Biometrics and classifier fusion to predict the fun-factor in video gaming

Andrea Clerico[1], Cindy Chamberland[2,3,4], Mark Parent[2], Pierre-Emmanuel Michon[3,4], Sébastien Tremblay[2], Tiago H. Falk[1], Jean-Christophe Gagnon[5] and Philip Jackson[2,3,4]
[1]Institut National de la Recherche Scientifique
[2]Université Laval
[3]Centre Interdisciplinaire de Recherche en Réadaptation et Intégration Sociale
[4]Centre de Recherche de lInstitut Universitaire en Santé Mentale de Québec
[5]Ubisoft Québec

***Abstract*—The key to the development of adaptive gameplay is the capability to monitor and predict in real time the players experience (or, herein, fun factor). To achieve this goal, we rely on biometrics and machine learning algorithms to capture a physiological signature that reflects the player's affective state during the game. In this paper, we report research and development effort into the real time monitoring of the player's level of fun during a commercially available video game session using physiological signals. The use of a triple-classifier system allows the transformation of players' physiological responses and their fluctuation into a single yet multifaceted measure of fun, using a non-linear gameplay. Our results suggest that cardiac and respiratory activities provide the best predictive power. Moreover, the level of performance reached when classifying the level of fun (70% accuracy) shows that the use of machine learning approaches with physiological measures can contribute to predicting players experience in an objective manner.**

I. Introduction

The video game industry has seen in the past decades an exponential market growth. Only in the last 10 years in the United States the revenues coming from computer and video games increased imposingly, from 7.3 in 2004 to 15.4 bn $ in 2014. If the money spent on accessories and hardware is also considered, the figure grows to 22.41 bn $ [1]. Players value that their money are more cleverly spent with videogames compared to movies or music [2]. One reason for the preference of videogames over other means of entertainment is that having an immediate feedback keeps the engagement high. One factor of retention of players is whether their gameplay experience is positive [1].

Quantifying the extent to which a players experience is positive throughout his or her gameplay remains a challenge. Efforts have been made by numerous playtest laboratories in the gaming industry but there are still shortcomings to the objective evaluation of the player's fun. Most experiments are conducted using empirical and subjective methods (i.e. interviews, focus groups questionnaires). In the academia, systematic methods have been explored, in particular the use of affective computing technologies [3], [4] and machine learning algorithms, but with limited combinations of physiological measures and its weak association with the fun-factor. In this context, the field of automated affective states computation has grown with the aim of creating affective video games. Most of the work has focused on the comprehension of simple parameters of emotions and cognitive states studied in the affective Brain Computer Interface field [3], such as valence and arousal. Some authors also suggest that fun is not a unitary concept, which might add to the challenge of quantifying it. Lazzaros [5] model proposes 4 types of fun (e.g.: players seeking hard fun enjoy challenges, but players seeking people fun play for the social interaction). Poels, Kort and IJsselsteijn [6] suggest up to 9 dimensions to describe the video game experience. Evidence suggests that the different dimensions of fun are associated with distinct neurophysiological patterns [7]. These various reactions might increase the difficulty of using physiological measures to assess fun. Additionally, some authors suggest that fun is not directly measurable. Sweetser and Wyeth [8] suggest a model of enjoyment based on the flow theory [9] while others (e.g.: Calleja [10]) center their models on incorporation (i.e.: assimilation in the game while giving a sense of embodiment to the player). In a first effort to capture the relationship between several physiological and behavioral markers with the players experience, we chose to conceptualize fun as unidimensional since it is the easiest way for players to report their experience in relation to a videogame.

