Modeling enjoyment preference from physiological responses in a car racing game

Simone Tognetti, Maurizio Garbarino, Andrea Bonarini, Matteo Matteucci

Abstract— **We propose a framework to estimate player enjoyment preference from physiological signals. This can produce objective measures that could be used to adapt dynamically a game to maintain the player in an optimal status of enjoyment. We present a case study on The Open Racing Car Simulator (TORCS) video game. In particular, we focus both on the experimental protocol, which we designed with special attention to produce physiological responses related to the game experience only, and on signal analysis, which produces a simple and general model good enough to estimate player enjoyment preference in real applications.**

I. INTRODUCTION

”A good game is the one you like to play”. With this motto in her mind, the digital game designer conceives rules and structure of a game to maximize the level of enjoyment for the target audience. Realistic ambiance and characters, intelligent and reactive opponents with a human-like behavior can provide enjoyable game experience, however this cannot be assumed a priori. Some evaluation can be done by considering performance parameters, often assuming that a better performance is related to higher enjoyment of the game experience; but, different players may react differently to game features, and enjoyment has to be considered as a personal experience. According to the affective computing, we can assume that physiological response can be related to enjoyment [1], and that it can be taken as an objective measure of it. Thus, we present in this paper the application of a methodological framework for the estimate of preference among game experiences, from the physiological state of the player, in the car racing video game scenario implemented by TORCS – The Open Racing Car Simulator [2] .

With the purpose of comparing and validating the proposed game scenario with respect to recent literature, we have carried out a correlation analysis between physiological data and subject preference during different variants of the game. Results show that features derived from some of the physiological signals (e.g., Galvanic Skin Response (GSR), Blood Volume Pulse (BVP) and Respiration (RESP)) have a high correlation with player reported preferences. On the other hand, as we expected from other studies in literature, only some physiological features show relevant high discriminating power. Therefore, signals such as Heart Rate (HR) or temperature are not really suitable as emotional input because of their poor correlation with user preference. Supported by these findings, we have estimated a linear model based on physiological signals able to predict the subject enjoyment preference during the game. Our approach focuses on the differential comparison of preference between game situations and the resulting model can be used, in a future experiment, to modify at runtime the game experience accordingly to the predicted user preference to optimize user enjoyment.

The ground truth about subject preference has been evaluated by questionnaire analysis. The players are asked to express a preference between two variants of the video game. This approach of eliciting emotion is named comparative affect analysis and it was first introduced by Yannakakis and Hallam [3], [4]. Affective models are then derived using preference learning techniques [5], [6] matching user reported preferences and features from physiological signals measured during the game session. We assume that physiological responses are not task dependent since the level of physical activity required for the interaction with our game is constant during the session. As presented by Yannakakis [7], the preference model can be learned using different computational methods. In this paper, we propose a different linear model obtained with Linear Discriminant Analysis (LDA) [8], which shows performance analogous to other models presented in literature [9], but lower complexity and lower computation demand.

In the next sections, we first give an insight into the state of the art, then we introduce the experimental protocol designed for our experiments, finally we present the methods we adopted for preference modeling and the results obtained.

A. State of the art

With the purpose of determining criteria that contribute to player satisfaction, Yannakakis and Hallam [10] proposed two techniques for modeling player satisfaction in real-time. They assume that player-opponent interaction primarily contributes to the entertainment in a computer game. Therefore, metrics based on qualitative considerations of what is enjoyable from in-game performances (e.g., time before the player loses a life) have been considered as indicators of the level of interest.

Another approach consists in modeling the entertainment by following the theoretical principles described by Malone [11], and concepts related to the Theory of Flow [12]. Qualitative factors such as challenge, fantasy and curiosity are the ones that, according to them, mostly account for player entertainment. Quantitative measures for challenge, curiosity and flow state can be derived from an empirical analysis of player responses to game mechanisms. All the

Politecnico di Milano IIT Unit, Dip. di Elettronica e Informazione, via Ponzio 34/5, 20133 Milano, Italy. E-mail: {tognetti,garbarino,bonarini,matteucci}@elet.polimi.it.

978-1-4244-6297-1/10/$26.00 ©2010 IEEE

metrics evaluated in the mentioned papers are specific for a given class of games (e.g., predator/prey, racing, football), therefore a general model can not be determined with these approaches. Moreover, player responses can be affected by factors such as user experience and motion or perception skills, which can be loosely related to enjoyment, while physiological responses seem to be better indicators.

Under a different perspective, emotion has been investigated in the past by many researchers, including philosophers, psychologists, sociologists, psychophysiologists, and engineers. Results from psychophysiological studies [1], [13] indicated that relationship between the stimuli presented to a person and observed physiological reactions may exist. Grounded on these findings, people working on Affective Computing aimed to design human-machine interfaces, with emotion recognition abilities, for real life applications [14], [15], [16]. Recently, this research line has been extended to video games in which experiments can be performed with a good trade-off between reality and control.

