User evaluation of affective dynamic difficulty adjustment based on physiological deep learning*

Guillaume Chanel[1][0000−0002−6184−8924] and Phil Lopes[2][0000−0002−9567−5806]

[1] Computer Science Department, University of Geneva, Switzerland
Guillaume.Chanel@unige.ch
cvml.unige.ch
[2] Immersive Interaction Group, EPFL, Switzerland
phil.lopes@epfl.ch
iig.epfl.ch

Abstract. Challenging players is a fundamental element when designing games, but finding the perfect balance between a frustrating and boring experience can become a challenge in itself. This paper proposes the usage of deep data-driven affective models capable of detecting anxiety and boredom from raw human physiological signals to adapt the fluctuation of difficulty in a game of Tetris. The first phase of this work was to construct several emotion detection models for performance comparison, where the most accurate model achieved an average accuracy of 73.2%. A user evaluation was subsequently conducted on a total of 56 participants to validate the efficiency of the most accurate model obtained using two diverging difficulty adaptation types. One method adapts difficulty according to raw values directly outputted by the affective model (ABS), while another compares the current value with the previous one to influence difficulty (REA). Results show that the model using the ABS adaptation was capable of effectively adjusting the overall difficulty based on a player's physiological state during a game of Tetris.

Keywords: Affective Computing, Machine Learning, Physiology, Player Modeling, Player Experience, Autonomous Difficulty Adaptation

1 Introduction

The accessibility of games is often a widely debated topic, where some argue that games should be accessible for the majority of the player-base to enjoy, while others suggest that games should be challenging but fair. Balancing difficulty in digital games has always been a challenging task for designers, where specific game-play elements are adapted in order to achieve an ideal balance for a wide varying player skill-sets. Traditionally players can select the amount of difficulty they want to face during play (e.g. easy, medium or hard difficulty). However, this solution still presents a few limitations, as some players may not directly fit into any of the categories (e.g. high difficulty still too easy, or easy difficulty still

* Co-financed by Innosuisse

too hard), or the game's actual difficulty progression during play may still be too steep leaving some players behind and frustrated, or even too slow making the game tedious.

A few games have attempted to address this problem by adjusting the difficulty of the game based on player performance. For example *Max Payne* (Remedy Entertainment, 2001) would slightly adjust the aim assist, provide additional ammunition and health kits when the player was having a hard or easy time to beat a particular level. The concept of a game reacting to a player's skill and progress is a fundamental idea of dynamic difficulty adjustment [10, 4, 33, 32], where the game attempts to manage mechanics, level layouts or the skill of non-playable opponents by using statistical game features to refine them towards the perfect challenge for the player.

Games are emotional experiences, they can induce a wide variety of emotions such as: frustration during challenging gameplay segments; anxiety in horror games; or even excitement. In fact, analyzing player reactions and emotions is of critical importance to evaluate games [27]. Thus, it makes sense for a game to also react to the players' emotions and not only to their performance. For this purpose, several researchers have proposed to develop models for automatic emotion recognition in games [7, 36, 20]. Unfortunately, the majority of research tends to be rooted specifically on the construction of accurate emotion recognition models, rarely addressing the question of game adaptation. In other words, a research question remain: how to employ emotion recognition models to adapt game difficulty and improve players' game experience?

Although this work mostly addresses emotion recognition for digital games, applications can be extended into other domains. Emotional models can allow systems to retain user engagement and attention by offering methods of detecting if a user is bored or frustrated, allowing the application to react accordingly and attempt to re-engage the user in some fashion.

This study attempts to answer this question by extending the work presented in [7]. It is based on models trained from physiological signals and which are capable of detecting anxiety and boredom during play. Given that most notable works using physiology to model affect in games have used skin conductance with a wide degree of success [21, 36, 25], this study applies deep learning to this modality for emotion recognition. The emotional information is then used to adapt the difficulty of a Tetris game. Two types of adaptation methods are explored, where one type directly uses the model decision, while the other compares the previous model output with the current one and then adjusts the difficulty accordingly. A user study was conducted exploring the viability of both approaches during a real-time gameplaying session.

2 Related Work

The following section provides literature of previous work exploring dynamic difficulty adjustment and affective modeling within the domain of digital games.

2.1 Dynamic Difficulty Adjustment

The concept of challenge and difficulty within the domain of digital games has always been a heavily debated topic, where several theories have been formulated. In particular the theory of flow [9], has been widely applied as both a design paradigm [16] and for Dynamic Difficulty Adjustment (DDA) systems [7]. DDA systems are capable of adjusting and eventually personalizing the difficulty during play. Thus, the ideal DDA system would still provide players with a degree of challenge, such that it is not overwhelmingly hard for the player but still challenging enough for the player to feel accomplished and engaged in the experience. According to the flow theory, a too high (respectively low) challenge should lead the player towards an emotional state of anxiety (respectively boredom). In [7] the concept of flow is used to balance difficulty by measuring players' anxiety and boredom: if the player is anxious (resp. bored) then reduce (resp. increase) difficulty.

One of the most common approaches towards designing DDA systems is through the measurement of performance metrics (e.g. number of deaths, damage received, damage done). One of the first academic example of DDA optimized adversarial agent strategy based on a function of challenge, by using an online evolutionary method [10]. Adversarial agents adapt their "ability" based on player's performance, which is calculated through specific heuristics derived from statistical game parameters (e.g. rate of missed or hit shots, games won and lost, time to finish a task). More recently, reinforcement Learning and evolutionary computation strategies for DDA were compared on a simplified racing game. A more data-driven approach was applied in [32], where the authors attempted to model player experience paradigms such as fun, frustration and challenge for the generation and personalization of *Super Mario Bros.* (Nintendo, 1985) levels.

As noted in [27], affective experiences are a fundamental part of digital games and should be one of the main evaluation factors during play-testing and game evaluation phases. Thus, it makes sense to approach the problem of difficulty management through the perspective of emotion, and player experience. This paper takes an affective computing approach exclusively, rather than using in-game performance metrics, by using physiological monitoring (i.e. Galvanic Skin Conductance) for the measurement of player affect.