In the gaming literature, fun has been related to positive player reactions during a gameplay session. It has been linked to emotional experience but it is not considered as an emotion itself [11]. It is generally linked to different affective states, but as described in Pagulayan et al. [12], since games are intended to be fun, assessing fun implies assessing the overall quality of the game. The same authors [12] also point out that there might be a need to consider fun as being different in every user, thereby attributing a high contribution of individual influences to its assessment. Moreover, the work of Nacke et al. [13] mentions that when studying video games with physiological signals, there is a need to connect also other affective measures (e.g. behavioral responses) to establish relationships between the players experience and physiological responses (here we used a continuous measure of fun that will be described in Sec. II-B3). Thus, the fun-factor can potentially be analyzed in a study that combines objective measures such as physiological responses with subjective components qualifying the player's

experience [14], [15].

Computational models of fun have been designed in the past with the purpose of generating personalized game levels. For instance, in the work of Pedersen et al. [16], the authors were able to predict player emotions (e.g. fun, challenge and frustration) using preference learning and neuroevolution and the well known console game Super Mario Bros. A weighted non-linear computational model (e.g. artificial neural network) for reported emotions was constructed and the authors concluded that fun is the hardest dimention to model with a nonlinear perceptron and the least correlated with the features they extracted. Moreover, Shaker et al. [17] modeled player's fun value in platform games using a Multi Layer Perceptron Model. The authors obtained 69.66% of accuracy when modeling fun (using 58 features), but they also highlighted the limitation of post-experience analysis. For this reason they propose the use of physiological measures for further investigation, together with an increased number of features.

Several studies used physiological signals to quantify affective states during video game play ([18], [19], [14]). For instance, in the work of [15], the authors quantified emotional experience under the two dimensions of valence and arousal, to determine real-time emotional states. The latter were estimated using physiological signals such as Electrodermal Activity (EDA), Electrocardiography (ECG) and Electromyography (EMG). From the data of the two affective dimensions (i.e. valence and arousal), the authors were able to determine five distinct states during gameplay: boredom, challenge, excitement, frustration and fun. The problem that the authors revealed was that there are no guidelines for transforming assessments of arousal and valence into levels of fun in a continuous scale. Moreover, several affective states (i.e. boredom, frustration, challenge, anxiety, excitement) have been shown to correlate with ECG, EDA and EMG [20], but not a lot of effort has been made on the evaluation of the fun-factor itself.

Furthermore, using only signal processing techniques reduces a meaningful and natural interaction with the game. The introduction of a machine with automated emotional intelligence based on physiological responses would be able to learn negative and positive inputs and take care of player's need. Many studies have focused on detecting and learning emotional states, combining biometric signals and machine learning algorithms. In the work presented by Liu et al. [21], the authors studied different machine learning techniques for affective computing tasks. The authors used anxiety, engagement, boredom, frustration and anger as affective states and a questionnaire for self-reporting. The best performance was obtained using the Support Vector Machine (SVM) classifier in comparison with K-Nearest Neighbor, Regression Tree, Bayesian Networks. But developing video game affective systems, however, is a challenging task and many open problems and questions persist. For example, which physiological signal modalities should be combined to measure an affective dimension directly related to the level of player's level of fun? Which features convey such task effectively? Which characteristics of the classifier are better adapted to the task at hand? Is it really relevant to address the detection of fun as an individual factor?

In recent years, effort has been made for real time adaptation of video games using biometrics. For instance in the work of Rani et al. [22], the authors classified three level of intensity (low, medium and high) for different emotions (engagement, anxiety, boredom, frustration and anger) using a Pong game and anagram puzzle. Parameters of the game were manipulated to elicit the required affective response. Cardiac activity, EDA, EMG and skin temperature were used along with four different classification methods. The best accuracy result was reached with the SVM classifier and 86% of accuracy. The results are promising but the work focused on the study of affective dimensions that only give an idea of a general fun level. Another example of work that used a more complex and dynamic system, designed to adjust several parameters in the game over time based on the player's physiological signals, was conducted by Chanel et al. [23]. In this case the authors reached 53% of accuracy using an SVM classifier when discriminating three emotional states (boredom, engagement and anxiety) using peripheral signals as EDA, blood pressure, heart rate, Respiration (RESP), temperature and self reported labels. The number of features extracted might be determinant for the final result. In this last example there is a few features (only 14 features were computed from the signals). The extraction of a large number of features allows the machine learning algorithm to have access to a larger amount of information, otherwise not detectable in the case of a limited number of features.