In Mandryk et al. [17], statistically significant correlation has been claimed between GSR and reported fun (from questionnaire) in adults playing video games. Other physiological signals such as jaw electromyography, electrocardiography and respiration have been analyzed, but they resulted not to be correlated to the reported enjoyment. A fuzzy model inspired by psychophysiology theory is introduced by Mandryk and Atkins [18]. They reported that high values of HR and GSR together with a smile detection from an electromyography (EMG) in jaw are correlated to high values of arousal and positive valence.

Rani et al. [19] used psychophysiological measures such as HR and GSR to discriminate anxiety level and adjusted a Pong game to respond accordingly. In the approach proposed, they handle the problem of enjoyment maximization by appropriately minimizing the anxiety level.

Tijs et al in [20] showed values of skin conductance, HR and respiration to be statistically correlated to different difficulty levels of PacMan. The correlation with enjoyment state of a player is not directly calculated but it is inferred from a questionnaire analysis. Prediction models based on physiological state (HR, GSR) have also been proposed for potential entertainment augmentation in computer games [21].

In this work, we have applied some of the techniques presented in literature for physiological signal analysis for video game enjoyment evaluation. Preference learning techniques have been applied as presented by Yannakakis [9].

II. Experimental setting

We propose a new gaming experimental protocol, tested on a car-racing computer game, in which affective computing techniques can be applied and validated. The protocol was designed to produce an affective computing benchmark dataset, that could be used also for further developments. This dataset is composed by physiological data, questionnaire answers regarding user data (i.e., game experiences and race preferences), game logs and 2 video camera recordings. 75 volunteers (60 males and 15 females) aged from 18 to 30 years old (57 from 19-25, 18 from 26-30) took part in this study.

A. Task Design

The cognitive task in the experiment concerns playing a video game. This makes it possible to reach a high repeatability and a high level of involvement among participants. TORCS [2] was chosen as reference game for the following reasons: it is a video game that requires the player to be sitting in front of a computer, therefore subjects experiment emotionally different situations characterized by a similar physical activity and the effects of movement artifacts on acquired data are negligible differently from what happened in [9] where the subjects had to move and jump; this game is an open source project, therefore, it has been possible to implement custom logging and AI for opponent drivers; it is easy enough, even for an inexperienced player, thus the game experience can be kept as homogeneous as possible among subjects involved in the experiment.

During a game session, each participant played 7 races versus one computer driver that is the only opponent during the race. The opponent skill has been changed among races considering that this has a high potential impact on player emotional state, and that it can be easily adapted in a real-time affective loop to maintain the enjoyment level on the player according to the general principles of game engagement proposed by Malone [11].

Three classes of game scenarios have been considered and a customized opponent driver has been implemented to match the skill of the player. It modulates its speed to keep a given distance from the human driven car. We call W (Winner) the driver that is more skilled than the player and that has the goal to keep a distance of +100 m (relative distance between the cars within the current track) from the player. C (Challenging) is the driver that is as skilled as the player and tries to keep a distance of 0 m from the player. Finally L (Loser) is the driver that is less skilled than the player and that keeps a distance of -100 m from the player. According to a priori considerations, the second variant of the race could be considered really challenging, and more interesting for the player. Race parameters such as type of track, environmental details, car model, and number of opponents have been chosen to keep the game easy to play and to make the opponent skill being the main difference among races.

B. Experimental Protocol

Most of the choices in the experimental protocol have been made to maximize the focus of the player on the task. The environment where the experiment took place has been conceived with the purpose of isolating the player and maximizing the game immersion so that no external event could influence the subject physiological state. The setting was a small room with a computer placed on a desk. The player was sitting in front of the monitor and was interacting with the computer through standard mouse and keyboard.

No other people were in the same room and the operator monitored the experiment from an external site.

Ahead of the experiment, all participants have been asked to fill out a general questionnaire, presented in computer-based form, and used to gather information about their experience with video games, game preference, TORCS prior knowledge, and personal data such as age and handedness.

Then, Participants have been fitted with sensors to measure peripheral physiological activity as explained in Section II-C. The players were asked to wear a headphone to guarantee a deeper game involvement through race sounds. After this setup phase, players were left alone listening to a relaxing music (i.e., sounds from nature) with the purpose of both decreasing the stress and the initial excitement for the test and bringing all the subjects to a similar starting condition.

Cameras and physiological signal acquisition were started while the players were waiting. Any misplacing of sensors was checked by the operator from the external site by looking at the real time signals. The subjects have been instructed to minimize movements during the task to avoid artifacts. To increase subject involvement during the game, players have been told that they were competing for a prize. Prizes were given basing on a series of parameters including in-game performance, but also on physiological features, so that potential advantages of skilled player were reduced. Note that, from this moment on, to avoid the effect of covert communication [22], no further interaction between operator and subject occurred. The protocol was carried on by an automatic script on the computer that started each race and managed the questionnaire (see Section II-D).

After about one minute of relaxing music, the participants were asked to read the instructions and then, to start the trial by pressing a button. At the end of each race, starting from the second one, the participants were asked by a script to express, via a computer-based form, the preference between the race just played and the previous one. To minimize any potential order effect on physiological and self-reported data, each pair of game variants have been presented in both orders. The sequence of driver classes was as follows: W C L W L C W. With this sequence, all permutations pairs of classes W, C, L could be voted by the player once. The duration of each race was 3 minutes. This provided enough time to eliminate past race effects on physiological signals and to produce a new arousal level before the overcoming of boredom caused by excessive race length.