2.2 Affective Modeling in Games

In the field of affective computing, games are often used as stimuli to induce players' emotional reactions which are then recorded and used to create emotion recognition models [21, 2]. The detection of an individual's emotions can be achieved through several means, such as facial expressions and vocal prosody [6], but also by using physiological signals [26, 15]. In order to target applications such as video games, a noticeable shift has been observed in the last decade towards assessing emotions in more natural and realistic situations [30],[3]. This shift has demonstrated that the majority of algorithms suffer from a drop of

performance when compared to the recognition of controlled and acted emotions. In addition, emotion recognition models can be trained to recognize any player's emotions (i.e user independent) or only those of the player for which the models were trained (i.e. user dependant) [26]. It is important to note that user independent models have the advantage that they do not necessitate any learning phase prior to play but they generally suffer from a significant drop in performance [3].

In the context of emotion recognition from physiological signals for video gaming, Rani et al. [29] proposed to classify three levels of intensity for different emotions, including anxiety, using several physiological signals measuring heart activity, muscle activity, electro-dermal activity and skin temperature. The best average accuracy obtained with this method was 86%. Physiological user-independent models for affect recognition in games were proposed in [7]. By combining peripheral physiological signals (electro-dermal activity, heart rate, skin temperature) with brain signals a performance of 63% was obtained to distinguish three emotions: boredom, engagement and anxiety. The current paper is a significant improvement of this work, which in particular now uses deep networks to improve performance and evaluates players perception of affective difficulty adaptation. A deep convolutional network was used in [22] to detect flow, boredom, and stress from heart rate and electrodermal activity during Tetris play (15.5 hours of data). The results demonstrate that user-indepedant models can achieve an accuracy around 70% when classifying two classes (stress vs. not stressed and flow vs. not flow). However the boredom state is the most difficult to detect with an accuracy of 57% when classifying boredom vs. other states. In addition, when flow was detected, the players obtained a better score than in the other emotional states showing that physiological flow can be an indicator of performance. Interestingly, a remote approach to physiology measurement have also been adopted to detect stress and boredom in games with an accuracy of 61.6% [5]. Remote physiological sensing consist of measuring physiological signals such as heart rate without using contact sensors [35] (e.g. a webcam is used to measure the slight changes of skin color produced by heart rate variability). The results also showed that despite of a small drop of performance compared to contact sensors, fusing physiological remote sensing with facial modalities was increasing emotion recognition performance compared to any model using a single modality.

Affective computing can bring substantial benefits to the domain of digital games, allowing designers and the game itself to construct personalized experiences for players based on predictions of their current emotional state. For example in the work of Lopes et al. [20], affective models were used to place sounds within a procedurally generated level to follow a specific emotional fluctuation for the player to experience. To our knowledge, emotion classifiers using physiological signals as inputs have been evaluated for DDA in two studies [19, 11]. The classification methods employed in [19] were the same as those presented in [29] but classifiers were retrained for the new players. The reported accuracy dropped to 78% but a user-study showed a significant improvement of player

experience compared to difficulty adjustment based on performance. Although this demonstrates the interest of using affective computing for the purpose of game adaption, the proposed model was user-dependent and required a one hour training session for each new player in order to tune the classifiers. In [11], the difficulty of a Tetris game was adapted depending on players' brain activity and using three adaptive strategies: conservative, moderate and liberal (ranging from the one presenting less adjustments to the one providing a lot of adaptations). In this experiment, players reported a higher alertness when playing with the conservative version compared to the liberal one, showing that adaptive systems should be designed with care and not overly adapt to the players physiology.

Even though a substantial amount of research exists for the construction of affective models, works that take these models and effectively integrates them within a game for user evaluation remain scarce. This study argues that although the conceptualization of affective models is an important process, it is also important to think and conceptualize what to do with the information obtained from such models and how to directly apply them within the applicative context. This work investigates methods for directly applying an emotion recognition model in games by observing both: the players actual perception of difficulty during play, and the actual behavior of said model during several play tests. Furthermore, this paper also explores two different adaptation methods which change difficulty according to either: the raw values obtained from the model (ABS); or by comparing the previous and current output to make a decision (REA).

3 Methodology

This section details the the experimental methodology utilized for the study of adaptive difficulty adjustment using affective modeling. In a first step, a model able to recognize player's emotions was developed. It followed the physiological data driven approach proposed in [12, 7]. However, instead of extracting features from the recorded signals we hereby propose an end-to-end learning solution based on the combination of deep convolutional [24] and Long Short Term Memory (LSTM) networks [13]. The proposed method was compared to the original method in [7] by using the same dataset. In a second step, the trained system was employed to test two difficulty adaptation strategies in real-time. These two strategies were tested and compared in a user study where the players experience was collected through self-reports.

3.1 Affective data

The data collection process used for this study is fully described in [7] and summarized below. The affective dataset was acquired by measuring the physiological activity of 20 participants playing a Tetris game for 6 sessions of 5 minutes each (10 hours of data). Each game session was played at a given difficulty level in order to elicit the following emotions: boredom (difficulty level lower than their skill), flow (difficulty level that matches their skill) and anxiety (difficulty

level higher than their skill). The players' skill was determined on previous play sessions. Each condition was played twice hence the 6 sessions. For this work only the boredom and anxiety emotional states are considered because: (*i*) the flow class was the most difficult to classify as reported in [7], (*ii*) by performing adaptation based on boredom and anxiety we expect that the difficulty level will oscillate around the ideal player difficulty. In addition, while [7] addressed the importance of fusing several physiological signals to improve performance, we preferred to adopt a user friendly interface which relies only on the most efficient physiological signal: electrodermal activity.

