, do not show good results which brings the usability of EEGs for emotion recognition into question. The Chanel 08 study shows good results but the classification only uses physiological signals other than EEG and as such it also suggests that EEG is not suited for emotion recognition.
The studies that do use EEG for emotion recognition do not provide good results. The Chanel 06 study shows unconvincing results. For classification in two classes using Bayes classifier the mean of all participants using EEG is 60%. That is 10 percentage point better than chance level. For classification

Study	Class. classes	Chance level	Findings	PP diff.
Chanel 06	2	50%	60%	10
Chanel 06	3	33%	45%	12
Bos 06 (arousal)	2	50%	65%	15
Bos 06 (valence)	2	50%	65%	15

Table 4.1: Comparison of EEG results. PP is an abbrev. for percentage points

into three classes using Bayes classifier the result is 45% which is 12 percentage points better than chance level. A comparison is shown in Table 4.1.

The Reuderink 13 study is not about classification so it cannot be easily compared with the other studies. The study instead finds weak correlates between specific brain areas, frequency bands and valence, arousal and dominance which does not indicate that EEG is a good enough solution. Furthermore Reuderink concedes that for many hypotheses few could be confirmed.

The Bos 06 study shows some confusing results. In classification with two classes using binary linear FDA the classification rate for both valence and arousal is around 65%. It is therefore 15 percentage points better than chance level. The downside is however that there seems to be large variance in the data since the minimum performance is around 50% and in most cases below 50% while the maximum performance is consistently around 90%.

For both Chanel 06 and Bos 06 studies the percentage points difference between chance level and findings are 10-15 percentage points. That is a quite modest result and on closer inspection it turns out that the two studies show large variance. For the Bos 06 study minimum performance is around 50% or a chance level and the maximum performance is around 90%. The Chanel 06 study has participants that perform worse than chance level and some who perform much better than the average. This is a major worry that not only are the results modest, the variance seems to indicate that performance could be person dependent.

One issue with EEG for emotion recognition is that there is no consensus on which frequency bands and brain areas that represent specific emotions. This would greatly complicate the current project. Furthermore as Hamann states, the usage of EEG for emotion recognition is not recommended since results from other studies show that emotions might be represented by a network of brain areas and not just a single brain area. This would require some spatial view of the brain which is one of the weaknesses of EEG and together with the results from the Chanel 06,Bos 06 and Reuderink 13 studies, using EEG is not recommended for the current project.

4.2 Other Physiological Measures

The human body is a massive signal generator. These signals can be measured with various instruments. The ones presented are candidates for emotion recognition in game applications.

GSR or Galvanic Skin Response which is also known as electrodermal activity or skin conductance level is a technique where a device will measure the conductivity of the skin through the sweat level of the skin. This is useful because when people experience physical arousal sweat is produced which leads to better conductivity.

HR or Heart Rate is measured through a heart rate monitor but can also be measured using many different methods.

BVP or Blood Volume Pulse is a method for measuring blood flow through skin capillary beds in the finger. A BVP sensor is put on a participant's finger and gives a measure on anxiety. This is due to the "cold feet" phenomenon where blood drains from the extremities a person during periods of emotional duress. The BVP sensor can also be used for HR measurements.

Another physiological signal is muscle activity. This is measured with EMG or Electromyography. EMG detects surface voltages when a muscle is contracted [MIC06]. In emotion recognition EMG are often used on facial muscles to detect smiles and frowns.

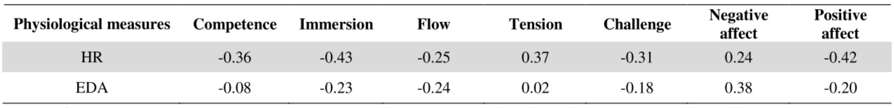
In this section four studies are presented. One by Drachen [DN10] is about finding correlates between physiological signals and game experience. The three others, Lisetti 04 [LN04], Kim 04 [KBK04] and Scheirer 02 [SFKP02], are about emotion recognition. The use of stimuli varies, Drachen and Scheirer use computer games, Lisetti uses movie clips and Kim uses audio, visual and cognitive stimuli.

The Drachen study [DN10] describes an experiment wherein 16 participants were playing 3 computer games (all in the First-Person Shooter genre) while wearing a HR monitor and a GSR device. Every 5 minutes the participants were asked to complete a questionnaire called iGEQ or ingame Game Experience Questionnaire. The iGEQ is a self-report scale for player experience and contains seven dimensions of player experience: Immersion, Flow, Competence, Tension, Challenge, Negative affect and Positive affect.

The data was filtered and the correlations were calculated with Pearson's correlation coefficient.

The results shown in Figure 4.11 show that there is a correlation between low heart rate and positive affect, flow, low challenge, immersion and feelings of competence. There is a correlation between high heart rate and tension and negative affect.

For GSR (EDA in the figure) there is a correlation between low skin conductivity and immersion, flow and positive affect while there is a correlation between



Physiological measures	Competence	Immersion	Flow	Tension	Challenge	Negative affect	Positive affect
HR	-0.36	-0.43	-0.25	0.37	-0.31	0.24	-0.42
EDA	-0.08	-0.23	-0.24	0.02	-0.18	0.38	-0.20

Figure 4.11: Pearson correlation coefficient for the seven dimensions of iGEQ and HR/GSR.

Classifier	Sadness	Anger	Surprise	Fear	Frustration	Amusement
KNN	70.4%	70.8%	73.9%	80.9%	78.3%	69.6%
DFA	77.8%	70.8%	69.6%	80.9%	72.7%	78.3%
MBP	88.9%	91.7%	73.9%	85.6%	77.3%	87.0%

Table 4.2: Comparison for classification rates of each classifier.

high skin conductivity and negative affect.

Drachen concludes that physiological measures correlate with game experience although it is dependent on context and approach.