The total time of a session was about 30 minutes, i.e., 21 minutes (7 races × 3 min.) of racing and about 7 minutes of setup, question answering and resting.

C. Acquired data

In this protocol four types of data have been acquired: physiological data, questionnaire answers (presented in Section II-D), game logs and video camera recordings.

Physiological data were gathered using the ProComp Infiniti device [23]. This device captured 5 physiological signals: BVP, Electrocardiogram (ECG), GSR, Respiration (RESP) and Temperature (TEMP). A sample rate of 256Hz has been used except for ECG and BVP signals that were sampled at 2048Hz. The hand not used for interacting with the game was fitted with GSR, BVP and TEMP sensors. The 3 terminal ECG sensor were placed around the chest, as well as the RESP sensor.

TABLE I

EXTRACTED FEATURES

Feature of x	Description
x_m	Mean
x_v	Variance
x_{min}	Min value
x_{max}	Max value
x_D	Max − Min
x_{tm}	Time of Min value
x_{tM}	Time of Max value
x_{dT}	ΔT of Max and Min
x_{fd}	Mean Absolute value of first differences
x_{sd}	Mean Absolute value of second differences
x_{ct}	Trend
x_{acf}	AutoCorrelation function at 10s

Based on previous literature [24], [25], [26], [9], several derived signals have been extracted from the basic ones at the original sampling rate. Heart rate has been derived both from ECG (HR_{ecg}) and BVP (HR_{bvp}); magnitude (SM) and duration (SD) of signal variation has been derived from GSR; inspiration/espiration time ($inTime$, $outTime$), apnea in/out time ($apneaup$, $apnealow$) and respiration interval ($rTime$) have been extracted from respiration signal; upper/lower envelope of BVP (BVP_{up},BVP_l) and their difference ($BVP_d = BPV_{up} - BVP_l$) have been also computed.

A feature vector $F = [f_1 f_2 \dots f_D] \in \mathcal{R}^D$ has been finally obtained by the union of features described in Table I computed for each mentioned signal during each race. We assume that the first part of each race is subject to transitory phenomena due to the transition from a race to the next one. Thus, these features have been computed by considering only the last 60 seconds of each race.

A log file containing timestamp and some game status variables was saved during each race. Note that information regarding the TORCS state has not been used to obtain the results reported in this paper, but the timestamps have been used for synchronization between races and physiological data. Two video cameras recorded the environment in which the player acted too. A frontal camera captured the player's face, the second camera was placed at the top right back corner of the room, with respect to player, and captured the player actions and the game output from the monitor. All captured frames have been associated with a timestamp that is used for synchronization with the other signals. These data have not been considered in the analysis presented in this paper, but will be used by further research activities.

D. Questionnaire

Enjoyment preference between races have been collected. At the end of each race, the subject was asked whether he/she enjoyed more the last race or the previous one. A pairwise preference scheme (2-alternative forced choice: 2-AFC) has been used in self reports. 2-AFC offers a main advantage to acquire objective enjoyment: it normalizes the different conception of enjoyment among subjects and it allows a fair comparison between the answers of different subjects. Since we are concerned with finding a general model for the relationship between physiological features and reported entertainment preferences that generalizes over different players, 2-AFC is preferred with respect to other approaches, such as ranking [27].

III. Methods

Because of the lack of absolute ranked answers about player subjective enjoyment, canonical classification methods based on learning a target output is inapplicable since the target output is not defined.

Several techniques that learn from a set of pairwise preferences exist. Such algorithms are based on Gaussian processes [28], support vector machines (SVNs) [29], and evolving artificial neural networks (ANNs) [9]. In this work we focused on a linear approach developing a linear classifier, based on subject reported preferences and physiological features, which has similar performance, simpler structure, and lower computational demand.

A. Preliminar Statistical Analysis

A statistical analysis has been performed to understand the relationship between physiological features and reported enjoyment. In the first part of this analysis a Pearson's chi-square test [30] has been performed to establish whether the null hypothesis of independence between a feature and the reported preference could be accepted or rejected. When the null hypothesis is rejected, the feature and the preference can be considered not independent, but the strength of the true relationship is still unknown.

Then, a correlation analysis has been performed as proposed in [9]. We use Cohen's Kappa coefficient [31] to evaluate correlation between features and reported enjoyment defined as follows:

$$k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{1}$$

where $Pr(a)$ is the relative observed agreement between the user preference and the difference of the mean of a single physiological feature beween a pair of races; $Pr(e)$ is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each feature to be randomly correlated. If the preferences and the variation of mean values between pair of races are in complete agreement then $k = 1$ (or $k = -1$ in case of anticorrelated features). If there is no correlation (other than what would be expected by chance) then $|k| \leq 0$.