3.2 Electrodermal activity

The Electrodermal Activity (EDA) is a measure of the sympathetic nervous system, specifically the sweat gland activity. An increase of activity in sweat glands usually occurs when one is experiencing arousing emotions such as stress or surprise [17, 31]. In the original dataset, electrodermal activity was recorded using the Biosemi Active 2 system with a sampling rate of 256Hz which was later undersampled to 8Hz (EDA is a slow varying signal). The Biosemi Active 2 electrodes are difficult to equip and we therefore favored an open source solution for the user study. We used the sensors proposed in [1] and inspired from [28], which measured EDA (more precisely skin resistance) at a sampling rate of 8Hz. For both experiments the same positioning of electrodes was used: two electrodes positioned on the medial phalanges of the index and middle finger.

Signals of each session were segmented into 20 second windows with an overlap of 19 seconds. This specific window duration was chosen as a compromise between a fast adaptation time and the necessary duration required to reliably infer user emotions. This overlap also provides a high number of samples which are generally necessary for the training of deep networks. This windowing procedure led to a total of 21918 windows with a slight skew towards anxiety (48.7% of windows belonging to the anxiety class), as it was necessary to discard two sessions due to recording errors.

3.3 Baseline affective models

In order to create baseline models (i.e. models to which the deep approach will be compared), the most relevant EDA features suggested in [7] were selected to characterize both the Phasic and Tonic components of an EDA signal. The average EDA was computed to reflect the Tonic component. The percentage of samples which included an increase of the EDA signal was used to indicate the importance and duration of conductance peaks. Lastly, the total number of peaks within an EDA signal was used.

All features are computed for each time window as described in section 3.4, and concatenated into a feature vector. Due to an abundance of inter-participant and inter-session variability when monitoring physiological signals [37], it is standard practice to record participants during a resting period, and subsequently

subtract these measurements from the signal obtained during the actual activity. For this study, a resting period of 20 seconds was recorded before each play session. A rest feature vector was constructed based on the rest signals and subtracted from all the play feature vectors of the same session.

Results from the previous study [7] suggest that discriminant analysis classifiers are superior on the proposed dataset. Thus, this work compares the proposed end-to-end learning approach to that of a linear discriminant analysis (LDA) and a quadratic discriminant analysis (QDA). Although the linear approach might show better generalization on the test set, the quadratic method allows to define non-linear boundaries between classes as is the case of the proposed deep neural network.

3.4 Affective end-to-end learning

The advantage of using deep neural networks is that it offers the possibility to train models capable of identifying the most useful features of a signal and build a representation of the problem through intermediary layers [18]. In other words, instead of including expert knowledge to define and select relevant features used for classification, the raw signals themselves are fed directly to the network, building a discriminative internal representation of the data. For this work, the derivative was calculated for each window of a signal and then used as input for the networks, so as to reduce inter-participant variability which tends to occur due to varying data range (i.e. the general amount of sweat on the fingers). Contrarily to the baseline methods presented in section 3.3 it does not need to rely on the recording of a rest period.

The following deep architecture, inspired from [34] and stacking convolutional layers with a LSTM layer, is proposed to model EDA:

- 1D convolutional layer with 16 kernels of size 2, a stride of 1, no padding and a linear activation function;
- max-pooling layer with size 2;
- the succession of 3 1D convolutional layers (each with 8 kernels of size 2, a stride of 1, no padding and a linear activation function) and max-pooling layers of size 2;
- LSTM layer with 16 recurrent neurons and a softsign activation function;
- 1 neuron fully connected layer acting as a logistic regression (i.e. a sigmoid activation function with a loss measured using binary cross-entropy).

With this neural architecture the receptive field of a neuron in the last convolutional layer corresponds to 31 samples of the input signal. This implies that a sample outputted by a neuron of the last convolutional layer corresponds to approximately 3.9 second of the input signal. This duration and the corresponding architecture was specifically chosen as it is enough to detect an EDA peak. In other words, the convolutional layers are expected to act as peak detectors and characterizers. The LSTM takes these temporal features as inputs and memorizes their temporal occurrence. The LSTM layer can be seen as a peak counter with the capabilities of establishing a temporal relationship among them.

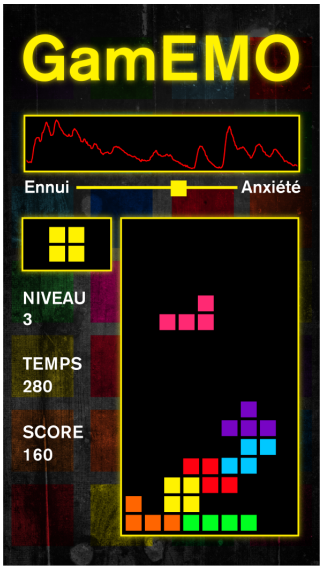

Fig. 1. Screenshot of the affective Tetris game. The upper window shows an example of an EDA signal with a duration of 70 seconds, which was recorded during the user study. Several galvanic responses (peaks) can be observed in the signal.

The network was implemented using Keras [8]. The decay rate was set to 10^{-4}, while the momentum to 0.5. The models were trained with 100 training epochs and a batch size of 2000. The performance of every classifier was evaluated using a leave-one participant out cross-validation (see 4.1). In addition, 70% of the training data was used to adjust the network weights while 30% of the data was used to control the validation error of each fold. Since all validation curves showed a plateau (even with a number of epoch higher than 100) it seemed that the model did not overfit the training data. The last trained model at epoch 100 was thus chosen as the model to be tested. Finally, in order to improve the performance of this model, we also performed ensemble learning by combining 20 neural networks with the same architecture. These 20 networks were trained on splits of the training set containing each 70% of the data randomly selected from the full training set. The decisions of these models were then averaged to take a final decision.

3.5 The Tetris Game

In Tetris, the difficulty is mainly controlled by the falling speed of the Tetrimino pieces. The simplicity of this particular difficulty mechanic is the main reason why Tetris was chosen, as it allows our adaptive models to easily manipulate the speed by which a chosen piece is falling. In addition, the Tetris game can be played with only one hand which is convenient to place the EDA sensor on the non-dominant hand.

The main difference between traditional Tetris and the version utilized within this work, is that the game will not terminate once a block reaches the upper "game-over" boundary, and instead will clear all the blocks and instantly reset the playing field. This was done to have play sessions of the same duration. Furthermore, the game also logs player and model information that occurs during play. All events are logged with an adjoining time-stamp and include: the predicted emotion (boredom vs. anxiety); horizontal line completion, and the moment of a game-reset (i.e. upper boundary reached).

