

# Simple Games – Complex Emotions: Automated Affect Detection Using Physiological Signals

Thomas Friedrichs<sup>1</sup>, Carolin Zschippig<sup>2</sup>, Marc Herrlich<sup>3</sup>,  
Benjamin Walther-Franks<sup>3</sup>, Rainer Malaka<sup>3</sup>, and Kerstin Schill<sup>2</sup>

<sup>1</sup> OFFIS – Institute for Information Technology, Oldenburg, Germany

<sup>2</sup> Cognitive Neuroinformatics, University of Bremen, Germany

<sup>3</sup> Digital Media Lab, TZI, University of Bremen, Germany

**Abstract.** Understanding the impact of interaction mechanics on the user’s emotional state can aid in shaping the user experience. For eliciting the emotional state of a user, designers and researchers typically employ subjective or expert assessment. Yet these methods are typically applied after the user has finished the interaction, causing a delay between stimulus and assessment. Physiological measures potentially offer more reliable indication of a user’s affective state in real-time. We present an experiment to increase our understanding of the relation of certain stimuli and valence of induced emotions in games. For this we designed a simple game to induce negative and positive emotions in the player. The results show a high correspondence between our classification of participants’ physiological signals and subjective assessment. However, creating a clear causality between game elements and emotions is a daunting task, and our designs offer room for improvement.

**Keywords:** Objective game evaluation, Psycho-physiology, Affective gaming, Valence detection.

## 1 Introduction

The role of emotions in human-computer interaction has received increased attention over the last years, and the user’s emotions are nowadays recognized as an important part of the overall user experience. While the emotional impact has always been central to entertainment applications it is no longer limited to this area, but also considered important in areas such as business applications.

Measuring the user’s emotional state and the impact of an interaction experience on it still relies mainly on subjective feedback from participants in user studies or observation and classification by experts. Questionnaires and similar tools are mostly suited for either pre- or post-experience assessment, but not for fine-grained real-time measurement or an online adaptation of the interaction. For example, when asking players after a game they might have difficulties remembering certain situations early in the game or certain events might overshadow the impact of others. Expert-based assessments are limited in the extent

and detail in which they can detect emotions, and require sophisticated alignment between assessors, as well as significant logistical overhead.

Psycho-physiological measurements of physical reactions can potentially allow objective, real-time assessment of the emotional state of users, thus enabling researchers and designers to make direct connections between design changes and the emotional impact. Once these connections are understood and formalized, the user experience can even be modified by automatically adapting to the user. With this goal in mind, the choice on which physiological parameters are to be measured becomes dependant on the ease of use of necessary sensors. It might not be practical to depend on the user wearing an electroencephalograph (EEG) cap or stick electrodes on facial muscles when playing a casual game on a mobile device, possibly in public.

We present further progress towards real-time, fine-grained measurement and classification of emotional valence in human-computer interaction without a prior calibration of the system to a specific player. In this we focus on video games, since emotions are particularly relevant to the gaming experience. We designed a game aimed at inducing positive and negative emotional states in the player at short but well defined intervals. Using this we observed and classified psycho-physiological data unobtrusively measured by electrocardiography (ECG) and electrodermal activity (EDA). We were able to successfully classify positive and negative emotional states according to the self-assessment of the participants across conditions by using psycho-physiological measures without prior knowledge on individual player reactions. We were less successful in inducing positive/negative emotional states in a controlled way with our game design.

## 2 Related Work

For a general overview of the recent work in Affective Computing, we point to the comprehensive review paper by Calvo [2] and the recently published *Oxford Handbook of Affective Computing* [3]. In addition to the studies mentioned in those publications, there has been a growing number of computer games that are either designed specifically as an affect manipulation tool or aim to utilize the affective state measured through physiological signals. Kivikangas et al. did an extensive literature review on psycho-physiological methods in game research in 2010 [8]. Since a comprehensive overview of the recent developments in the field is outside the scope of this paper, we limit ourselves to two highlights in order to illustrate the current state-of-the-art in affect-adaptive gaming research. Nogueira et al. have proposed the “Emotion Engine biofeedback loop system” to study and manipulate the affective player experience [11]. The Engine uses EDA, cardiovascular measures and EMG to infer the player’s emotional state and can achieve 78% classification accuracy for valence if the player undergoes a personal calibration process. Chanel et al. studied the affective reaction of players to the game TETRIS via an EEG [4] aiming at implementing a dynamic adaptation of game difficulty. They succeeded in classifying valence with an accuracy of about 60%, leaving adapting the game as a future perspective.

