Generic Physiological Features as Predictors of Player Experience

Héctor P. Martínez[1], Maurizio Garbarino[2], and Georgios N. Yannakakis[1]

[1] Center for Computer Games Research,
IT University of Copenhagen,
Rued Langgaards vej 7, 2300, Denmark
{hpma,yannakakis}@itu.dk
[2] IIT Unit Dipartimento di Elettronica ed Informazione,
Politecnico di Milano,
Piazza Leonardo Da Vinci 32,20133 Milano, Italy
garbarino@elet.polimi.it

Abstract. This paper examines the generality of features extracted from heart rate (HR) and skin conductance (SC) signals as predictors of self-reported player affect expressed as pairwise preferences. Artificial neural networks are trained to accurately map physiological features to expressed affect in two dissimilar and independent game surveys. The performance of the obtained affective models which are trained on one game is tested on the unseen physiological and self-reported data of the other game. Results in this early study suggest that there exist features of HR and SC such as average HR and one and two-step SC variation that are able to predict affective states across games of different genre and dissimilar game mechanics.

Keywords: generality, heart rate, skin conductance, affective modeling, preference learning, feature selection

1 Introduction

One of the primary goals pursued by affective computing [1] is the creation of accurate computational models of user experience. Studies in psychophysiology [2], in particular, contribute to this aim by drawing the relationships between user physiological signals, context-based cues and affective and/or cognitive states derived from empirical data. Psychophysiology focuses on the collection of physiological data from a group of subjects performing a set of controlled tasks which are designed to elicit a palette of dissimilar user emotional responses. Key characteristics of the physiological data obtained are then mapped to expected, observed or reported affective user states to form the desired affective models. While the generality of physiology-based affective models across different tasks, groups of subjects and experimental protocols comprises a necessary step towards reliable affective interaction, research in psychophysiology has not yet focused on the generic attributes of particular physiological features.

This paper examines the hypothesis that there exist features extracted from physiological signals which can act as inputs of accurate computational models that are able

to predict affective states of players across dissimilar games. For this purpose artificial neural network (ANN) models are trained via *neuroevolutionary preference learning* [3] on the signals to predict affective states of players expressed as pairwise preferences. Automatic feature selection chooses a subset of the extracted features that maximise the prediction accuracy of the model.

To test the generality of physiology-based affective models across games we collected heart rate (HR) and skin conductance (SC) signals and self-reports of two independent groups of participants playing two dissimilar games. In particular, a prey/predator and a car racing game are used in the experiments presented in this paper. The models are trained on data from one game and tested on the data of the other game revealing statistical features of physiology that are generic predictors of affect across different games. The differences between the experimental protocols, the biofeedback devices, the experience questionnaires and, most importantly, the games themselves are expected to affect the validation accuracy of the models when presented to unseen data from a dissimilar game. However, results show that there exist key physiological features such as average HR and SC variation features that successfully predict dissimilar affective states across the two games.

2 Related Work

The field of psychophysiology [2] explores the relationship between emotions and physiological signals of participants on their daily lives or in a laboratory experiment in which different mental states are induced. Several tasks have been employed to elicit emotional responses in experiment participants including watching videos, reading and solving mathematical problems (see [4–6] for extensive reviews) but also playing videogames (e.g. [7, 8]).

More recently, with the establishment of the new field of Affective Games [9, 10], research on detecting affect in computer game players is rapidly growing in several directions including psychophysiology [11]. Generally, these studies apply the concepts and methodologies of traditional psychophysiology [12] to analyse the effect of different game aspects such as social experiences (e.g. playing with friends or strangers [13]), game events (e.g. killing an enemy [14]) and game features (e.g. camera viewpoint [3]) to the player's state.

For example, Rani et al. [15] explore the correlation between anxiety, engagement, boredom and frustration and several physiological signals (HR and SC among others) while playing Pong while Nacke and Lindley [16] investigate the correlations between flow, boredom and immersion, and jaw electromyography (EMG) and SC in a first person shooter. In this paper we do not focus on linear psychophysiological relationships but, instead, we apply machine learning to create non-linear models that approximate the function between a set of physiological signal attributes and self-reported affective states. While most studies in machine learning within psychophysiology ([17, 18, 7, 19] among others) focus on the classification accuracies of different methods and disregard the particular models built, this paper analyses the effect of various physiological features in the prediction of affective states.