3.6 User study

To validate the effectiveness of our DDA approach a user-study was conducted with the Tetris game. The objective of this investigation was to observe the behavior of the emotion recongition model during an entire game of Tetris and how the difficulty was perceived by each participant during play. The experiment was conducted in a public setting during an open science and research demonstration day. Thus, it was tested in a ecological situation where emotion recognition models generally performs worst than in a laboratory (for instance players are playing in front of many people which can bias physiological reactions due to the emerge of social emotions). Two computers were used, where each one was running a different version of the Tetris game: the ABS and the REA adaptation variants, which are further described below. Playtime was limited to 120 seconds. All participants gave their informed consent, while the EDA sensors were being prepared. At the start of each game both the REA and ABS versions would equally cycle through 3 diverging starting difficulties: *Easy*, *Medium* and *Hard*. This was done in order to visualize the efficiency of each adaptation type towards situations of extreme difficulty, or simplicity. During play, all participants were required to use noise-canceling headphones to avoid physiological reactions due to the environment. At the end of the play session each participant was asked to rate on a five point Likert scale the perceived difficulty of the first and last 60 seconds of the game.

The proposed affective models output a value between the interval $[0, 1]$, where values close to 0 indicate a strong certainty on boredom while values close to 1 represents a higher confidence towards anxiety. Classifiers outputs were computed using 20s of signal. The first output was thus obtained 20s after the play begun. However, using a signal buffer, the following decisions were taken every 10s based on the last 20s of signal. Based on classifiers outputs two game adaptation strategies were developed. The first strategy called absolute (ABS) consists of increasing the game difficulty if the classifier output is lower than 0.5 (i.e. boredom is detected), or decreasing it if otherwise (i.e. anxiety is detected). By using this method it is expected that difficulty will start to converge towards an "ideal" challenge for players, lying within the boundary between boredom and anxiety. The second strategy referred to as relative (REA), consists of changing the difficulty at step t with respect to the previous classifier output at step $t - 1$. If the classifier output is superior to the previous one, than adaptation assumes that the player is more anxious and thus decreases

the difficulty, while if the contrary occurs the adaptation assumes the player is bored and thus increases difficulty. The idea behind this strategy is that the model would utilize historical information to make the decision to either increase or decrease the overall difficulty by simply comparing output values rather than use the absolute model value. Thus, with this adaptation style we consider that the fluctuation of boredom/anxiety directly relates to the previous measurement and not a global scale. Given the short length of a playing session, using the direct previous value $(t-1)$ was considered to be sufficient.

4 Results

4.1 Affective Modeling

Models previously presented in section 3 are trained and subsequently tested on the dataset described in section 3.1. Each classifier is cross-validated using a leave one participant out method. The classifiers performance is presented in table 1. Three performance measures are used: the accuracy measures the percentage of samples correctly classified, while Cohen's Kappa and F1 scores are measures which are not sensitive to the slight class imbalance present in the dataset.

Table 1. Classifiers performance for anxiety and boredom recognition

Classifier	Accuracy	Kappa	F1
QDA	66.1%	0.315	0.647
LDA	69.5%	0.388	0.693
DeepNet	70.4%	0.408	0.704
DeepNet ensemble	73.2%	0.465	0.732

Table 1 showcases the performance of each classifier trained on the data collected in a laboratory. Results suggest that the proposed deep network performed just as efficiently, and even slightly better than the baseline models, achieving an accuracy of 69.5% and 70.4% for LDA and the deep network, respectively. In addition to validating the superiority of the Deep network model, the kappa and F1 score demonstrate that none of the classifiers were sensitive to the slight class imbalance.

Given the favorable results obtained from deep networks, an ensemble of these networks was also constructed to better generalize on the test data. Overall, an improvement of 2.8% was observed when using ensemble learning on the deep networks, showcased in Table 1. Table 2 shows the confusion matrix of this model. A slight bias can be observed towards the anxiety class, which suggests that this network tend to over-estimate anxiety. Taking into consideration the improved performance of this model, it was decided to integrate it specifically in the subsequent user studies for the detection of boredom and anxiety during a playthrough of Tetris. For this use case, the ensemble deep network architecture was re-trained on the full data set (i.e. without cross validation)

Table 2. Confusion matrix (in %) of the ensemble of deep networks

	Estimated boredom	Estimated anxiety
True boredom	36%	15%
True anxiety	12%	37%

4.2 User study

Although over 100 individuals participated in the experiment, it was necessary to discard almost half of this collected data because signals were quite noisy and several participants were below the age of consent (i.e. minors). In total 56 of the participants presented usable data, where 26 and 30 of these were assigned to both the ABS and REA variations of the game, respectively. Table 3 showcases the distribution of the starting difficulty according to each adaptation method experimented.

Table 3. The total number of participants and their respective Tetris variation and starting difficulty.

Difficulty	ABS	REA
Easy	13	14
Medium	5	9
Hard	8	7
Total	26	30

In terms of familiarity with the Tetris game, the majority of participants stated that they had frequently played the game (32%), while 14% and 20% claimed to have never or just occasionally played the game. 14% asserted that they were playing regularly, while the remaining participants did not directly state their skill levels.

Decision Accuracy during Play: A Binomial significance test was conducted to check if the difficulty of the second half was harder than the first half. The most difficult half of the game was determined by selecting the half with the highest self-reported difficulty on the likert scales (see section 3.6). This ranking approach was used to remove participant bias, as the interpretation of such scales may vary substantially among different individuals [14]. Furthermore, participants who gave the same rating for both halves of the game were discarded from the analysis. The significance of the result was tested using a binomial test.