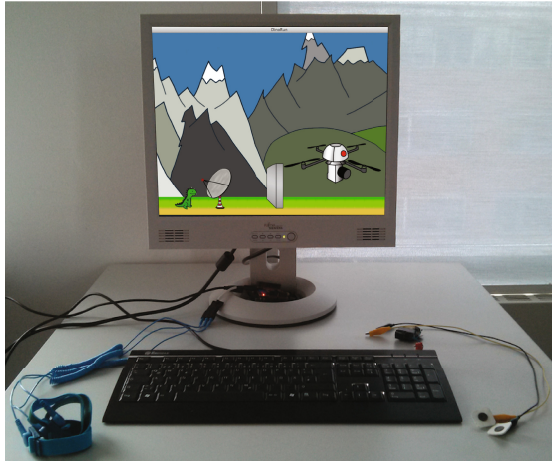
### 3 Designing a Game for Measuring Valence

We specifically designed and implemented *Dino Run* as a tool for inducing positive or negative emotional valence. The tool should have clear influencing variables to allow us to manipulate the emotional impact dynamically. Measurement and classification of the emotional state is realized using psycho-physiological measures. *Dino Run* is a simple casual game, its core mechanics modelled after various successful games from the mobile game market. Great care was taken in game design to maximize the emotional impact and reduce noise in the measurements induced by complex and only intricately traceable cognitive processes. The main goal of *Dino Run* is to steer a little dinosaur through an obstacle course by jumping or ducking (cf. figure 1). The game is a typical side-scrolling game, with only jumping and ducking as vertical motion.

The goal of a tool for valence induction and measurement imposed two constraints that directed the design process. It had to be appealing to a broad audience, and the duration of gaming sessions should be kept to a minimum. Both constraints are satisfied by casual games, especially the so-called “one button” games popular on mobile devices. They are played by a large and diverse audience and the threshold to new players is very low. It can be assumed that many users are already familiar with such games, reducing the potential emotional impact of learning. Furthermore, the limited controls and the fixed set of game mechanics allow better control of game parameters in order to induce positive or negative emotional states.

In contrast to standard entertainment games, for our study design it was crucial to eliminate any redundant mechanics. In most games many different elements such as puzzles, collectables, power-ups, enemies, score systems are used in conjunction to make the game *fun*. Yet it is not apparent how these elements exactly affect players of different target groups. For our purposes we require the positive and negative conditions to be as symmetric and comparable as possible. We therefore pre-tested different mechanics before creating the final game design, including specific visual feedback, and special negative/positive items. The final game design included two mechanisms for inducing either positive or negative emotions based on the core locomotion mechanic of the game. They were designed to create a noticeable emotional change without the players consciously recognizing the change in the underlying game parameters. The first is an *adjustable collision detection and jumping force* of the character. Pre-tests confirmed that subtle changes regarding the hitbox sizes lead to increased collisions and negative performance of the players, which should impact their emotional state negatively, while not being noticed by the players. To reinforce the positive/negative affect, *positive and negative auditive feedback* was included in the game in the form of fanfares and buzzer sounds, respectively. Pre-tests confirmed this to work better than additional visual feedback, as this can easily be overlooked by players. The auditive channel is independent of the visual channel and has been successfully used to affect the emotional state of users [7].

We chose cardiac and electrodermal activity as psycho-physiological signals for valence detection, which can be unobtrusively measured by electrocardiography



**Fig. 1.** Experimental setup with ECG and EDA sensors

(ECG) and skin conductance measurement (SC). Heart rate variability (HRV) has been shown to be significant for valence detection [9] and in combination with EDA carries information about the respiration pattern, which has been shown to be influenced by valence [6].