The results indicate that game experience does correlate somewhat with physiological signals. One worry is that the correlations are not strong although it might not be a problem for the current project but this may not prove a problem since the correlations are for the seven dimensions of game experience. Positive and negative affect does however show the strongest correlation which is good since positive and negative affect could relate a lot with valence.

The Lisetti study [LN04] attempts to recognize emotion from physiological signals using movie clips as stimuli. The experiment had 29 participants watch movie clips while having their HR, GSR and temperature measured. Another study had preceded this where the movie clips were chosen. 5 movie clips were chosen from that study to represent sadness, anger, amusement, fear and surprise and the participants were also asked to solve a difficult mathematical problem without aids which was to represent frustration. After each clip/task the participants were asked to fill out a questionnaire about whether the participant felt the intended emotion, rate the intensity of that emotion and whether other more intense emotions were felt.
The data resulted in 12 features and for each of the six classes three classifiers were trained. The three classifiers were k-nearest neighbor algorithm (KNN), discriminant function analysis (DFA) and the Marquardt backpropagation algorithm (MBP). The KNN was tested with leave one out cross validation.
The results are shown in Table 4.2 and the results for the best classifier, MBP, are shown in Figure 4.12.

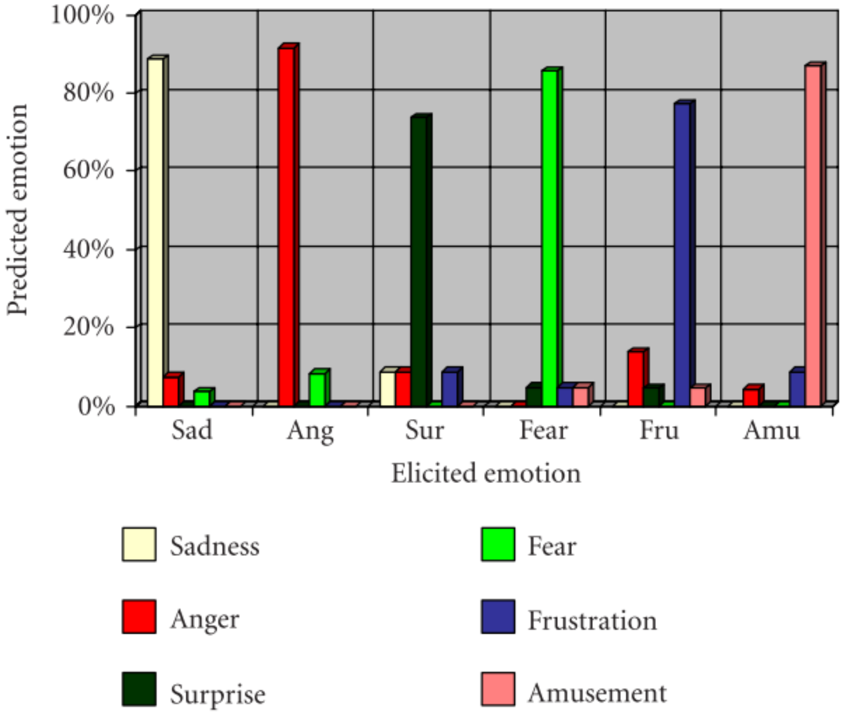


Figure 4.12: The results of classification using the MBP algorithm. For each emotion on the x-axis there is a bar for each emotion that has been classified. The height of the bar corresponds to the percentage that the emotion has been classified.

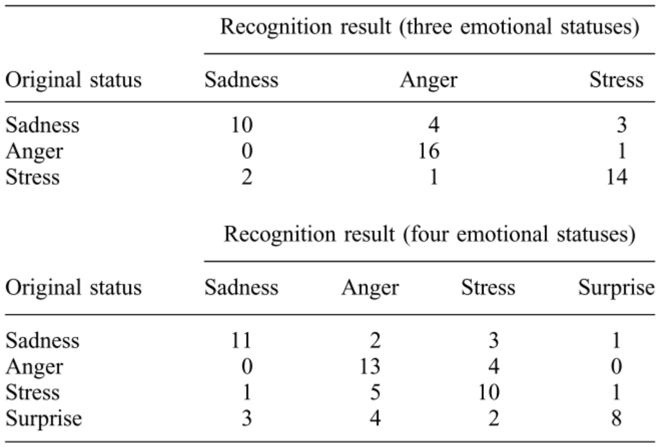
The results showed that the DFA algorithm was better than KNN for sadness, frustration and amusement, KNN was better for surprise and MBP was best in all classes except surprise and frustration.
The study concludes that the three algorithms can categorize emotions with 72.3%, 75.0% and 84.1% accuracy for KNN, DFA and MBP, respectively.

The study present some great results since the average classification rates are 72.3%, 75.0% and 84.1% for a six class problem.
There are however a problem that could have had an effect on the classification rates. The problem is that Lisetti only use one clip to represent each emotion. This could lead to the classifier to classify according to which clip was shown and not an emotion. Lisetti asked the participants whether or not they thought the clip evoked the intended emotion. For some of the clips the agreement rate are around 50-60%. This means that a specific clip only invoked the intended emotion in half of the participants or they associated another emotion with the movie clip. In connection the classifier could have classified according to the physiological footprint of the clips instead of classification of the emotion felt. As such this study cannot definitely show that GSR, HR and temperature are good choices for emotion recognition.
Finally the KNN classifier was used with leave one out cross validation which results in a classifier that is not user specific.

The study Kim 04 [KBK04] describes an experiment where participants were subjected to auditory, visual and cognitive stimuli in an attempt to recognize emotion. The participants were 175 children in the age of 7-8 years and the experiment was conducted in two occasions with 125 children in the first and 50 children in the second. The children had their GSR, HR and temperature measured.
The first experiment was to recognize sadness, anger and stress while the second was to recognize the three emotions mentioned and the emotion surprise. Apart from measuring an additional emotion the second experiment was also conducted with instructions for the children to be as still as possible.
The data was classified with support vector machines and the result of the second experiment is shown in Figure 4.13.
The results for the first experiment returned a classification rate of 55.2% for three classes. The second experiment returned a classification rate of 78.4% for three classes and 61.8% for four classes.