Mandryk et al. [13] use fuzzy rules to map HR, SC, respiration and EMG of the jaw muscles to an arousal-valence space and to levels of fun, boredom, challenge, excitement and frustration during a hockey computer game. Yannakakis et al. [20, 21] model the fun pairwise preferences of children playing physical interactive games from HR and SC using neuro-evolution. In all aforementioned studies psychophysiological models are built and the relationship between physiology and emotion in computer games is examined but the generality of the reported physiological models in similar or different game genres is not further investigated.

In [3] models for a set of different affective states (e.g. frustration and fun) built on HR, SC and blood volume pulse (BVP) features and game context cues are tested in a different version of the game from the one they are trained on. This paper is novel in that models are trained and evaluated in games of dissimilar genres which allow one to identify features of physiology that are generic predictors of affect in games.

3 Data Collection

The two datasets used in this study were gathered via two independent experimental surveys in which a group of participants played a sequence of different variants of the same game during which a set of physiological signal data is collected. The players are asked to report their affective preferences about these game variants in a post-experience manner. This section provides an overview of the experiments highlighting the similarities and the differences between them. The reader is referred to [3, 22] for a more detailed description of the games and the experimental protocols used.

3.1 Games

The two test beds used in the experiments are short single player games controlled with the arrow-keys (one-handed). In the first test-bed game, named Maze-Ball[3], the player controls a green ball in a 3D maze with the goal of picking up as many pellets as possible in 90 seconds while avoiding a group of enemies. The eight variants of the game correspond to different virtual camera profiles. The second test-bed game, TORCS[4], is an open source racing car simulator customized in the experiment reported in this paper to allow 3 minute-long races against a computer-controlled car opponent whose performance skills change across variants.

3.2 Participants

Thirty six subjects (80% males) aged from 21 to 47 years (mean and standard deviation of age equal 27.2 and 5.84, respectively) played Maze-Ball at the Center for Computer Games Research (IT University of Copenhagen) and 75 subjects (80% males) aged from 18 to 30 years (mean and standard deviation of age equal 23.40 and 4.12, respectively) played TORCS at the IIT Unit Dipartimento di Elettronica ed Informazione (Politecnico di Milano).

[3] http://itu.dk/people/hpma/MazeBall.html
[4] http://torcs.sourceforge.net/

3.3 Experimental Protocol

In the TORCS experiment subjects played 7 games and reported (at the end of each game) whether they liked more the game they just played than the previous one using a 2 alternative forced choice (2-AFC) questionnaire. On the other hand, Maze-Ball participants played 8 games grouped in pairs after completing a tutorial. After each pair, subjects had to report whether the first or the second game felt more *anxious*, *boring*, *challenging*, *exciting*, *frustrating*, *fun* and *relaxing* or whether the affective state was felt equally in both games or in neither of them (4-AFC).

The final TORCS dataset contains 450 samples (pairs) and the Maze-Ball datasets contain, respectively, 97, 92, 90, 90, 86, 83 and 54 for challenge, fun, frustration, relaxation, anxiety, excitement and boredom after removing the unclear (equal or neither) preferences.

3.4 Physiological Signals and Extracted Features

Skin conductance, s, and blood volume pulse are collected at 32Hz from the Maze-Ball participants using the IOM biofeedback device. Skin conductance signal is collected at 256Hz and blood volume pulse at 2048Hz from the TORCS participants using the ProComp Infiniti hardware (signal artifacts are removed and the signals are filtered and re-sampled at 64Hz — see [22]). In both datasets, the heart rate signal, h, is inferred by the BVP signal (see [3] for details) and the magnitude (SM), m, and the duration (SD), d, of signal variation have been derived from SC [23].

The following set of features is extracted from the last 60 seconds of each signal ($\alpha \in \{h, s, m, d\}$) inspired by previous studies on physiological feature extraction [21, 22]:

- average ($E\{\alpha\}$) and variance ($\sigma^2\{\alpha\}$) of the signal;
- initial (α_{in}) and final (α_{last}) recording and difference between them (Δ^α);
- minimum ($min\{\alpha\}$) and maximum ($max\{\alpha\}$) signal recording and difference between them (D^α);
- time when maximum ($t_{max}\{\alpha\}$) and minimum ($t_{min}\{\alpha\}$) samples were recorded and difference between those times (D_t^α);
- average first and second absolute differences ($\delta^\alpha_{|1|}$ and $\delta^\alpha_{|2|}$, respectively);
- Pearson's correlation coefficient (R_α) between raw α recordings and the time t at which data were recorded;
- autocorrelation (lag equals the sampling rate of α) of the signal (ρ^α);

All features are normalized to the [0,1] interval using standard min-max normalization.