B. Preference Learning

Preference learning [6] is a technique that aims to predict the subject's preference among different elements. A subject expresses a preference $A \succ B$ (we say A is preferred over B) between two elements A and B when he/she is able to order them with respect to a personal preference criterion. Each element is characterized by a set of features $F = [f_1 f_2 \dots f_D] \in \mathcal{R}^D$. The goal of preference learning is to estimate a preference function P that respects the set of N constraints:

$$\textbf{if } A_i \succ B_i \; i = 1 \dots N \quad \textbf{then } P(F_i^A) > P(F_i^B)$$

Where F_i^A and F_i^B are the features vectors of elements A and B respectively in the i-th comparison and N is the total number of comparisons. Preference learning is a general approach and can be used to model subject's preference from physiological data. There are different techniques that can be used to estimate P depending on the function used. The Large Margin Algorithm (LMA) [29] can be used as linear preference learning classifier, and comes from Support Vector Machine theory (SVMs). This algorithm considers a linear combination of individual features F as emotional preference function $P(F) = FW^T$ where the weight vector $W = [w_1 w_2 \dots w_D]$ binds the user preferences to the physiological features. This method was first applied by Fiechter and Rogers [29] to a routing problem where the particular structure of the problem lead to a simplification of SVM approach to a linear problem. The main simplification was given by the fact that, in routing problems, preference decreases with costs represented by features F. Thanks to this hypothesis it was possible to add a new set of constraints $w_j \geq 0 \; j = 1, \dots, D$ that brought the quadratic problem into a linear optimization problem.

This method has been applied also to model subject preferences from physiological features as reported in [9]. However, features in our problem cannot be considered as costs and thus, the hypothesis $w_j \geq 0 \; j = 1, \dots, D$ would lead to a non optimal solution: weights greater than zero are given only to positively correlated physiological features i.e., negatively correlated features are ignored. We propose then to use a linear approach for preference learning based on Linear Discriminant Analysis (LDA) A.K.A Fisher's projection [8].

Given a set of N race pairs $R_i^A \succ R_i^B \; i = 1, \dots, N$ where the subject prefers R_i^A over R_i^B the goal is to estimate W in such a way that the user preference is preserved:

$$\textbf{if } R_i^A \succ R_i^B \; i = 1, \dots, N \quad \textbf{then } F_i^A W^T > F_i^B W^T$$

where F_i^A and F_i^B are the feature vectors associated to R_i^A and R_i^B respectively. We can rewrite the previous inequality as $(F_i^A - F_i^B)W^T = F_i^d W^T > 0$. Where F_i^d is the feature difference between preferred and not preferred races of pair i. We thus reformulate the problem of estimating W as a linear classification problem by considering the data set $X = \{x_i | i = 1 \dots N\}$, $C = \{c_i | i = 1 \dots N\}$, where $x_i = [F_i^d, -F_i^d]$ and $c_i = 0, 1$ (i.e., we assign class 0 to positive examples F_i^d and class 1 to negative examples $-F_i^d$). The problem can be solved with Fisher's projection, which finds a projection direction W in which classes are well separated. In this way, W can be used to predict one of the two classes by evaluating the inequality $XW^T < K$ (in our case K=0 since the mean of our data is 0).

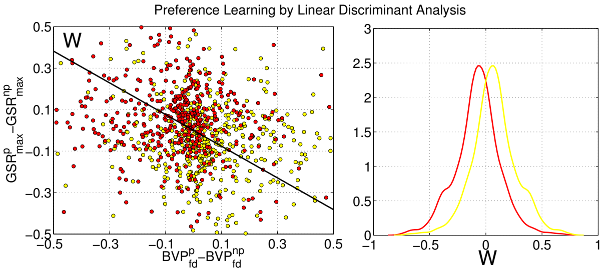

Fig. 1. Preference Learning by Linear Discriminant Analysis. A 2-D scatter plot of data from BVP and GSR is presented on the left. Sample points belonging to different classes have different colors. The black line W represents the direction computed by LDA in which data are projected. Probability densities of data by class with respect to the projection direction W is presented on the right. The direction W found by LDA is the one that best separates such densities.

To obtain Fisher's projection, we first define a within-class scatter matrix S_c as:

$$S_c = \sum_{c \in [0,1]} \sum_{x \in c} (x - \mu_c)(x - \mu_c)^T \tag{2}$$

where μ_c is the mean value of sample points that belong to class c. Then we define a between-class scatter matrix

$$S_b = \sum_{c \in [0,1]} N_c (\mu - \mu_c)(\mu - \mu_c)^T \tag{3}$$

where N_c is the number of sample points in the class c and μ is the mean value of all sample points. The best W that separates all the classes is the one that maximizes

$$J(W) = \frac{W^T S_c W}{W^T S_b W}. \tag{4}$$

Finding W is a one step process that requires much less time than running the optimization algorithm required by LMA.

Fisher's projection produces good classification performance when the distributions of data belonging to the same class are unimodal and classes are well separated. Figure 1 shows an example of application of LDA (based on real data obtained during our experiment) to classify the preferences of a subject by using two biological features BVP_{fd} and GSR_{max}. We can observe that both the distributions of $BVP^{p}_{fd} - BVP^{np}_{fd}$ and $GSR^{p}_{max} - GSR^{np}_{max}$ are unimodal but they are also partially overlapping, thus some of the preferences will be misclassified. Since all the data we are using in this work have an unimodal distribution similar to the one shown in Figure 1, the obtained performance depends on how well classes are separated.