Kim concludes that a novel emotion recognition system has been developed based on the physiological signals of HR, GSR and temperature since the classification rate is significantly higher than chance level.

The results presented are great with classification rates of 78.4% for three classes and 61.8% for four classes. This is from the second experiment and the difference



	Recognition result (three emotional statuses)		
Original status	Sadness	Anger	Stress
Sadness	10	4	3
Anger	0	16	1
Stress	2	1	14

	Recognition result (four emotional statuses)			
Original status	Sadness	Anger	Stress	Surprise
Sadness	11	2	3	1
Anger	0	13	4	0
Stress	1	5	10	1
Surprise	3	4	2	8

Figure 4.13: Recognition results for the second experiment. The numbers indicate number of participants.

between the first and second experiment can be a worry. For the current project the system is intended to be used in a living room scenario and if the system places restrictions then it might not be widespread. For the second experiment the children were asked to sit as still as possible and the results improved from 55.2% to 78.4%. In reality it might not pose a problem but it is something has to be considered before the current project can be released.

An issue is that the classification was based on a training set consisting of a third of the entire sample. The classification rate was then calculated on the rest of the sample. What would have been interesting is if Kim ran cross-validation on the data set. This is however a minor problem since Kim did use one third of the children as training and the rest as testing. This ensures that the classifier is not user specific.

Questions could also arise on the use of children in the study. Would the results change if adults were used instead?

Although results are not as great as the Lisetti 04 study it does point toward the results in the Lisetti 04 as being representative of emotion recognition.

Scheirer 02 [SFKP02] is a study that attempts to classify whether a participant is frustrated or not. In the experiment conducted by Scheirer 36 participants were to play a computer game where the best participant would receive a cash prize. Frustration was induced by rigging the mouse to ignore clicks and the participants wore a GSR sensor and a blood volume pulse sensor.

The data was classified with Hidden Markov Models (HMM) using data from 24 participants. The ground-truth classes were frustration and non-frustration. The frustration classes consisted of the 10 seconds after a frustration event and

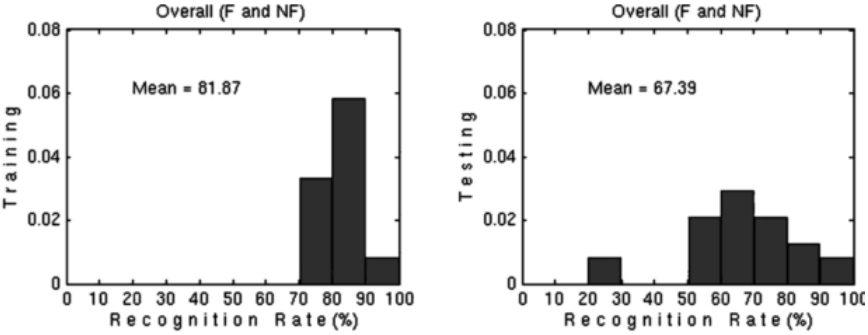


Figure 4.14: Overall classification rate for frustration and no frustration. Results are from 24 participants and therefore from 24 HMM structures. The plot on the left are classification rates for the training and the right are for the testing.

the rest of the samples were labelled as non-frustration. A HMM was trained for each participant using the data from the other participants.
The results which can be seen in Figure 4.14 show an overall classification rate of 81.87% for the training set and 67.39% in the testing set. The classification rate for frustration is 53.26% and for no frustration the classification rate is 72.44% which can be seen in Figure 4.15. Scheirer concludes that results might have been influenced by the labelling of ground-truth classes. The strategy of labelling the following 10 seconds after a frustration event is perhaps not the best as evidenced in Figure 4.15, top right plot where some show a recognition rate well below chance level. Scheirer however also concludes that for 21 out of 24 participants the classifier could classify frustration and non-frustration with a classification rate above chance level.

The results are very modest. For recognizing frustration it only performs slightly above chance level with 53.28%. Overall classification rate is 67.39%. The study mentions the ground-truth classes as a problem and this can perhaps be seen in Figure 4.15 where the classifier completely fails to recognize frustration for some participants with classification rates below 30% and well below chance level. Although the classification rate is 67.39% that results might improve with better ground-truth class labelling.

Overall the four studies using physiological other than EEG present some good results.
The study by Drachen seeks to find correlates between physiological signals and player experience in computer games. They succeed somewhat, finding moderate correlates between HR, GSR and player experience. The results are

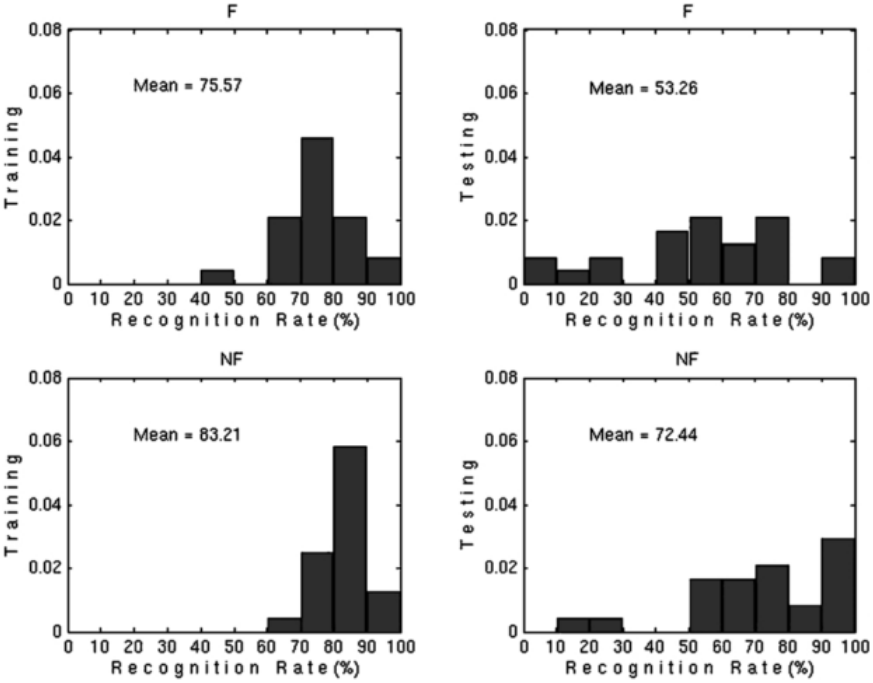


Figure 4.15: The classification rate for frustration in the top row. Bottom row is for no frustration. The top row is the classification rate for frustration when the samples are labeled frustration and the bottom row is the classification rate for no frustration when the samples are labeled no frustration. The results are from 24 participants. Plots on the left are classification rates for the training set and for testing set on the right.