## 4 User Study

For the study, students were recruited from the university campus. The participants (26 males and 21 females) were from different fields, backgrounds and had diverse gaming habits. Participants played the game while their electrodermal activity (EDA, i.e. skin conductance) and an electrocardiogram (ECG) were recorded. The ECG was done using an Olimex SHIELD-EKG-EMG on an Arduino Mega 2560, the Bluetooth-operated EDA sensor was a custom design [12]. After introducing the overall procedure, the EDA and ECG sensors were applied to fingers and lower arms, respectively. Participants were given instructions to avoid any obstacles in the game. Prior to beginning data acquisition, the participants had the opportunity to practice the controls in a basic test level that was free of any manipulations or feedback. Each game was internally divided into three phases: a neutral phase without any manipulation or feedback, the game phase  $E^1$  where the players played either a positive ( $E_p^1$ ) or negative game condition ( $E_n^1$ ), and phase  $E^2$ , which could be positive or negative as well ( $E_p^2$  or  $E_n^2$ ). The phases lasted 60s, 150s and 150s. They were not explicitly communicated to the player, nor were there obvious indicators in game. From this set-up with full permutation, four participant groups result: people playing only the positive game condition ( $E_p^1 \& E_p^2$ ), people playing the positive, then the negative condition ( $E_p^1 \& E_n^2$ ), people playing negative, then positive ( $E_n^1 \& E_p^2$ ) and people playing negative only ( $E_n^1 \& E_n^2$ ). All participants were distributed randomly across the groups. Four groups were chosen to be able to compare

a) changing conditions within the game (positive to negative and vice versa) and b) changing reference conditions across trials (only positive or negative).

After finishing the task, the participants were asked to fill out an online questionnaire containing questions from the *Game Experience Questionnaire (GEQ)* [1] and a 9-point *Self Assessment Mannequin (SAM)* scale [10]. These questionnaires were chosen as they are based on two different models of emotion. There is currently no consensus in the affective systems research community on the most performant theoretical model of emotion in the human-machine interaction context, other than a strong tendency to use dimensional rather than discrete emotional models[2]. We support dimensional models, but view it as an open question to define the details of the dimensional space. Consequently we chose to attain the subjective ratings of our participants through a questionnaire based on a model with two axes for valence, labelled positive and negative affect (GEQ) and another questionnaire modelling valence on a single axis with the poles positive and negative (SAM).

## 5 Results

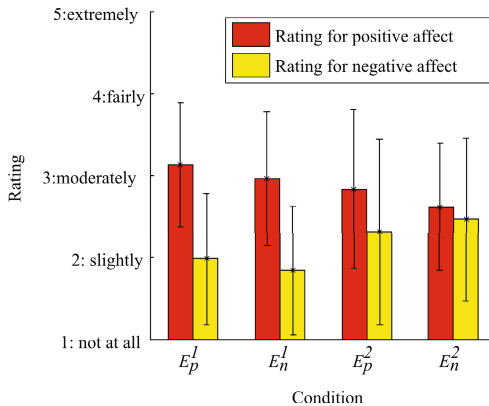
For the evaluation, data of 47 different participants was used. 38 data sets comprise of two conditions per person and 9 data sets comprise of one condition per person. Due to noise in the measurement, some data sets had to be removed from the analysis because the sensors failed randomly during the trials. However, the identification of the corrupted data was feasible because it was clearly distinguishable from the non-corrupted data (completely distorted signal).

The physiological data was preprocessed with digital filtering algorithms before a set of 13 features was extracted. The Biosig-toolbox [14] provided the algorithms to extract the RR-intervals in the ECG signal. The feature set comprised features derived by time domain methods as well as frequency domain methods, which have been shown to have psychophysiological significance in various publications [2]. The features chosen were: the standard deviation of the tonic and phasic component of the skin conductance, the slope of a linear approximation of the tonic component, the mean, standard deviation and root mean square of the RR-intervals, the root mean square and standard deviation of differences in interval lengths, as well as the power and normalized power in the high (0,15-0,4Hz) and low (0,04-0,15Hz) frequency range, plus the ratio of low to high frequency power. To improve the frequency resolution of the power density estimation, we used an autoregressive model with the model order of 16, as suggested in the literature [13]. A principal component analysis of the feature space reduced the number of features to 10, which explain 99% of the variance. This feature matrix was evaluated using a support vector machine (SVM), in order to infer the user's affective state from the physiological data during gameplay. The SVM was implemented using the libSVM library [5] in MATLAB.