4 Method

The computational models of affect constructed and compared in this paper are trained on self-reported pairwise preferences. The inputs of the ANN models are selected automatically from the set of the above-mentioned statistical features using *sequential*

forward feature selection (SFS). Feature selection is essential in scenarios where the available features do not have a clear relationship and, thus, impact to the prediction of the target output (i.e. it is not easy to decide *a priori* which features are useful and which are irrelevant for the prediction). Moreover the computational cost of testing all available feature sets is combinatorial and exhaustive search might not be computationally feasible in large feature sets. Under these conditions, FS is critical for finding an appropriate set of model input features that can yield highly accurate predictors [24]. Additionally, we would like our models to be dependent on as few features as possible to make it easier to analyze and to make it more useful for incorporation into future implementations of real-time applications.

4.1 Sequential Forward Feature Selection

SFS is a bottom-up search procedure where one feature is added at a time to the current feature set. The feature to be added is selected from the subset of the remaining features such that the new feature set generates the maximum value of the fitness function over all candidate features for addition. The fitness function used in this paper is given by the average cross validation performance of the ANN model on unseen folds of data.

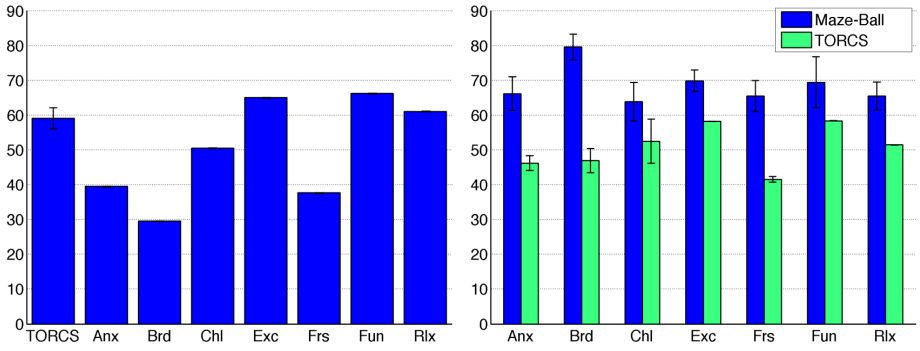
4.2 Neuroevolutionary Preference Learning

We apply preference learning [25] to build affective models that predict users' self-reported emotional preferences based on the subsets of features selected by the FS algorithm. In this study, the models are implemented as single layer perceptrons (SLPs) that are trained via neuroevolutionary preference learning (as in[3]) to map the selected features to a predictor of the reported pairwise emotional preferences. Note that the pairwise preference relationship of the training data is known (e.g. game A is preferred to game B) but the value of the target output is not (i.e. the magnitude of the preference is unknown). Thus, any gradient-based optimization algorithm is inapplicable to the training problem since the error function under optimization is not differentiable.

The expressivity of SLPs allows us to analyse the impact of each one of the selected features to the reported affective preferences. For instance, when a feature with a corresponding high connection weight value increases from one game to another, the magnitude of the predicted preference is increased or decreased depending on the sign of the weight value. On the other hand, weight connections with low values have a small impact on the prediction of preferences.

5 Experiments

In this section we test the generality of the psychophysiological models which are trained on one dataset and evaluated on the other. The performance of the models has been evaluated using cross-validation (10 and 3 folds when training on TORCS and Maze-Ball, respectively, which produces validation sets of acceptable size for all datasets while keeping the total number of runs low). Consequently each run produces a subset of features and a set of ANN models built on those features (1 per fold) that



(a) Model trained on TORCS HR features (b) Model trained on Maze-Ball HR features

Fig. 1. Average cross-validation and standard error of performance on the training dataset and corresponding average accuracy on the testing datasets of the best FS run out of 10. Anx, Brd, Chl, Exc, Frs, and Rlx represent, respectively, anxiety, boredom, challenge, excitement, frustration and relaxation.

are tested on the unseen dataset. The baseline prediction for pairwise preferences depends on the distribution of reported preferences between different pairs of variants. We approximate this baseline with the average accuracy of 100 SLPs — initialized with random weights and different number of randomly selected features — which lies between 45% and 55%. In the remaining of this paper, accuracies (and their standard error) on unseen data outside the 45-55% interval are considered to be significantly different from baseline. Note that, generally, a model that predicts correctly X% (e.g. 45%) of preference samples yields a prediction accuracy of 100-X% (e.g. 55%) if those preferences are inverted.