Study	Class. classes	Chance level	Findings	PP diff.
Chanel 08 (boredom)	3	33%	72.5%	39.5
Chanel 08 (anxiety)	3	33%	68.4%	35.4
Lisetti 04 (KNN)	6	17%	72.3%	55.3
Lisetti 04 (DFA)	6	17%	75%	58
Lisetti 04 (MBP)	6	16%	84.1%	67.1
Kim 04	3	33%	78.4%	45.4
Kim 04	4	25 %	61.8%	36.8
Scheirer 02	2	50%	67.4%	17.4

Table 4.3: Comparison of the results of other physiological measures. PP is an abbrev. for percentage points

encouraging and provide good guidelines for a future implementation of the current project.
The other three studies are about classifying emotion based on physiological signals. All of the studies use stimuli which occupy two or more senses. For the Lisetti 04 study the stimuli was movie clips, the Kim 04 study used a mix of cognitive, visual and aural stimuli and the Scheirer 02 study used a visual puzzle. Another study is the Chanel 08 study described in the EEG chapter. Chanel used a computer game which involves cognitive and visual stimuli.
A comparison of the results is found in Table 4.3 below. Here the findings are much better compared to emotion classification using EEG signals. The worst result here (Scheirer 02) is still better than the best result using EEG signals.

The Lisetti 04 study presents the best results though one worry is that the classifiers might have been made to classify according to which clip the participants were watching instead of their emotion when watching the clip.
There are also differences in how the classification was done in the different studies that could impact the classification rates. Most notably if the test set was generated with samples from all participants the classifier is then trained for all participants and the classifier is less useful for predicting emotion for others. The Kim 04 study divided the participants into two groups: test and training and had some good results. The Scheirer 02 study trains a classifier for each participant which is not useful since each classifier only works for one participant. The Lisetti 04 study does not describe how the training set was generated while the Chanel 08 created a classifier for each participant which consisted of a training set which was composed of the other participants.
This is important when judging the results. The results from the Scheirer 02 study would not be usable for practical purposes because a classifier has to be trained for every new person. In comparison the Kim 04 classifier is plug and play – that is, it is usable on new people.

Study	Measures	Class. Method	PP diff.
Chanel 06 (2 class)	EEG	FDA	10
Chanel 06 (3 class)	EEG	FDA	12
Bos 06 (arousal)	EEG	FDA	15
Bos 06 (valence)	EEG	FDA	15
Chanel 08 (boredom)	GSR, BVP, HR, temp	RBF SVM	39.5
Chanel 08 (anxiety)	GSR, BVP, HR, temp	RBF SVM	35.4
Lisetti 04	GSR, HR, temp	KNN	55.3
Lisetti 04	GSR, HR, temp	DFA	58
Lisetti 04	GSR, HR, temp	MBP	67.1
Kim 04 (3 class)	GSR, HR, temp	SVM	45.4
Kim 04 (4 class)	GSR, HR, temp	SVM	36.8
Scheirer 02	GSR, BVP	HMM	17.4

Table 4.4: Comparison of the measures and classification methods. PP is an abbrev. for percentage points

Looking at the data analysis methods the different studies choose different methods. Table 4.4 summarizes the results and data analysis methods of the studies. The results show that the most successful studies uses both GSR, HR and temperature. This suggests that these three measures are very useful in recognizing emotion. Four of the studies use SVM and the results range from 35.4 to 45.4 which highly suggests that SVM is preferred. KNN, DFA and MBP show greater results but they are all from the Lisetti 04 study which says more about the study than the methods. The choice is between the MBP which is the best performing method and SVM which show great results across two studies.

The two EEG studies (Chanel 06 and Bos 06) furthermore trained classifiers for each participant while Chanel 08, Kim 04, Scheirer 02 and Lisetti 04 trained classifiers that were not user-specific which definitely is in favor other physiological measures and not EEG.

The mean of the difference in percentage points is 44.4 for physiological signals without EEG compared to the mean of 13 for EEG signals suggests that using a mixture of GSR, HR and skin temperature as physiological measures and classifying using SVM or MBP is the best solution for this project.