First, we checked our working hypothesis that the game induces a positive affective state in the player during condition  $E_p$  and a negative affective state during condition  $E_n$  by training an SVM with the feature matrix and the conditions as labels. After a 10 fold cross validation with a training data/test data

ratio of 9:1, the accuracy on training data was 60%. For the prediction the SVM did not reach chance level. This low prediction accuracy led us to the interpretation that the game might not have induced the expected valence during play, despite our careful design and pretesting.

We consequently tested our hypothesis by evaluating the subjective ratings reported by the players by means of the GEQ and by SAM. We specifically looked at GEQ items that rate the negative and positive affect induced by gameplay. Comparing the ratings for positive affect to those of negative affect revealed that for both game conditions the users rate positive affect higher than negative. Looking at the changes over time, we found the positive affect rating declining from first to second condition while the negative affect rating inclined. This was true for all games, when the condition was meant to induce a negative affect but also when it was meant to induce a positive affect. Figure 2 shows the mean and standard deviation of the GEQ ratings. We then performed a  $\chi^2$ -test (hypothesis of independence and a normal distribution with  $\mu = \text{mean}(\text{GEQ})$ ) to see in more detail whether the valence of the condition has any significant influence on the GEQ votes, or if the votes are independent of the game condition. The  $\chi^2$  values of the votes for  $E^1$  and  $E^2$  are:  $\chi^2_{E^1, p.a.r.} = 5,63$ ,  $\chi^2_{E^1, n.a.r.} = 12,01$ ,  $\chi^2_{E^2, p.a.r.} = 6,70$  and  $\chi^2_{E^2, n.a.r.} = 8,76$  (p.a.r. = GEQ positive affect rating, n.a.r. = GEQ negative affect rating). We interpreted these results as follows: in the first minutes of gameplay ( $E^1$ ), playing the game induces positive affect in the players. If this condition is a negative one ( $E_n^1$ ), the GEQ ratings on negative affect are significant (on a significance level of 5%). For the second game condition ( $E^2$ ), both affect ratings depend upon the game condition on a significance level of 5%. While this indicates that the game's mechanisms



**Fig. 2.** Mean and standard deviation of the GEQ rating for positive and negative affect.

induce a discriminative player experience, the values also suggest that the game conditions are experienced significantly different depending on the time already played. Analysis of the SAM ratings showed these to be independent of the game

condition label ( $\chi_{E^1, SAM}^2 = 4.94$ ,  $\chi_{E^2, SAM}^2 = 6.98$ , both not significant on a 5% level). This led us to the hypothesis that a classifier trained with the subjective game experience obtained through SAM should produce better classification results, if the intended valence induction does indeed not match the objective overall experience of the player. The SVM trained with SAM labels grouped into  $SAM \geq 5$ : +1 and  $SAM < 5$ : -1 had a training accuracy of 75% and a test accuracy of 68.9%, supporting our hypothesis regarding the player experience.

## 6 Discussion and Future Work

The initial low performance of the classifier was substantially improved by using the SAM subjective rating of the game experience. Compared to literature values, a classification accuracy of almost 70% on data with very high subject variability is a promising result. Yet there is room for improvement. On the classification side, we will look into the temporal resolution of the features and take into account prior game events, in-game player actions and emotional stages. Also, we will look into alternative classification methods that allow for a more detailed resolution of the valence space, not just a binary positive/negative clustering. With these alternative methods, we are looking to classify on the basis of the two affect dimensions evaluated by the GEQ to see if this model will result in higher classification rates in our setting. A second point of discussion is the game design and its success in inducing the intended valence in a substantial intensity. The data obtained through the post-game questionnaires indicate that the reduction of game complexity may have been a key contributing factor to the increase in negative affect in condition  $E^2$ , since one dimension of the negative affect was attributed to increased boredom. Comparing our game to successful casual games, we identified the graphical design as one element to improve the game experience without introducing uncontrollable disturbance factors. Many side scrolling games change the visuals of the game, which contributes to the motivational aspect of curiosity without any impact on the game mechanics. As a next step, we will redesign the game in order to improve the affect induction and transfer our results towards the development of an adaptive game, manipulating the game parameters during gameplay to achieve a desired game experience.