5.1 Heart Rate features

To find the heart rate features that can predict preferences across the two games, FS runs 10 times on each dataset and the resulting models are tested on the unseen dataset. Figure 1 depicts the accuracies of the ANN models yielding the highest cross-validation accuracy on the training dataset.

TORCS data as training set The best ANN models trained on data obtained from TORCS are able to yield prediction performances above the baseline for all affective states of Maze-Ball excluding challenge (see Fig. 1(a)). Average and minimum HR are the only two features that form the input of these ANN models, both connected with high positive weights. The ANN models yield a prediction accuracy of 65.06%, 66.30% and 61.11% for excitement, fun and relaxation and 60.47%, 70.37% and 62.22% for the inverse of anxiety, boredom and frustration, respectively, while reaching a 59.11% average cross-validation performance on TORCS preferences. These accuracies are comparable to the training performances reached when training models on the Maze-Ball set

(see Figure 1(b)). These results suggest that average and minimum HR (both indicators of sympathetic arousal) are good predictors of preferences in TORCS but also efficient predictors of most affective states examined in Maze-Ball. Unsurprisingly, higher values of these features indicate higher preference for TORCS games but also predict heightened fun, excitement and relaxation and lower anxiety, boredom and frustration in Maze-Ball.

Maze-Ball data as training set The highest performing models built on excitement (69.89%) and fun (69.46%) predict preferences in TORCS with accuracies comparable to the models trained directly on that dataset (58.22% and 58.37%, respectively). The excitement model contains solely average HR while the fun model's input contains two features: $E\{h\}$ (positive weight) and $\sigma^2\{h\}$ (negative weight).

The inverse of TORCS preferences are predicted with a performance of 58.44% by the model built on Maze-Ball frustration. The minimum and average HR features are present once again in the model accompanied by the difference between the times when maximum and minimum are recorded; all three features are connected with a negative weight. Finally, the models trained on anxiety, challenge and relaxation do not appear to be able to predict TORCS players' preferences with accuracies higher than the baseline.

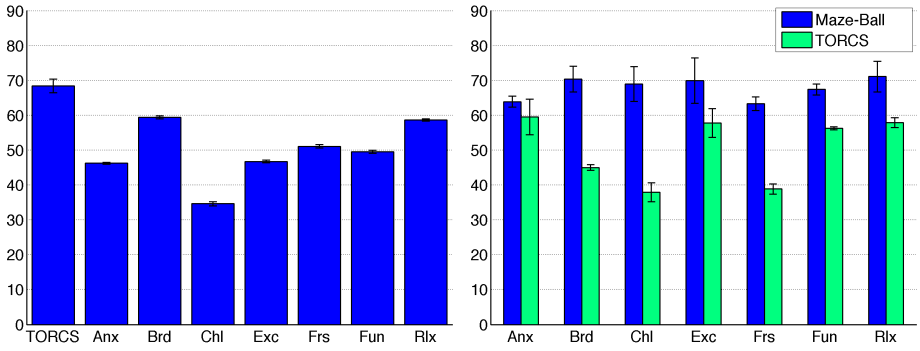
Findings from this set of experiments suggest that HR features such as minimum HR, average HR and HR variance are selected to predict affective states in the Maze-Ball game but are also good predictors of TORCS preference. Specifically, it appears that heightened average HR suggests higher excitement and fun and lower frustration in Maze-Ball but also predicts higher preference in TORCS.

5.2 Skin Conductance features

To find skin conductance features that can predict accurately preferences across games (as in the HR experiments), FS runs 10 times on each dataset and the resulting models are tested on the unseen dataset. Figure 2 shows the accuracies of the ANN models with the highest cross-validation accuracy on their training dataset.