Chapter 5

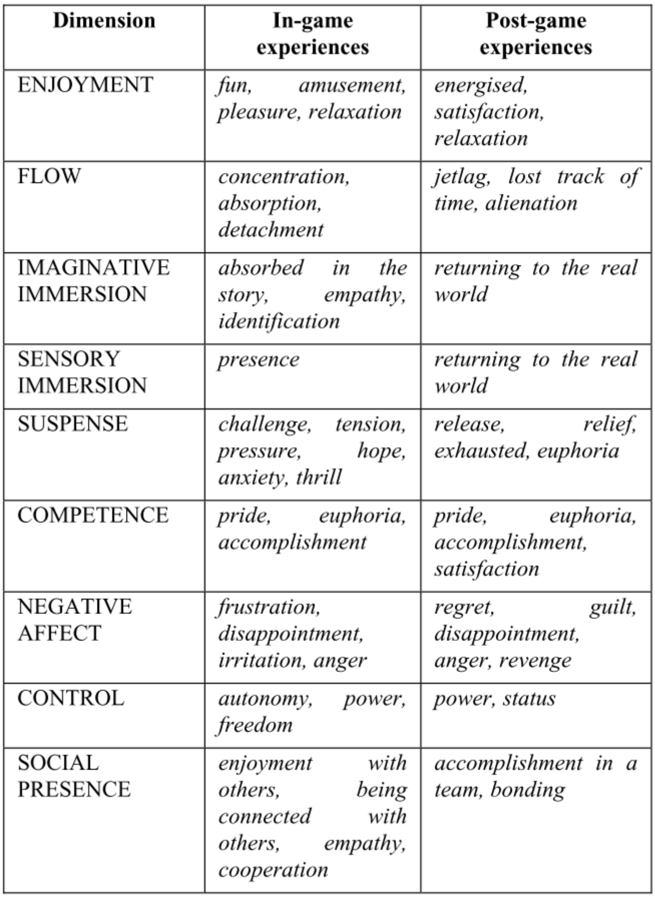
Behavioral Measures

To create a classifier to recognize emotions it has to be trained first. Therefore there has to be a method for label the physiological samples with the emotion felt. In the section Physiological measures, a lot of the studies use the Game Experience Questionnaire or the Self-Assessment Manikin to establish the relation between a person's self-reported emotions and the physiological reaction of the same person. These two measures will be described below.

5.1 GEQ

GEQ or Game Experience Questionnaire is a questionnaire created by Poels, Kort and Ijsselsteijn [PKI07]. The GEQ was created because of the underrepresentation of game experience in academic circles. The GEQ provides a common framework for measuring game experiences.

The GEQ was created by using focus groups that were tasked to individually reflect on game experiences before a group discussion was held. Five game researchers were also invited to perform the same procedure as the focus groups. After the focus groups testing the results were gathered and Poels et al. combined them to form nine different game experience dimensions. The dimensions are shown in Figure 5.1.



Dimension	In-game experiences	Post-game experiences
ENJOYMENT	*fun, amusement, pleasure, relaxation*	*energised, satisfaction, relaxation*
FLOW	*concentration, absorption, detachment*	*jetlag, lost track of time, alienation*
IMAGINATIVE IMMERSION	*absorbed in the story, empathy, identification*	*returning to the real world*
SENSORY IMMERSION	*presence*	*returning to the real world*
SUSPENSE	*challenge, tension, pressure, hope, anxiety, thrill*	*release, relief, exhausted, euphoria*
COMPETENCE	*pride, euphoria, accomplishment*	*pride, euphoria, accomplishment, satisfaction*
NEGATIVE AFFECT	*frustration, disappointment, irritation, anger*	*regret, guilt, disappointment, anger, revenge*
CONTROL	*autonomy, power, freedom*	*power, status*
SOCIAL PRESENCE	*enjoyment with others, being connected with others, empathy, cooperation*	*accomplishment in a team, bonding*

Figure 5.1: Dimensions of game experience.

The study Drachen 10 [DN10] uses some of the dimensions to find correlates between game experience and GSR and HR and shows that elements of game experience does correlate with physiological measures.

5.2 SAM

The SAM or Self-Assessment Manikin [BL94] is a pictorial instrument to assess pleasure or valence, arousal and dominance in response to an object or event. The SAM was created to address issues with the Semantic Differential Scale which is also used to assess valence, arousal and dominance. The issues with the Semantic Differential Scale is that contains 18 ratings on a 9-point scale which is cumbersome to use if a participant would have to rate each stimulus on the scale. Another issue is that it is verbal e.g. the participants are therefore asked to rate whether a stimulus is unhappy to happy on a 9-point scale. This can cause problems in non-English studies due to translation or with children who might not understand the nuances of different emotions.
The SAM directly assess valence, arousal and dominance using a 9-point scale. The SAM can be seen in Figure . Since there are only 5 images on each dimension the participant can also choose to rate a stimulus as being in between the images. This leads to a 9-point scale.

Of the two the GEQ is the more game-specific. The problem with GEQ is that it has 9 dimensions compared to the SAM with 3 dimensions. This means that the GEQ has 3 times the dimensions of the SAM and if the game has to ask the player to fill out the GEQ 6 times then the player would have had to answer 45 questions compared to just 15 with the SAM. Apart from being annoying, it is also immersion-breaking to have to answer a large questionnaire 6 times in a row.
Furthermore using the arousal/valence model the SAM is much more suited since it measure arousal and valence directly. It even does it pictorially which more or less sidesteps the issues with different languages and cultures that are present in the GEQ.
The recommendation for the current project is to use SAM because it is less intrusive and therefore less immersion-breaking, directly measure arousal and valence and does not rely on definitions of words and concepts.

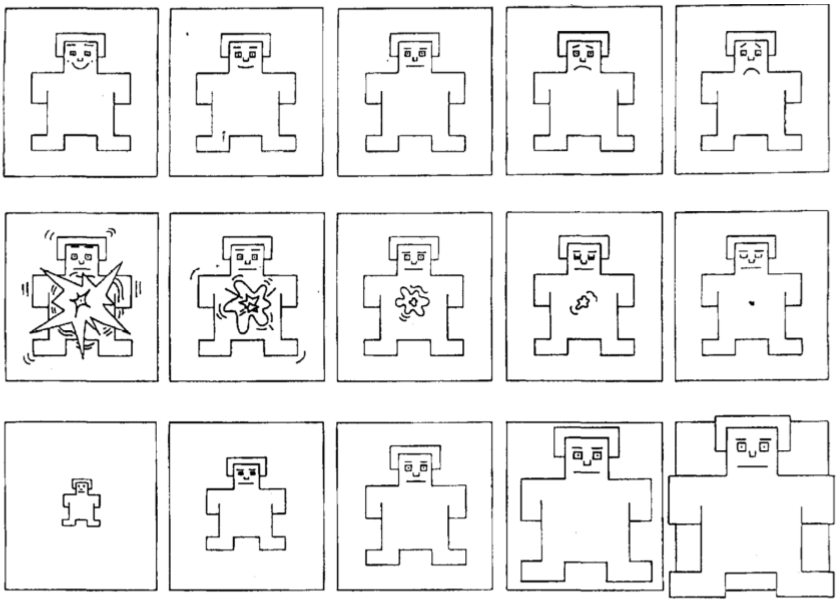


Figure 5.2: The SAM. The top row is valence, middle is arousal and bottom is dominance.