Figure 2 shows that there is a discrepancy between the ABS and REA adaptation methods. For REA the majority of participants (80%, $p < 0.05$) ranked the second half of the game as more difficult than the first half. Early impressions from these results suggest that the REA adaptation method was not particularly effective in managing the overall difficulty of the participants, with the majority

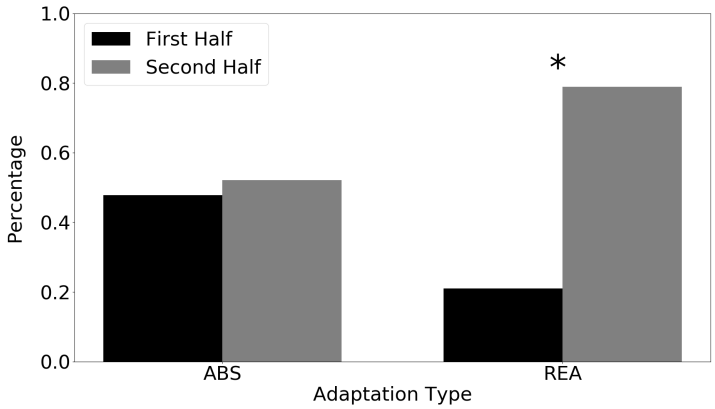

Fig. 2. Distribution of difficulty (i.e. first and second halves) chosen by participants, according to the different adaptation variants. Significant results are indicated by *, where $p < 0.05$.

Table 4. The classification accuracy of the model in the user-study, computed for participant self-reports, using different adaptation methods for each starting difficulty game variant (i.e. Easy, Medium and Hard). p-value is indicated in parenthesis

Adaptation	Acc_{Easy}	Acc_{Medium}	Acc_{Hard}
ABS	**31.6% (0.008)**	69.6% (0.09)	62.9% (0.18)
REA	51.7% (0.89)	64% (0.23)	55% (0.83)

of individuals struggling with the second half of the game. Contrarily to the REA results, participants presented more ambiguous answers for the ABS version of the game, with a slight non-significant bias towards the second half being considered more difficult. Unlike REA, the ABS results show that participants had a more balanced experience with this version of the game, although these specific results were not be significant it does suggest that this particular model adapted more accurately towards the players skill.

To measure the detection performance in the user-study according to participants reports, we defined that the classifier performed well if it increased (resp. decreased) difficulty in the second half when the participants reported a low (resp. high) difficulty in the first. To facilitate the analysis and concentrate specifically on how the classifier adjusted difficulty based on an initial perception, runs where participants perceived the first half of the game to be "okay" were discarded. Table 4 shows the accuracy of the model for each starting difficulty game variant. The ABS adaptation type presents the most varied classification accuracy, where the easy difficulty variant obtains a statistically significant average accuracy of 31.6% ($p < 0.05$), a substantially lower accuracy when compared to both medium (69.6%, $p < 0.1$) and hard (62.9%, $p < 0.2$) variants. No significant results were obtained for the REA adaptation.

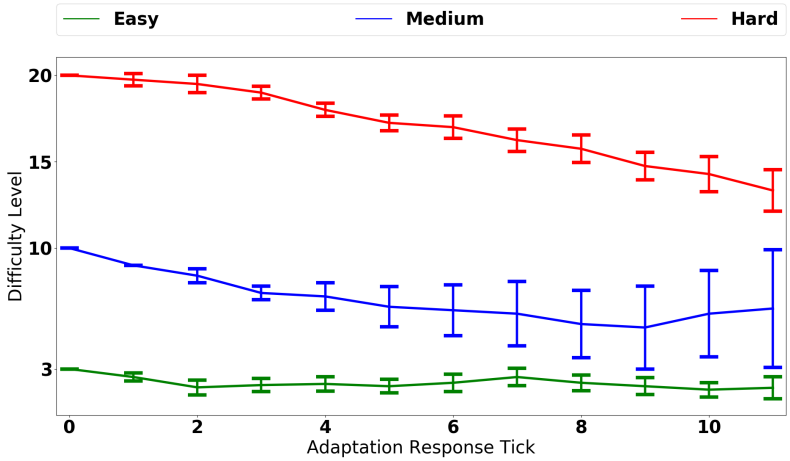

Fig. 3. Difficulty fluctuation during play using the ABS model. Each tick represents the average and standard error of each increase or decrease in difficulty level done by the model. The different colors represents the game starting difficulties.

Difficulty Adaptation over Time: During play both models attempt to slowly adapt the player difficulty based on the physiological responses obtained from the player. In order to visualize how the models adapted play, the average difficulty of the game (and the standard error) was plotted in Figure 3 and Figure 4 for each type of adaptation and starting difficulty. In these figure, each tick of the x-axis represents a change of difficulty done according to the detected player's emotion. The difficulty (y-axis) is represented as a positive natural number, where a higher level represents a more difficult state of the game. Unfortunately, due to an error with the logging system within the REA version, we were unable to record the final level modification of this version of the game.

Figure 3 shows that the ABS adaptation did influence the difficulty fluctuation during play. For the *hard* condition in particular the ABS model detected anxiety and progressively lowered the difficulty until almost reaching the "medium" starting difficulty. Interestingly, in the *medium* condition a similar pattern is observed to that of the *hard* condition, where the early decisions of the model is to consistently lower the difficulty level. However, contrarily to the previous difficulty during the mid-game a larger standard error is observed, and during the latter stages the average difficulty actually increased. This does suggest that for some participants the adaptation did not just constantly reduce the difficulty, in fact during the latter stages of the game the model attempted to adjust the overall game difficulty, which makes sense considering that players potentially proved more proficient in the latter portions of the game. Out of all game variants the *easy* condition proved to be the most stable, with little fluctuation throughout the game. This goes in-line with the low accuracy results