TORCS data as training set The highest performing subset of features (See Figure 2(a)) contains the second absolute difference of SC (positive weight), the second absolute difference of SD (positive weight) and the maximum SC (positive weight). While the second absolute differences on SC and SD are indicators of skin conductance variability, the maximum SC is an indicator of heightened arousal. This ANN model trained on TORCS data is able to predict well Maze-Ball boredom (59.44%), relaxation (58.6%) and inverse of challenge (65.36%) indicating the influence of the three aforementioned features in the prediction of these affective states in Maze-Ball.

Maze-Ball data as training set The best ANN models obtained generate performances that range from 64.44% in predicting frustration to 71.11% in predicting relaxation in Maze-Ball. These models' validation performances in TORCS are rather close to the baseline for most affective states: 59.56%, 45.04%, 62.07%, 57.78%, 58.15%, 56.30%



(a) Model trained on TORCS SC features (b) Model trained on Maze-Ball SC features

Fig. 2. Average cross-validation and standard error of performance on the training dataset and corresponding average accuracy on the testing datasets of the best FS run out of 10. Anx, Brd, Chl, Exc, Frs, and Rlx represent, respectively, anxiety, boredom, challenge, excitement, frustration and relaxation.

and 57.93% for anxiety, boredom, challenge (inverse), excitement, frustration (inverse), fun and relaxation respectively (see Figure 2(b)).

The ANN of the anxiety model contains the average SC and the first absolute difference of the SC signal. Both are connected with a positive weight to the model predicting TORCS preferences with 64.67% accuracy. This result suggests that heightened sympathetic arousal (increased average SC) and high SC variation are linked to preference in TORCS and anxiety in Maze-Ball.

The Maze-Ball challenge model, which did not show a prediction accuracy better than baseline on TORCS when trained on HR features, predicts the inverse of TORCS preferences well (62.07%). The input vector of that ANN model contains the second absolute differences of SM and SC (negative weights), the first absolute difference of SD (negative weights), the maximum SC (positive weight), the average SD (positive weight) and the time when maximum is recorded (negative weight) as inputs. It, therefore, appears that lower SC variation, higher SC and shorter times to reach maximum SC contribute to higher predicted challenge in Maze-Ball and lower predicted preference in TORCS.

The reported frustration ANN model trained on Maze-Ball data predicts the inverse of TORCS preferences with 61.11% accuracy and contains the first absolute difference of SC (negative wight), the SD variance (negative weight) and the autocorrelation of SC signal (positive weight). Maze-Ball frustration and TORCS inverted preference appear to be increased when there is lower SC variation but also when the SC signal's level of randomness is lower (higher autocorrelation).

The fun, excitement and relaxation models present lower TORCS validation accuracies while the boredom model does not even reach a performance which is significantly different from baseline performance. None of the above models contains either the first

or the second absolute difference of SC indicating the importance of these features for predicting TORCS preferences.

6 Conclusions

This paper investigated the generality of affective preference models built on physiological features using two dissimilar games as test-beds. More specifically, the paper explores the existence of heart rate and skin conductance signal features that predict reported player affective states across dissimilar games. Results obtained show a strong dependency between the subsets of features used as inputs to a computational affective model and its prediction accuracy on different datasets highlighting the impact of automatic feature selection in the process of creating these models.

This initial study shows that average and minimum heart rate, indicators of sympathetic activity, can yield good estimators of player reported experiences across different games. Similar results have already been reported in a physical interactive environment in which average heart rate is picked as a predictor of reported fun in those games [20]. Observing the results obtained through the SC experiments, it appears that the 1 and 2-step differences of the SC signal — corresponding to the level of fluctuation existent in the SC signal — are good predictors of affective states across both games tested.

This explorative study on the generality of physiology-based preference models covers two physiological signals, a limited set of statistical features and a limited number of games. Future work includes an extended study on more physiological signals such as BVP and features such as those derived from heart rate variability (average inter beat interval among others) as well as models fusing inputs from different physiological signals and models combining physiology with generic game context features. Furthermore, more complex models such as multi-layer perceptrons will be employed and explored since those might be able to better approximate the mapping between game-independent physiological features and reported affective states.