Chapter 6

Conclusion

The computer game industry has grown tremendously in the past years. The reason is that games provide people with entertainment and a sense of escapism. Most games are created to be fun although not all are. A system was proposed in the beginning of the report which could help improve the fun in games. This system is to measure the player's mental state and improve the game played to optimize the fun for the player. This system is composed of several parts. The first part is to establish how emotions are represented in the brain. Here Hamann [Ham12] proposes a network representation where each emotion activates several parts of the brain instead of each emotion corresponding to one part of the brain. This however is best measured with fMRI which is not a portable technology and therefore not suitable. Left is the dimensional theory where all emotions exist on a two dimensional map. This representation is chosen since the two dimensions, arousal and valence, can be measured with heart rate, facial muscles or skin conductance. These measures receive stronger signals than an EEG which could be used to measure emotions according to the other theory where an area of the brain equals an emotion.

Now that a view of emotions has been established as being dimensional, a clear view of fun in games must also be established. Fun in games can be compared to being in flow. The psychology of flow is a concept coined by Csikzentmihaly and defines a flow state as a state where a person is in complete focus of the task and feels a loss of sense of time, merging of awareness and action and a loss of self-consciousness. Csikzentmihaly has furthermore plotted flow in an abil-

ity/challenge map where great ability and low challenge will lead to boredom, great challenge and low ability will result in anxiety and equal challenge and ability will lead the player to be in a flow state.
Fun in a game can therefore be induced by leading the player into a flow state. Salen and Zimmerman [SZ04] discusses Dynamic Difficulty Adjustment (DDA) as a technique to dynamically regulate the level of challenge posed to a game player. A study by Hunicke [Hun05] and the use of DDA in commercial games show that DDA is a tool which can be used.
With the dimensional view of emotions and the concept of flow and DDA the emotions of the player has to be measured. Two methods are presented. One with EEG and measurement on the electrical signals of the brain and one with heart rate sensors, galvanic skin response sensors and other sensors which does not measure on the brain.
Comparing the results the EEG clearly lose to other physiological measures where EEG on average fared 13 percentage points better than chance level while other physiological measures were 44.4 percentage points better than chance level. Dissecting the performance of the other physiological measures show that galvanic skin response, heart rate and temperature were used by most studies and support vector machines were the most commonly used method for classifying the signals into emotions. This suggest that measuring galvanic skin response, heart rate and temperature and using support vector machines to classify emotions is the preferred method.
The measuring of emotions has to be classified using support vector machines. Since every person shows different patterns when reacting to events the game must include the training of the support vector machines. This can be done in the starting phase of the game and typically by asking the players how they feel. Two self-report schemes are useful. One is the game experience questionnaire which is a 9-dimensional scale where the dimensions relate to game experience and the other is the self-assessment manikin which is a pictorial 3-dimensional scale and relates to arousal, valence and dominance. The self-assessment manikin is preferred since it is faster to complete, is pictorial and completely relates to the chosen view of emotions as being on an arousal/valence map.

In the end the system is composed of physiological sensors that measure heart rate, galvanic skin response and temperature. The system will ship with a classifier that has already been trained. The training consists of players that will fill out self-assessment manikins while playing a game that will evoke different emotions. This data plus the signals from the sensors are used to train a support vector machine that can classify emotions based on arousal/valence. Based on studies the results should be around 44.4 percentage points better than chance level. When the game is played by a normal consumer the system will continuously measure the player emotions and if the player falls outside the flow zone the dynamic difficulty adjustment system can vary the challenge so that the player can reenter the flow zone and ultimately provide a fun and engaging game experience for the player.

Appendix A

Index of Terms

Arousal - Arousal is a measure of how alert a person is and in the valence/arousal dimensional emotion theory it ranges from calm to excited.
BVP - Blood Volume Pulse is a method to measure the blood flow using infrared light.
DDA - Dynamic Difficulty Adjustment. A technique to dynamically alter the difficulty in games.
Dominance - Dominance is an additional dimension that can be used with the valence/arousal dimensional emotion theory. It ranges from submissive to dominant.
EEG - Electroencephalography is the recording of electrical activity on the scalp.
EMG - Electromyography is the recording of electrical activity produced by muscles.
EOG - Electrooculography is a technique to measure the electricity used to move the retina. In reality it provides a measure on how the eye has moved.
Flow - A concept coined by Mihaly Csikszentmihalyi to represent a state of complete and energized focus in an activity with a high level of enjoyment and fulfillment.
GEQ - Game Experience Questionnaire. A questionnaire created by Poels, Kort and Ijsselsteijn [PKI07]. The questionnaire contains seven dimensions of gameplay: enjoyment, flow, imaginative immersion, sensory immersion, suspense, competence, negative affect, control and social presence.

GSR - Galvanic Skin Response. Other names include Skin Conductance Level and electrodermal activity. GSR is a technique to measure the conductivity of the skin through the sweat level of the skin.
HR - Heart Rate. Heart rate can be measured through a heart rate monitor but also with BVP or other methods.
IADS - International Affective Digital Sounds. A database compiled by Bradley and Lang containing sounds that are rated with arousal and valence values.
IAPS - International Affective Picture System. A database compiled by Bradley and Lang containing pictures that are rated with arousal and valence values.
SAM- Self-Assessment Manikin. A non-verbal pictorial assessment technique that directly measures the pleasure, arousal and dominance associated with a person's affective reaction to stimuli.
Valence - Valence is a measure of how pleasurable an event is to a person and in the valence/arousal dimensional emotion theory it ranges from highly positive to highly negative.