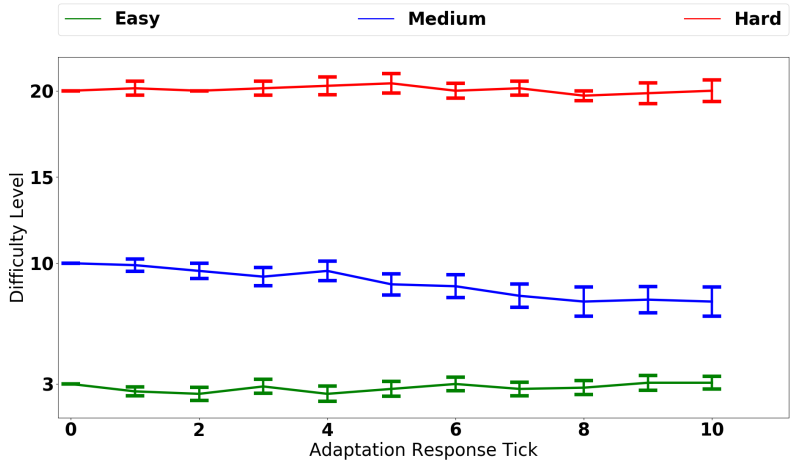

Fig. 4. Difficulty fluctuation during play using the REA model. Each tick represents the average and standard error of each increase or decrease in difficulty level done by the model. The different colors represents the game starting difficulties.

observed in this condition and further suggests the models struggle in increasing difficulty (i.e. detecting the boredom state).

Contrarily to the previous adaptation type, for the REA plot (Figure 4) a more stable difficulty fluctuation can be observed across all game variants. The *medium* condition presented the most fluctuation, with the game slightly reducing the overall game difficulty during the second half of play, which may be a reason why a higher accuracy was previously observed for this condition compared to the other two. Both *hard* and *easy* conditions did not present substantial differences, where for the majority of play the adaptation method does not change the difficulty of the game. The reason for this stabilization is due to how this model operates, specifically when comparing the previous and current outputted values by the model for the decision making process. More precisely, this adaptation method attempts to treat outputted values as a relative measure according to the previous one, even if the difference between the two values are minimal a decision is still made. This consequently manifests as a constant state of difficulty fluctuation that does not allow the adaptation to commit towards a "long-term" objective, and thus why it often gets stuck within a range of around three levels from the initial starting difficulty.

5 Discussion

5.1 Model Performance

The performance obtained from the classifiers trained on laboratory data show the superiority of the deep network architecture. As previously mentioned in section 3.3, there are two factors which increase the performance of LDA/QDA:

feature selection and the rest period subtraction. Feature selection was done according to our previous study, which was accomplished on the same dataset [7]. Thus, it is possible that the baseline models slightly over-fitted the database using knowledge extracted from the full dataset. This could result in artificial improvement of the QDA/LDA performance. In addition, the LDA/QDA used additional information (i.e. the rest period) to take decisions on anxiety and boredom. However, despite these potential advantages of the baseline models, the deep network still reached an higher accuracy than both QDA and LDA. Importantly, avoiding a rest period allows deploying emotion recognition systems without any calibration phase to the user. This is particularly relevant to reach market deployment or to perform experiment in an ecological situation such as the one we presented.

It is important to note that the data used for training was collected from a total of 20 participants, aggregating only 590 minutes of total play-time. Although the number of samples is relatively high (21918 samples), there is a lot of data redundancy since consecutive windows overlap by 95%. This contrasts with the typical tendency observed in previous work on deep networks, where such models are often trained on large subsets of data, which typically is a requirement for these types of network architectures [23]. This demonstrates the capabilities of the proposed network in generalizing on unseen data without requiring of huge quantities of training data.

It remains difficult to compare the obtained results with other studies because the number and type of emotions to be detected is not the same. However, the Kappa score of 0.465 obtained by the ensemble of deep networks demonstrates the feasibility of user-independent affect recognition contrarily to the 0.01 Kappa score obtained in [3]. Although the best reported accuracy of 73% (for two classes) is lower than the 78% (for three classes) reported in [19], our model is user-independent contrarily to the one proposed in [19]. The closest study to our work is probably [22] as anxiety, boredom and flow were detected using a deep architecture trained on 15.5 hours of physiological signals (heart rate and EDA). A very similar performance was obtained using this approach (around 70% to classify 2 classes). Together with our results, this reinforces the statement that deep architectures can be trained with only a 10 to 15.5 hours of physiological activity.

Finally, the performance reported by the players in the user study is much lower than the one computed on the original database, dropping to a value between 31% to 70%, depending on the original difficulty. The low performance in the easy difficulty could be explained by a bias of the classifier toward the anxiety class which is already observed in the confusion matrix computed on the laboratory data (see Table 2). Interestingly a similar bias was reported in [22] who showed that boredom is more difficult to detect than flow and stress. However this bias might not be the only responsible for the drop of performance. An additional explanation is that the EDA sensors were not the same for training and for the real time measurements. It is thus possible that the features captured by the deep networks on the original sensors do not have the same shape and

occurrence in the signals recorded by the real-time sensors. To circumvent this problem, a transfer learning approach [37] might allow to adjust the results to the new sensors.

5.2 User Perception of Adaptation Methods

Early results show a clear distinction between both the ABS and REA models. The ABS adaptation in particular also showed a few distinctions between the different game difficulty conditions. For the *hard* game type, the ABS adaptation method did attempt to adjust the difficulty towards more sensible levels considering that the majority of participants were mostly casual players, which can be observed from the degree of accuracy obtained (62.9%). Although the latter results were not significant, it important to note that for both *hard* and *medium* conditions a lower amount of participant data was collected. The tendency for ABS was to guide participants towards the *medium* level difficulty interval, as the model attempted to gradually lower the difficulty as much as possible (i.e. one level per "tick"). Furthermore, for this particular difficulty variant the ABS adaptation presented a clear bias towards decreasing the game level (i.e. detecting anxiety), which may be the reason why this model performed better in the *Hard* compared to *Easy* condition. In a way this does show that the model is capable of self-regulating the difficulty based on the participant to a certain degree, however these initial results suggest that for the model to do so efficiently the participant must already be sufficiently challenged.