References

1. Picard, R.: Affective computing. The MIT press (2000)
2. Cacioppo, J., Tassinary, L., Berntson, G.: Psychophysiological science. Handbook of psychophysiology 2, 3–23 (2000)
3. Yannakakis, G., Martínez, H., Jhala, A.: Towards affective camera control in games. User Modeling and User-Adapted Interaction 20, 313–340 (2010), 10.1007/s11257-010-9078-0
4. Andreassi, J.: Psychophysiology:: Human Behavior and Physiological Response (2000)
5. Cacioppo, J., Berntson, G., Larsen, J., Poehlmann, K., Ito, T., et al.: The psychophysiology of emotion. Handbook of emotions pp. 119–142 (1993)
6. Calvo, R., D'Mello, S.: Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Transactions on Affective Computing pp. 18–37 (2010)
7. Rani, P., Liu, C., Sarkar, N., Vanman, E.: An empirical study of machine learning techniques for affect recognition in human–robot interaction. Pattern Analysis & Applications 9(1), 58–69 (2006)
8. Fernandez, R., Picard, R.: Signal processing for recognition of human frustration. In: Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on. vol. 6, pp. 3773–3776. IEEE (1998)

9. Gilleade, K., Dix, A., Allanson, J.: Affective videogames and modes of affective gaming: assist me, challenge me, emote me. In: Proc. DIGRA. vol. 2005 (2005)
10. Hudlicka, E.: Affective game engines: motivation and requirements. In: Proceedings of the 4th International Conference on Foundations of Digital Games. pp. 299–306. ACM (2009)
11. Kivikangas, J., Ekman, I., Chanel, G., Jarvela, S., Salminen, M., Cowley, B., Henttonen, P., Ravaja, N.: Review on psychophysiological methods in game research. Proc. of 1st Nordic DiGRA
12. Fairclough, S.: Psychophysiological inference and physiological computer games. In: ACE Workshop-Brainplay. vol. 7 (2007)
13. Mandryk, R., Atkins, M.: A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. International Journal of Human-Computer Studies 65(4), 329–347 (2007)
14. Ravaja, N., Saari, T., Laarni, J., Kallinen, K., Salminen, M., Holopainen, J., Järvinen, A.: The psychophysiology of video gaming: Phasic emotional responses to game events. In: Proceedings of the DiGRA conference Changing views: worlds in play
15. Rani, P., Sarkar, N., Liu, C.: Maintaining optimal challenge in computer games through real-time physiological feedback. In: Proceedings of the 11th International Conference on Human Computer Interaction. pp. 184–192 (2005)
16. Nacke, L., Lindley, C.: Flow and immersion in first-person shooters: measuring the player's gameplay experience. In: Proceedings of the 2008 Conference on Future Play: Research, Play, Share. pp. 81–88. ACM (2008)
17. Picard, R., Vyzas, E., Healey, J.: Toward machine emotional intelligence: Analysis of affective physiological state. IEEE transactions on pattern analysis and machine intelligence pp. 1175–1191 (2001)
18. Nasoz, F., Alvarez, K., Lisetti, C., Finkelstein, N.: Emotion recognition from physiological signals using wireless sensors for presence technologies. Cognition, Technology & Work 6(1), 4–14 (2004)
19. McQuiggan, S., Lee, S., Lester, J.: Early prediction of student frustration. Affective Computing and Intelligent Interaction pp. 698–709 (2007)
20. Yannakakis, G.N., Hallam, J., Lund, H.H.: Entertainment capture through heart rate activity in physical interactive playgrounds. User Modeling and User-Adapted Interaction 18(1), 207–243 (2008)
21. Yannakakis, G.N., Hallam, J.: Entertainment Modeling through Physiology in Physical Play. International Journal of Human-Computer Studies 66, 741–755 (October 2008)
22. Tognetti, S., Garbarino, M., Bonanno, A., Matteucci, M., Bonarini, A.: Enjoyment recognition from physiological data in a car racing game. In: Proceedings of the 3rd international workshop on Affective interaction in natural environments. pp. 3–8. ACM (2010)
23. Tognetti, S., Garbarino, M., Bonarini, A., Matteucci, M.: Modeling enjoyment preference from physiological responses in a car racing game. In: Computational Intelligence and Games (CIG), 2010 IEEE Symposium on. pp. 321–328. IEEE
24. Martínez, H.P., Yannakakis, G.N.: Genetic search feature selection for affective modeling: a case study on reported preferences. In: Proceedings of the 3rd international workshop on Affective interaction in natural environments. pp. 15–20. ACM (2010)
25. Fürnkranz, J., Hüllermeier, E.: Preference learning. Künstliche Intelligenz 19(1), 60–61 (2005)