Bibliography

[BL94] MM Bradley and PJ Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental . . .*, 25(I), 1994.

[Bos06] DO Bos. EEG-based emotion recognition: The influence of visual and auditory stimuli. *Emotion*, 2006.

[Che07] Jenova Chen. Flow in games (and everything else). *Communications of the ACM*, 50(4):31, April 2007.

[CKGP06] Guillaume Chanel, Julien Kronegg, Didier Grandjean, and Thierry Pun. Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals. *. . . , classification and security*, pages 530–537, 2006.

[CRBP08] Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun. Boredom, engagement and anxiety as indicators for adaptation to difficulty in games. *Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era - MindTrek '08*, page 13, 2008.

[DN10] Anders Drachen and LE Nacke. Correlation between heart rate, electrodermal activity and player experience in first-person shooter games. *. . . on Video Games*, 2010.

[ESA12] ESA. Essential facts about the computer and video game industry, 2012.

[Gau12] John Gaudiosi. New Reports Forecast Global Video Game Industry Will Reach $82 Billion By 2017, 2012.

[Ham12] Stephan Hamann. Mapping discrete and dimensional emotions onto the brain: controversies and consensus. *Trends in cognitive sciences*, 16(9):458–66, September 2012.

[Hic11] Jon Hicks. Bethesda confirms no dynamic difficulty in Skyrim, 2011.

[Hun05] Robin Hunicke. The case for dynamic difficulty adjustment in games. *Proceedings of the 2005 ACM SIGCHI International . . .* , pages 429–433, 2005.

[KBK04] K H Kim, S W Bang, and S R Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical & biological engineering & computing*, 42(3):419–27, May 2004.

[LN04] Christine Læ titia Lisetti and Fatma Nasoz. Using Noninvasive Wearable Computers to Recognize Human Emotions from Physiological Signals. *EURASIP Journal on Advances in Signal Processing*, 2004(11):1672–1687, 2004.

[MIC06] RL Mandryk, KM Inkpen, and TW Calvert. Using psychophysiological techniques to measure user experience with entertainment technologies. *. . . & Information Technology*, 2005, 2006.

[MW13] Merriam-Webster. Merriam-Webster, 2013.

[PKI07] Karolien Poels, Yvonne De Kort, and W Ijsselsteijn. It is always a lot of fun!: exploring dimensions of digital game experience using focus group methodology. *. . . of the 2007 conference on Future Play*, pages 83–89, 2007.

[R* 12] R* K. Rockstar Game Tips: Difficulty Settings - Getting Started in Max Payne 3, 2012.

[RMP13] Boris Reuderink, C Mühl, and M Poel. Valence, arousal and dominance in the EEG during game play. *International Journal of Autonomous . . .* , 2013.

[Sal09] Adam Saltsman. Game Changers: Dynamic Difficulty, 2009.

[SFKP02] Jocelyn Scheirer, Raul Fernandez, Jonathan Klein, and Rosalind W Picard. Frustrating the user on purpose: a step toward building an affective computer. *Interacting with Computers*, 14(2):93–118, February 2002.

[Sie06] Joe Siegler. Prey Weekly Development Update #7, 2006.

[SZ04] Katie Salen and Eric Zimmerman. *Rules of Play: game design fundamentals.* MIT Press, 2004.

[Tol08] Josh Tolentino. Good Idea, Bad Idea: Dynamic Difficulty Adjustment, 2008.

[Uni11] University of Florida. THE CENTER FOR THE STUDY OF EMOTION AND ATTENTION, 2011.

Appendix B

Experiment 2 results

Participant	Heart rate DDA	Performance DDA
1	4 ± 1	4.33 ± 1.53
2	3.67 ± 1.15	5 ± 1
3	5 ± 0	3.67 ± 0.58
4	2.67 ± 1.53	1.33 ± 0.58
5	3.67 ± 0.58	3 ± 1.73
6	3.33 ± 1.15	2.33 ± 0.58
7	3.67 ± 0.58	4.33 ± 1.53
8	3.33 ± 0.58	3 ± 0
9	3 ± 1	1.33 ± 0.58

Table B.1: Comparison of valence ratings for heart rate based DDA and performance-based DDA

Participant	Heart rate DDA	Performance DDA
1	3.67 ± 1.15	4.33 ± 2.31
2	3.33 ± 1	5 ± 1
3	5.67 ± 1.15	4 ± 1
4	1.33 ± 0.58	2 ± 0
5	4.67 ± 2.08	3.67 ± 1.15
6	3 ± 2.65	5 ± 1
7	5.67 ± 0.58	4.33 ± 1.53
8	4.67 ± 0.58	3.33 ± 0.58
9	6.67 ± 0.58	5.33 ± 0.58

Table B.2: Comparison of arousal ratings for heart rate based DDA and performance-based DDA

Participant	Heart rate DDA	Performance DDA
1	5.67 ± 0.58	5.33 ± 1.53
2	3.33 ± 0.58	2.33 ± 0.58
3	6 ± 1	4.67 ± 1.53
4	2.67 ± 2.08	4.33 ± 0.58
5	6 ± 0.58	5.33 ± 0.58
6	5.33 ± 1.53	5 ± 1.53
7	5 ± 1	4.67 ± 0.58
8	4.67 ± 0.58	3.33 ± 0.58
9	6.67 ± 0.58	5.33 ± 0.58

Table B.3: Comparison of challenge ratings for heart rate based DDA and performance-based DDA

Participant	Heart rate DDA	Performance DDA
1	0.33 ± 0.58	0.33 ± 0.58
2	0.67 ± 0.58	1.67 ± 0.58
3	0 ± 0	0 ± 0
4	1 ± 1	1 ± 0
5	0.33 ± 0.58	1 ± 0
6	1.33 ± 0.58	0.67 ± 0.58
7	1 ± 0	0.67 ± 0.58
8	0.67 ± 0.58	1.33 ± 0.58
9	0.67 ± 0.58	0.33 ± 0.58