## 7 Conclusion

We designed and implemented a simple video game as a controllable environment for the study of the player's emotional reaction during gameplay. In an empirical study with 47 participants, we mapped physiological data to the player's subjective game experience with an accuracy of approximately 70%. Our work demonstrates that automatic detection of affect valence using non-invasive physiological sensors is possible, and gives first insights into stimulating negative and positive emotional responses through game mechanics design. The results imply that the evaluation of a specific game design element or game mechanic can be facilitated. Thus, the choice of game mechanics and design elements for the

evocation of an intended specific emotion (e.g. does this scene really scare the player?) can be grounded on data which, in contrast to post-game questionnaires, have a more direct temporal mapping between stimulus and response.

## References

1. Brockmyer, J.H., Fox, C.M., Curtiss, K.A., McBroom, E., Burkhart, K.M., Pidruzny, J.N.: The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *J. of Experimental Social Psychology* 45(4), 624–634 (2009)
2. Calvo, R.A., D’Mello, S.: Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Trans. on Affective Computing* 1(1), 18–37 (2010)
3. Calvo, R.A., D’Mello, S., Gratch, J., Kappas, A.: *The Oxford Handbook of Affective Computing*. Oxford University Press (2014)
4. Chanel, G., Rebetz, C., Pun, T.: Emotion Assessment From Physiological Signals for Adaptation of Game Difficulty. *IEEE Trans. On Systems, Man, and Cybernetics* 41(6), 1052–1063 (2011)
5. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. *ACM Trans. on Intelligent Systems and Technology* 2, 27:1–27:27 (2011)
6. Ganglbauer, E., Schrammel, J., Deutsch, S.: Applying Psychophysiological Methods for Measuring User Experience: Possibilities, Challenges and Feasibility
7. Hébert, S., Béland, R., Dionne-Fournelle, O., Crête, M., Lupien, S.J.: Physiological stress response to video-game playing: the contribution of built-in music. *Life Sciences* 76(20), 2371–2380 (2005)
8. Kivikangas, J.M., Chanel, G., Cowley, B., Ekman, I., Salminen, M., Järvelä, S., Ravaja, N.: A review of the use of psychophysiological methods in game research. *J. of Gaming & Virtual Worlds* 3(3), 181–199 (2011)
9. McCraty, R., Atkinson, M., Tiller, W.A., Rein, G., Watkins, A.D.: The effects of emotions on short-term power spectrum analysis of heart rate variability. *The American Journal of Cardiology* 76(14), 1089–1093 (1995)
10. Morris, J.: Observations: SAM: the Self-Assessment Manikin; an efficient cross-cultural measurement of emotional response. *J. of Advertising Research* (December 1995)
11. Nogueira, P.A., Rodrigues, R., Nacke, L.E.: Guided Emotional State Regulation: Understanding and Shaping Players’ Affective Experiences in Digital Games. In: *AIIDE* (2013)
12. Rachuy, C., Budde, S., Schill, K.: Unobtrusive data retrieval for providing individual assistance in aal environments. In: *Int. Conf. on Health Informatics* (2011)
13. Rajendra Acharya, U., Paul Joseph, K., Kannathal, N., Lim, C.M., Suri, J.S.: Heart rate variability: a review. *Medical & Biological Engineering & Computing* 44(12), 1031–1051 (2006)
14. Vidaurre, C., Sander, T.H., Schlögl, A.: Biosig: The free and open source software library for biomedical signal processing. *Computational Intelligence and Neuroscience* (2011)