IV. Data Analysis

In this section, the performance of the model for predicting reported preference is discussed. First we introduce the statistical analysis, then we present the results of the linear classifier based on LDA technique. The result will be finally extended by using different features selection methods.

A. Statistical Analysis

Results from the statistical analysis indicate that not all the features are dependent on the preference. This is confirmed by the fact that all p-values obtained from Pearson's χ^2 test are close to 0 ($< 10^{-5}$); thus, we cannot accept the hypothesis of independence. The second part of the analysis characterizes the kind of dependence between features and preference by using the correlation coefficients k. The results of this analysis are presented in Table II where correlation coefficients k and χ^2 values from Pearson's χ^2 test are shown. The table is organized as follows: features derived from the same physiological signal are grouped together. For each group, the 5 most correlated features (higher absolute values of k) are shown.

The physiological measurement that best correlates with reported user enjoyment is GSR, which achieved a correlation coefficient $k = 0.332$ followed by BVP with $k = -0.285$. This result indicates that players tend to prefer games in which features from GSR increase and features from BVP decrease. Similar findings have been reported by Mandryk et al in [32] where GSR values resulted correlated with fun. Features related to respiration reported in Table II, represent time differences between respiration events and have also high negative correlation with enjoyment (i.e., $apnealow_m$ has a $k = -0.234$). Thus, players preferred games in which breathing rate is increased (the time is decreased). Finally, good values of correlation have been obtained for HR ($k = 0.166$) and temperature ($k = -0.197$). Similar correlation results are reported also by Yannakakis and Hallam in [9] in their experiment on a physical playground. Our experiments have been performed on a computer video game that does not require physical activity, thus, features that best match user preferences in our work do not completely match the ones that were previously reported in [9].

B. Classification by Linear Discriminant Analysis

In the previous section, we have shown significant correlation between physiological features and the reported subject enjoyment . In this section, we present a quantitative evaluation of how each physiological feature can be used independently to predict the user preference. For each feature $f_j \quad j = 1 \dots D$, an emotional preference function $P(f_j) = f_j w_j$ is estimated through LDA techninque [8] as explained in the previous section. Note that, since we are using only one feature, the estimated w_j could be only -1 or 1 depending on the type of correlation. Given a pair of races R^A, R^B, the function $P(f_j)$ classifies R^A as more entertaining if $P(f^A_j) > P(f^B_j)$ where f^A_j and f^B_j are the $j - th$ feature of preferred and non preferred race respectively. The performance of the preference function is defined as the number of correct pairwise classifications with respect to the total number of pairs (i.e., Correct Classification Rate CCR).

To guarantee significant values of performance, a leave-one-subject out cross validation has been applied to the classification process as follows: data relative to one subject

TABLE II

Correlation of feature and preference. Higher absolute values of k mean higher correlation between the single feature and the user preference.

Bvp			HR			GSR			Resp			Temp		
f	k	χ^2	f	k	χ^2	f	k	χ^2	f	k	χ^2	f	k	χ^2
bvp_{fd}	-0.285	36.9	hr_{dT}	0.166	12.5	gsr_{sd}	0.332	49.7	$apnealow_m$	-0.234	24.6	$temp_{ct}$	-0.197	17.5
bvp_{sd}	-0.285	36.9	hr_m	0.145	9.6	gsr_{fd}	0.328	48.3	$apnealow_{min}$	-0.234	24.6	$temp_{tm}$	0.188	15.9
bvp_v	-0.244	26.9	hr_{min}	0.145	9.6	SM_{fd}	0.310	43.2	$rrate_{max}$	-0.229	23.7	$temp_v$	-0.159	11.6
bvp_m	0.232	25.2	hr_{max}	0.111	5.6	SM_{sd}	0.310	43.2	$inrate_m$	-0.213	20.4	$temp_D$	-0.161	11.8
bvp_{min}	0.232	25.2	hr_{tm}	-0.103	4.7	gsr_{max}	0.265	31.8	$inrate_{min}$	-0.213	20.4	$temp_{fd}$	-0.143	9.4

TABLE III

Correct classification rate using LDA algorithm as learning technique for single features.

Bvp		HR		GSR		Resp		Temp	
f	CCR	f	CCR	f	CCR	f	CCR	f	CCR
bvp_{fd}	0.638	hr_{dT}	0.582	gsr_{sd}	0.667	$apnealow_m$	0.616	$temp_{ct}$	0.596
bvp_{sd}	0.638	hr_m	0.571	gsr_{fd}	0.664	$apnealow_{min}$	0.616	$temp_v$	0.582
bvp_v	0.618	hr_{min}	0.571	SM_{fd}	0.656	$rrate_{max}$	0.611	$temp_D$	0.582
bvp_m	0.613	hr_{max}	0.556	SM_{sd}	0.656	$inrate_m$	0.604	$temp_{fd}$	0.573
bvp_{min}	0.613	hr_{tm}	0.551	gsr_{max}	0.636	$inrate_{min}$	0.604	$temp_{sd}$	0.569