Table B.4: Comparison of valence ratings for heart rate based DDA and performance-based DDA

Appendix C

List of terms

Arousal - Arousal is a measure of how alert a person is and in the valence/arousal dimensional emotion theory it ranges from calm to excited.
BVP - Blood Volume Pulse is a method to measure the blood flow using infrared light.
DDA - Dynamic Difficulty Adjustment. A technique to dynamically alter the difficulty in games.
EEG - Electroencephalography is the recording of electrical activity on the scalp.
EMG - Electromyography is the recording of electrical activity produced by muscles.
Fast drop - In Tetris a fast drop is a move that accelerates the speed that a block is moving down with.
Flow - A concept coined by Mihaly Csikszentmihalyi to represent a state of complete and energized focus in an activity with a high level of enjoyment and fulfillment.
GSR - Galvanic Skin Response. Other names include Skin Conductance Level and electrodermal activity. GSR is a technique to measure the conductivity of the skin through the sweat level of the skin.
Hard drop - In Tetris a hard drop is a move that immediately moves a block down.
HR - Heart Rate. Heart rate can be measured through a heart rate monitor but also with BVP or other methods.

IADS - International Affective Digital Sounds. A database compiled by Bradley and Lang containing sounds that are rated with arousal and valence values.
IAPS - International Affective Picture System. A database compiled by Bradley and Lang containing pictures that are rated with arousal and valence values.
SAM- Self-Assessment Manikin. A non-verbal pictorial assessment technique that directly measures the pleasure, arousal and dominance associated with a person's affective reaction to stimuli.
Valence - Valence is a measure of how pleasurable an event is to a person and in the valence/arousal dimensional emotion theory it ranges from highly positive to highly negative.

Bibliography

[Amb11] Mike Ambinder. Biofeedback in gameplay: how Valve measures physiology to enhance gaming experience. *Game Developers Conference*, 2011.

[BL94] MM Bradley and PJ Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental . . .* , 25(I), 1994.

[Bos06] DO Bos. EEG-based emotion recognition: The influence of visual and auditory stimuli. *Emotion*, 2006.

[Che07] Jenova Chen. Flow in games (and everything else). *Communications of the ACM*, 50(4):31, April 2007.

[CKGP06] Guillaume Chanel, Julien Kronegg, Didier Grandjean, and Thierry Pun. Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals. *. . . , classification and security*, pages 530–537, 2006.

[CRBP08] Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun. Boredom, engagement and anxiety as indicators for adaptation to difficulty in games. *Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era - MindTrek '08*, page 13, 2008.

[ESA12] ESA. Essential facts about the computer and video game industry, 2012.

[FC86] AJ Fridlund and JT Cacioppo. Guidelines for human electromyographic research. *Psychophysiology*, 1986.

[HC04] Robin Hunicke and Vernell Chapman. AI for dynamic difficulty adjustment in games. *Challenges in Game Artificial Intelligence AAAI ...*, 2004.

[Hic11] Jon Hicks. Bethesda confirms no dynamic difficulty in Skyrim, 2011.

[Joh09] Bobbie Johnson. How Tetris conquered the world, block by block, 2009.

[Juu10] Jesper Juul. In search of Lost Time: on Game Goals and Failure Costs, 2010.

[KBK04] K H Kim, S W Bang, and S R Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical & biological engineering & computing*, 42(3):419–27, May 2004.

[Kos05] Ralph Koster. *Theory of fun for game design.* Paraglyph Press, 1 edition, 2005.

[LASC09] Changchun Liu, Pramila Agrawal, Nilanjan Sarkar, and Shuo Chen. Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback. *... Journal of Human–Computer ...*, 25(March 2013):37–41, 2009.

[LeB99] Marc LeBlanc. Feedback Systems and the Dramatic Structure of Competition, 1999.

[LN04] Christine Læ titia Lisetti and Fatma Nasoz. Using Noninvasive Wearable Computers to Recognize Human Emotions from Physiological Signals. *EURASIP Journal on Advances in Signal Processing*, 2004(11):1672–1687, 2004.

[McL13] Martyn McLaughlin. New GTA V release tipped to rake in £1bn in sales, 2013.

[Ngu] Du Nguyen. The use of physiological measurements to assess user involvement in computer games. •.

[Ols09] Scott Olstad. 25 Years of Tetris: From Russia With Fun!, 2009.

[R* 12] R* K. Rockstar Game Tips: Difficulty Settings - Getting Started in Max Payne 3, 2012.

[Ran05] Pramila Rani. Maintaining optimal challenge in computer games through real-time physiological feedback mechanical engineering. *Pulse*, 2005.

[RLSV06] Pramila Rani, Changchun Liu, Nilanjan Sarkar, and Eric Vanman. An empirical study of machine learning techniques for affect recognition in human–robot interaction. *Pattern Analysis and Applications*, 9(1):58–69, April 2006.

[RMP13] Boris Reuderink, C Mühl, and M Poel. Valence, arousal and dominance in the EEG during game play. *International Journal of Autonomous . . .*, 2013.

[Sal09] Adam Saltsman. Game Changers: Dynamic Difficulty, 2009.

[SFKP02] Jocelyn Scheirer, Raul Fernandez, Jonathan Klein, and Rosalind W Picard. Frustrating the user on purpose: a step toward building an affective computer. *Interacting with Computers*, 14(2):93–118, February 2002.

[Sie06] Joe Siegler. Prey Weekly Development Update #7, 2006.

[Sma13] Smartviking. Matris, 2013.

[SW05] Penelope Sweetser and Peta Wyeth. GameFlow: a model for evaluating player enjoyment in games. *Computers in Entertainment (CIE)*, 3(3):1–24, 2005.

[SZ04] Katie Salen and Eric Zimmerman. *Rules of Play: Game Design Fundamentals*. MIT Press, 2004.

[Tol08] Josh Tolentino. Good Idea, Bad Idea: Dynamic Difficulty Adjustment, 2008.