For example the *Medium* version of the game presents an increased variability in the standard error, compared to the other game difficulty variants, in the latter portions of the game. This can be expected when managing difficulty within this range, as the majority of players may feel more comfortable playing the game. The range itself is also complicated to measure effectively, as it lies exactly within the space of being "just right" and, either "too difficult" or "too easy". Interestingly, this game condition presented a similar pattern to *hard* during the initial instants of the game, but in latter portions ABS did attempt to re-increase the difficulty. This could be explained by the fact that the game effectively adjusted to the player's competence and the the plateau observed at the end of the curve corresponds to the average competence of the recorded players. For the *easy* condition on the other hand it was apparent that the ABS adaptation model struggled in comparison due to the present anxiety bias. There are several instances where the model attempts to lower the difficulty even though the game was not considered challenging at that point in time. This suggests that the model does have a higher difficulty in detecting the boredom state in relation to anxiety as explained in section 5.1.

Unlike the ABS adaptation, the REA variation presented a more stable difficulty fluctuation in comparison. This phenomenon was observed on all three starting difficulty variations, which suggests that on average the model lingered within the same difficulty range over the course of a playing session. This is also reflected on the average accuracy of the REA adaptation method, which in general presented ambiguous accuracies ($\approx 55\%$). This lack of adaptation was also

apparent to the participants themselves, where a substantial bias on perceived difficulty is present specifically on the second half of a playing session. The reasoning for this bias is due to the game itself and how it tends to be played. At the latter points of the game players tend to have a more Tetrimino cluttered playing area due to previous player mistakes, this will consequently slowly increase the chances of a potential game-over situation, which may be perceived as the game becoming substantially more difficult.

5.3 Future Work

Given that the current adaptation methods have only two options, i.e. either increase or decrease difficulty, a potential third class might be advantageous for ambiguous values outputted by the model. For future approaches on adaptation, including an hysteresis on the change of decision might actually help mitigate some of the bias that may exist within the model towards anxiety. Furthermore, for the ABS adaptation it might make sense to discard predictions that are within close proximity of 0.5 (i.e. the boundary between both emotional states), as they are too ambiguous to make an actual decision on difficulty.

Although, this paper focused specifically on experimenting with two diverging types of difficulty adjustment systems, an alternate study focusing specifically on the enjoyability while using such a system might be advantageous, where a comparison can be made between the classical approach (i.e. classic Tetris) and adaptive difficulty. A common argument against these types of dynamic adjustment systems usually focus on how certain players actually enjoy overcoming these arduous challenges; offering the player a sense of accomplishment. Although this is certainly true for certain types of games, this does not discount the advantages that such adaptive systems may offer in other applications for which having a task challenge that matches users' competences is relevant.

Future improvements to the proposed models could also possibly be obtained if in addition to physiological data, information from the gameplay itself could be utilized as input. In the current version of the model, input data consist purely of raw GSR data. The accuracy of these predictions could potentially be enhanced if in conjunction to the physiological data, some form of gameplay information could also be processed into features informing the model of: current difficulty level, number of lines, Tetriminos used and among others. However, in order to pin-point exactly what features influence emotion, and how to specifically train a model to use these gameplay features and relate it to affect is not a straightforward task. The simplest method would be to simply have participants play a game of Tetris, while logging all gameplay elements and measuring GSR synchronously. From this data, attempt to correlate certain gameplay characteristics to specific traits obtained from GSR signals (i.e. phasic peaks). Another method could be through real-time annotation, where participants annotate replays of Tetris playing sessions through an annotation system like RankTrace [21]. From these annotations a relation could possibly be made from the gameplay features and the annotated data.

Lastly, it is important to keep in mind that the models explored within this work can also be applied to several other domains. For example, applications that require constant engagement and monitoring from their users can benefit by reacting accordingly if an individual's attention starts to wane due to boredom. The software may be able to thus stimulate such individuals in order to retain their attention to the task at hand.

6 Conclusions

This paper presented a comparison between two diverging approaches for an affective model to adapt difficulty within the Tetris game. The affective model is capable of adjusting difficulty using human physiological responses, specifically electrodermal activity, and making an emotional prediction of the player's state, i.e. anxious or bored. Several algorithms were tested, where a higher performance was observed with an ensemble deep neural-network approach (73.2%). This model was then tested with 56 participants using two different adaptation approaches: the absolute (ABS) and relative (REA). Results show that the ABS outperformed the REA method. The ABS model was particularly efficient at adapting the game difficulty when players were stating to play in hard or medium difficulties. This paper builds upon previous research in dynamic difficulty adjustment by testing the affective models out of the laboratory.

Acknowledgments

We would like to thank the Blue Yeti[3] company for developing a first version of the affective Tetris game which was adjusted for the purpose of this study.

References

1. Abegg, C.: Analyse du confort de conduite dans les transports publics. Master thesis, University of Geneva (2013)
2. Alhargan, A., Cooke, N., Binjammaz, T.: Affect recognition in an interactive gaming environment using eye tracking. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 285–291. IEEE (2017)
3. Alzoubi, O., D'Mello, S.K., Calvo, R.A.: Detecting naturalistic expressions of nonbasic affect using physiological signals. IEEE Transactions on Affective Computing **3**(3), 298–310 (2012). https://doi.org/10.1109/T-AFFC.2012.4
4. Andrade, G., Ramalho, G., Santana, H., Corruble, V.: Extending reinforcement learning to provide dynamic game balancing. In: Proceedings of the Workshop on Reasoning, Representation, and Learning in Computer Games, 19th International Joint Conference on Artificial Intelligence (IJCAI). pp. 7–12 (2005)
5. Bevilacqua, F., Engström, H., Backlund, P.: Game-calibrated and user-tailored remote detection of stress and boredom in games. Sensors (Switzerland) **19**(13), 1–43 (2019). https://doi.org/10.3390/s19132877