have were and remaining data are used for training; the LDA is trained on the training data set and the performance of the model is tested on the data of the removed subject. These steps are iterated for each subject and the mean values are reported in Table III. The table is organized as follows: features derived from the same physiological signal are grouped together. For each group are shown the 5 features that give the best classification performance using LDA. Note that the values are consistent with the analysis of correlation reported in Table II: The best classification performance ($CCR = 0.667$) is achieved by GSR, which is the most correlated signal having the highest absolute value of k. BVP obtained a CCR=0.638, respiration reported CCR=0.616 and finally, for HR and temperature the CCR was 0.582 and 0.596, which are closer to 0.5 characteristic of a random classifier.

The results confirm that the Preference Learning approach can be successfully applied to model player preference in video games where the physical activity is kept as constant as possible over different type of games. Moreover, these results support the proposed experimental protocol as a successful way to produce reliable and replicable data for affective computing experiments.

C. Improvements by Considering Multiple Features

To improve single feature performance, a linear combination of all physiological features F by LDA has been used to predict subject enjoyment as explained in previous section. This is a generalization of the analysis performed over a single feature since a combination of features may introduce information useful for classification. Leave-one-subject out cross validation has been used to evaluate the performance. The LDA algorithm takes the full set of features as input and tries to find the best combination of weights W that separates classes. Correct classification rate achieved using all features is 0.633. This result is lower than the one obtained with the best single feature. This is likely, since some of the features might have introduced noise instead of adding useful information.

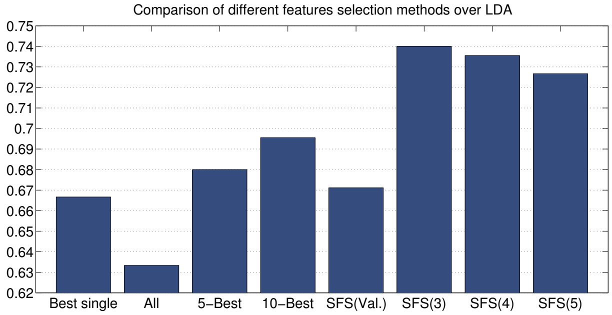

Fig. 2. Classification performance achieved with different subset of features: Best single feature, All features, Selection of best 5 or 10 features (5-Best and 10-Best), SFS on validation data (SFS(Val.)), SFS with pseudo-intersection (n=3,4,5)

Since including all features did not provide better performance than the best single feature, we tried to find a subset of features that performed better than single feature classification. We selected the first 5 and 10 features that individually gave the best CCR and the performance achieved was 0.68 for 5-best and 0.696 for 10-best. CCR is higher than the one obtained with the complete set likely because some of the noisy features are not included by this rough selection technique. A comparison of performances is reported in Figure 2

A better, still greedy, method for feature selection is the Sequential Feature Selection (SFS) [33] that finds the minimal subset of features that maximize the classification performance. It is a bottom-up algorithm where, at each step, an additional feature is added to the current feature set. The selected feature is added if its marginal value with respect to the classification function is positive. To evaluate the performance of a feature selection algorithm like SFS, we used an external cross-validation procedure. This avoids the overfitting of selection for the current dataset. A K-fold (K=15) cross-validation by subject has been performed. Data belonging to all the players in the original dataset Ds have been randomly split into K folds $Ds_i \quad i = 1 \dots 15$ (containing data of 5 subjects each). Folds have been then assembled into 15 data set containing training data $Ds_i^{tr} = Ds \backslash Ds_i$ and validation data $Ds_i^{val} = Ds_i$. The SFS has been executed K times independently over each different data set. For each iteration, the best feature selection with respect to the training data Ds_i^{tr} has been evaluated on validation data Ds_i^{ts} in order to estimate the selection performance on previously unseen data. Each selection found on training has been evaluated with the same leave-one-subject out cross validation method presented in Section IV-A. Each independent run selected the best subset of features that obtained different performance on validation data with mean value 0.671 of correct preference prediction over all the data set. Due to the different type of cross validation, this estimate of performance results the most general and least overfitted, hence the most significant. This result, labeled as SFS(Val.), is compared in Figure 2 with other feature selection methods.

The performance over validation gives an estimation of how well the classifier would perform with previously unseen data.However, with the presented SFS approach it is not possible to know which is the best subset of features that maximizes the CCR since each fold may yield to a different selection. To give an indicative measure of performance of the best general subset of features, a pseudo-intersection of the selections produced by each of the 15 independent runs has been performed. The pseudo-intersection merges the features that have been selected by at least n independent runs of SFS. We compared pseudo-intersection feature subsets with $n = 3, 4, 5, 6, 7$. We finally tested the obtained selections over the full data set with the previous described leave-one-subject out cross validation. In Figure 2 SFS(3)=0.74, SFS(4)=0.7356 and SFS(5)=0.7267 represent results obtained by the pseudo-intersections when features where selected by 3, 4 or 5 runs out of 15. The highest value is obtained with SFS(3) since it uses almost all the features that each fold have selected. This is the highest performance we achieved through leave-one-subject out cross validation by applying a linear function for preference modeling on these data.

In Table IV, we reported the features selected by some of the used methods. In the case of pseudo-intersection the performance decreases as much as we increase the number of folds from which features have been selected. This is due to the fact that pseudo-intersection exploits the information coming from each independent run. The less runs are considered, the more performance is overestimated, since we are using more data to obtain the selection. However, the more a feature has been selected from a fold the more it is an invariant measure of preference among subjects. This is the case of BVP_{fd}, GSR_{fd} and $outTime_v$ that have been selected at least by 7 out of 15 runs of SFS, so they are likely to be invariant features among subjects to predict enjoyment. Table IV shows also how the selections produced by SFS are different with respect to the 5-Best or 10-Best approach. SFS obtains higher performance, fixed the number of used features (i.e., SFS(4) vs 5-Best), since it removes variables that are dependent on the ones that have been already selected (i.e., GSR_{sd} and GSR_{fd} or SM_{sd} and SM_{fd}) and introduces other variables that can be more useful even if they do not produce high performance when used alone.

TABLE IV

SELECTED FEATURES BY DIFFERENT METHODS

Type	Selection	Performance
SFS(3)	BVP_{fd} BVP_{sd} GSR_{fd} GSR_{sd} SM_v SD_{fd} $inrate_m$ $outrate_v$ $apnealow_v$	74%
SFS(4)	BVP_{fd} GSR_{fd} SM_v SD_{fd} $outTime_v$	73.56%
SFS(5)	BVP_{fd} GSR_{fd} SD_{fd} $outTime_v$	72.67%
SFS(6)	BVP_{fd} GSR_{fd} SD_{fd} $outTime_v$	72.67%
SFS(7)	BVP_{fd} GSR_{fd} $outTime_v$	69.33%
SFS(8)	GSR_{fd}	66.67%
5-Best	BVP_{fd} GSR_{sd} GSR_{fd} SM_{sd} SM_{fd}	68%

By means of this analysis we have now a clear picture of which are the most relevant features that combined linearly can predict the reported preference of video game players with a CCR up to 0.74% on previously unseen data.

V. CONCLUSION

We have presented a way to identify the preference of players among different variants of a video game. We have proposed a new video game scenario in which 3 different game conditions have been obtained by controlling the opponent skill. The proposed scenario requires the player to be sitting in front of a computer, therefore the effects of movement artifacts on acquired data are reduced. Moreover, thanks to the adaptive controller, the game experience during the test has been homogeneous among different subjects. A data set composed by physiological data, questionnaire answers regarding user data, game experiences, race preferences, game logs, 2 video camera recordings have been acquired with the purpose of building a benchmark data set in which different Affective Computing techniques can be applied and validated.

We performed correlation analysis between physiological data and subject preference, showing that features from GSR have a high correlation with player reported preferences ($k = 0.332, \chi^2 = 49.7, p{-}value \approx 10^{-5}$). Similar results have been shown by Mandryk et al. in [32] and are slightly different from the ones presented by Yannakakis and Hallam in [9] probably due to the different nature of the task involved (computer video game vs physical playground). We have estimated a linear model of preference that maps a subset of physiological features to the preference level, trough a novel approach that makes use of Linear Discriminant Analysis (LDA). Results showed that this model is able to predict reported preference with an accuracy up to 0.74% on previously unseen data.

The analysis of questionnaires highlighted that for less than 42% of subjects there was a consistent agreement on the order of preference of game variants. That means that for them, a situation where the opponent is as skilled as the player was always preferred. However, the remaining players either did not answered coherently among different repetition of the same stimulus or they expressed a different order of personal preference. This interesting result indicates that the game preference is personal and it is hard to design a priori game experience (e.g., by changing opponent skills) that results suitable for everyone. This motivates the current research that aims to evaluate the enjoyment from biological signals without any assumption on players game preferences.

Player satisfaction from physiological data, is perhaps one of the most promising application area of affective computing. Classic and canonical games could be enhanced to adapt to the player affective states, and entirely new types of games could be created. Results from this experiments can be used as starting point for a follow up experiment in which game experience is modified in realtime with the purpose of keeping the player satisfaction to a high level.

Acknowledgments

The research activity described in this paper has been partially supported by IIT - Italian Institute of Technology.

References

[1] P. Ekman, R. Levenson, and W. Friesen, "Autonomic nervous system activity distinguishes among emotions," *Science*, vol. 221, no. 4616, pp. 1208–1210, 1983.

[2] "The open racing car simultaor website," http://torcs.sourceforge.net/.

[3] G. Yannakakis and J. Hallam, "Towards capturing and enhancing entertainment in computer games," *Lecture Notes in Computer Science*, vol. 3955, p. 432, 2006.

[4] G. Yannakakis, H. Lund, and J. Hallam, "Modeling children's entertainment in the playware playground," in *Proceedings of the IEEE Symposium on Computational Intelligence and Games*, 2006, pp. 134–141.

[5] J. Furnkranz and E. Hullermeier, "Pairwise preference learning and ranking," *Lecture Notes in Computer Science*, vol. 2837, pp. 145–156, 2003.