[3] http://www.blueyeti.fr/en/

6. Calvo, R., D'Mello, S., Gratch, J., Kappas, A., Calvo, R., D'Mello, S., Gratch, J., Kappas, A.: Introduction to Affective Computing. In: The Oxford Handbook of Affective Computing. Oxford University Press (jan 2015). https://doi.org/10.1093/oxfordhb/9780199942237.013.040
7. Chanel, G., Rebetez, C., Bétrancourt, M., Pun, T.: Emotion Assessment From Physiological Signals for Adaptation of Game Difficulty. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans **41**(6), 1052–1063 (November 2011). https://doi.org/10.1109/TSMCA.2011.2116000
8. Chollet, F., et al.: Keras. https://github.com/keras-team/keras (2015)
9. Csikszentmihalyi, M.: Beyond boredom and anxiety. Jossey-Bass (2000)
10. Demasi, P., Adriano, J.d.O.: On-line coevolution for action games. International Journal of Intelligent Games & Simulation **2**(2) (2003)
11. Ewing, K.C., Fairclough, S.H., Gilleade, K.: Evaluation of an Adaptive Game that Uses EEG Measures Validated during the Design Process as Inputs to a Biocybernetic Loop. Frontiers in Human Neuroscience **10** (may 2016). https://doi.org/10.3389/fnhum.2016.00223
12. Guillaume, C., Konstantina, K., Thierry, P.: GamEMO: How Physiological Signals Show your Emotions and Enhance your Game Experience. In: Proceedings of the 14th ACM international conference on Multimodal interaction - ICMI '12. pp. 297–298. ACM Press, New York, New York, USA (October 2012). https://doi.org/10.1145/2388676.2388738
13. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation **9**(8), 1735–1780 (November 1997). https://doi.org/10.1162/neco.1997.9.8.1735
14. Holmgård, C., Yannakakis, G.N., Martinez, H.P., Karstoft, K.I.: To rank or to classify? annotating stress for reliable ptsd profiling. In: Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on. pp. 719–725. IEEE (2015)
15. 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. Journal of Gaming & Virtual Worlds **3**(3), 181–199 (sep 2011). https://doi.org/10.1386/jgvw.3.3.181_1
16. Koster, R.: Theory of fun for game design. O'Reilly Media, Inc. (2013)
17. Lang, P.J., Greenwald, M.K., Bradley, M.M., Hamm, A.O.: Looking at pictures: affective, facial, visceral, and behavioral reactions. Psychophysiology **30**(3), 261–273 (1993)
18. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015). https://doi.org/10.1038/nature14539
19. Liu, C., Agrawal, P., Sarkar, N., Chen, S.: Dynamic Difficulty Adjustment in Computer Games Through Real-Time Anxiety-Based Affective Feedback. International Journal of Human-Computer Interaction **25**(6), 506–529 (2009)
20. Lopes, P., Liapis, A., Yannakakis, G.N.: Framing tension for game generation. In: Proceedings of the International Conference on Computational Creativity (2016)
21. Lopes, P., Yannakakis, G.N., Liapis, A.: Ranktrace: Relative and unbounded affect annotation. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE (2017)
22. Maier, M., Elsner, D., Marouane, C., Zehnle, M., Fuchs, C.: DeepFlow: Detecting Optimal User Experience From Physiological Data Using Deep Neural Networks. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. vol. 2019-Augus, pp. 1415–1421. International Joint Conferences on Artificial Intelligence Organization (aug 2019). https://doi.org/10.24963/ijcai.2019/196

23. Malik, J.: What led computer vision to deep learning? Communications of the ACM **60**(6), 82–83 (May 2017). https://doi.org/10.1145/3065384
24. Martínez, H.P., Bengio, Y., Yannakakis, G.N.: Learning Deep Physiological Models of Affect. IEEE Computational Intelligence Magazine **8**(2), 20–33 (2013). https://doi.org/10.1109/MCI.2013.2247823
25. Martínez, H.P., Garbarino, M., Yannakakis, G.N.: Generic physiological features as predictors of player experience. In: International Conference on Affective Computing and Intelligent Interaction. pp. 267–276. Springer (2011)
26. Mühl, C., Allison, B., Nijholt, A., Chanel, G.: A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. Brain-Computer Interfaces **1**(2), 66–84 (May 2014). https://doi.org/10.1080/2326263X.2014.912881
27. Pagulayan, R.J., Keeker, K., Wixon, D., Romero, R.L., Fuller, T.: User-centered design in games. In: The human-computer interaction handbook. pp. 883–906. L. Erlbaum Associates Inc. (2002)
28. Poh, M.Z., Swenson, N.C., Picard, R.W.: A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. IEEE transactions on bio-medical engineering **57**(5), 1243–52 (may 2010). https://doi.org/10.1109/TBME.2009.2038487
29. Rani, P., Sarkar, N., Liu, C.: Maintaining Optimal Challenge in Computer Games through Real-Time Physiological Feedback. In: 11th HCI International. Lawrence Erlbaum Associates, Inc, Las Vegas, USA (2005)
30. Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication **53**(9-10), 1062–1087 (2011). https://doi.org/10.1016/j.specom.2011.01.011
31. Sequeira, H., Hot, P., Silvert, L., Delplanque, S.: Electrical autonomic correlates of emotion. International Journal of Psychophysiology **71**, 50–56 (2009)
32. Shaker, N., Yannakakis, G.N., Togelius, J.: Towards automatic personalized content generation for platform games. In: AIIDE (2010)
33. Spronck, P., Ponsen, M., Sprinkhuizen-Kuyper, I., Postma, E.: Adaptive game ai with dynamic scripting. Machine Learning **63**(3), 217–248 (2006)
34. Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Schuller, B., Zafeiriou, S.: Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5200–5204. IEEE (March 2016). https://doi.org/10.1109/ICASSP.2016.7472669
35. Wang, C., Pun, T., Chanel, G.: A Comparative Survey of Methods for Remote Heart Rate Detection From Frontal Face Videos. Frontiers in Bioengineering and Biotechnology **6**(MAY) (may 2018). https://doi.org/10.3389/fbioe.2018.00033
36. Yannakakis, G.N., Martínez, H.P., Jhala, A.: Towards affective camera control in games. User Modeling and User-Adapted Interaction **20**(4), 313–340 (2010)
37. Zheng, W.l., Zhang, Y.q., Zhu, J.Y., Lu, B.l.: Transfer components between subjects for EEG-based emotion recognition. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 917–922. IEEE (September 2015). https://doi.org/10.1109/ACII.2015.7344684