[6] J. Doyle, "Prospects for preferences," *Computational Intelligence*, vol. 20, no. 2, pp. 111–136, 2004.

[7] G. Yannakakis, "Preference Learning for Affective Modeling," in *Proceeding of the international conference on Affective Computing and Intelligent Interaction, ACII 2009*, Amsterdam, Netherland, 2009.

[8] R. Duda, P. Hart, *et al.*, *Pattern classification and scene analysis*. Wiley New York, 1973.

[9] G. Yannakakis and J. Hallam, "Entertainment modeling through physiology in physical play," *International Journal of Human-Computer Studies*, vol. 66, no. 10, pp. 741–755, 2008.

[10] ——, "Capturing Player Enjoyment in Computer Games," *Computational Intelligence (SCI)*, vol. 71, pp. 175–201, 2007.

[11] T. Malone, "What makes computer games fun?" in *Proceedings of the joint conference on Easier and more productive use of computer systems.(Part-II): Human interface and the user interface-Volume 1981*. ACM New York, NY, USA, 1981.

[12] M. Csikszentmihalyi, *Flow: The psychology of optimal experience*. Harper & Row New York, 1990.

[13] J. Cacioppo, L. Tassinary, and G. Berntson, *Handbook of Psychophysiology*. New York, NY: Cambridge University Press, 2000.

[14] R. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: analysis of affectivephysiological state," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[15] C. Lisetti, F. Nasoz, C. LeRouge, O. Ozyer, and K. Alvarez, "Developing multimodal intelligent affective interfaces for tele-home health care," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 245–255, 2003.

[16] J. Kim and E. Andrè, "Emotion recognition based on physiological changes in listening music," *IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 30 (12), pp. 2067-2083, December*, vol. 30, no. 12, pp. 2067–2083, December 2008.

[17] R. Mandryk, K. Inkpen, and T. Valvert, "Using psychophysiological techniques to measure user experience with entertainment technologies," *Behaviour & information technology(Print)*, vol. 25, no. 2, pp. 141–158, 2006.

[18] R. Mandryk and M. Atkins, "A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies," *International Journal of Human-Computer Studies*, vol. 65, no. 4, pp. 329–347, 2007.

[19] P. Rani, N. Sarkar, and C. Liu, "Maintaining optimal challenge in computer games through real-time physiological feedback," in *Proceedings of the 11th International Conference on Human Computer Interaction*, 2005, pp. 184–192.

[20] T. Tijs, D. Brokken, and W. IJsselsteijn, "Dynamic game balancing by recognizing affect," in *Proceedings of the 2nd International Conference on Fun and Games*. Springer, Berlin, 2008, p. 93.

[21] S. McQuiggan, S. Lee, and J. Lester, "Predicting user physiological response for interactive environments: an inductive approach," in *Proceedings of the 2nd Artificial Intelligence for Interactive Digital Entertainment Conference*, 2006, pp. 60–65.

[22] R. Rosenthal, "Covert communication in laboratories, classrooms, and the truly real world," *Current Directions in Psychological Science*, pp. 151–154, 2003.

[23] "Thought technology ltd., 2002. website," http://www.thoughttechnology.com/.

[24] A. Bonarini, L. Mainardi, M. Matteucci, S. Tognetti, and R. Colombo, "Stress recognition in a robotic rehabilitation task," in *Proc. of "Robotic Helpers: User Interaction, Interfaces and Companions in Assistive and Therapy Robotics", a Workshop at ACM/IEEE HRI 2008*, vol. 1. Amsterdam, the Netherlands: University of Hertfordshire, March 2008, pp. 41–48.

[25] S. Tognetti, C. Alessandro, A. Bonarini, and M. Matteucci, "Fundamental issues on the recognition of autonomic patterns produced by visual stimuli," in *Proceeding of the international conference on Affective Computing and Intelligent Interaction, ACII 2009*, Amsterdam, Netherland, 2009.

[26] R. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE transactions on pattern analysis and machine intelligence*, pp. 1175–1191, 2001.

[27] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 140, pp. 1–55, 1932.

[28] W. Chu and Z. Ghahramani, "Preference learning with Gaussian processes," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, p. 144.

[29] C. Fiechter and S. Rogers, "Learning subjective functions with large margins," in *ICML 2000: Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, pp. 287–294.

[30] H. Chernoff and E. Lehmann, "The use of maximum likelihood estimates in $\chi 2$ tests for goodness of fit," *The Annals of Mathematical Statistics*, pp. 579–586, 1954.

[31] J. Cohen, "Coefficient of agreement for nominal scales. Educational and Psychological Measurement." *Psychological bulletin*, vol. 20, pp. 37–46, 1960.

[32] R. Mandryk, M. Atkins, and K. Inkpen, "A continuous and objective evaluation of emotional experience with interactive play environments," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 2006, pp. 1027–1036.

[33] P. Pudil, J. Novoviová, and J. Kittler, "Floating search methods in feature selection," *Pattern recognition letters*, vol. 15, no. 11, pp. 1119–1125, 1994.