The purpose of this paper is the design of a predictive model that is able to discriminate the fun experience of players, based on theirs physiological responses, as measured by indicators of ECG, EMG, EDA and RESP, together with a self-reported continuous measure of fun and the best classification system. Our main goal is to identify a physiological signature of the fun-factor associated with positive gaming experiences, together with the best classifier traits in order to create an adaptive video game according to prediction of players' affective and cognitive states. To achieve this, we used an innovative method that allows for continuous rating of fun during gameplay. This method provides an advantage over subjective measures by providing a high-temporal resolution of the fun rating instead of a single value for a given time period. Furthermore, fun ratings are converted to trends (i.e.: ordinal scales), which is considered to reduce biases associated with human self-rating of emotions [24]. Finally, the present study uses an off-the-shelf and modern video game, thus increasing its ecological validity and application to future work.

II. Methods and Materials

A. Participants

Sixty-two participants (5 women and 57 men) aged between 18 and 35 years (M = 25.9, SD = 4.9) were recruited from Université Laval and from Ubisoft Québec's volunteer

database to participate in a single two-hour experiment session. They all had prior game experience with the Assassin's Creed series, but had never played the specific title used in the current study. They all had normal or corrected-to normal vision and audition, and reported having no cognitive or neurological impairment. Participants received 20$ for their participation at the end of the experiment.

B. Apparatus and procedure

1) Computer game: Participants were asked to play the computer version of Assassin's Creed Unity -an action/adventure game developed by Ubisoft in 2014 (see a screenshot of the game at the top of Fig. 1)- with an Xbox 360 Controller. Two missions were specifically selected for this experiment: 'The prophet' and 'The escape'. The objective of this action-adventure game taking place in Paris during the French Revolution, played from a third-person view, is to complete pre-determined objectives to progress through the story. It is a non-liner gameplay, meaning that outside of the prefixed quests, the player can freely roam in the open world; thus giving the player more degrees of freedom compared to a linear gameplay.

2) Procedure: Participants read through a tutorial displayed on the computer screen describing the gameplay mechanics and explaining the procedure required to perform the different possible actions with the Xbox Controller. Participants were then familiarized with the game environment during a period of 5 minutes in which they had to complete seven objectives, all associated with the gameplay mechanics described in the tutorial (e.g., use a smoke bomb, climb up a building, assassinate an enemy while using a firearm). After successful completion of the objectives, a physiological resting baseline was recorded during a 3-minute period. Participants were instructed to remain calm and to refrain from moving while they were looking at a black fixation cross on a white screen and heard white noise via their headphones. Participants were then asked to play the first mission (presented in a counterbalanced order). The mission ended either when it was completed or after 15 minutes if participants failed completing it. The same procedure was repeated for the second mission.

Fig. 1: Graphic interface developed in order to give players a visual feedback of their fun ratings.

Fig. 2: USB controller (PowerMate, Griffin Technology) used to rate the level of fun.

3) Continuous measure of fun: After each mission, participants were required to watch a playback of their game session and to rate continuously the fun they felt during the game using a USB controller (PowerMate, Griffin Technology) shown in Fig. 2. This USB controller was an infinite control knob with no feedback (clicks) on the knob position. This controller was linked to a custom-made visual interface that allowed online graphic representation of the participant's evaluation of fun in real-time. As shown at the bottom of Fig. 1, the green areas correspond to positive levels of fun, while negative levels were depicted in the red areas. The level of fun was sampled at 30 Hz. Fun ratings were then transposed to a -100 to 100 scale for analysis.

4) Psychophysiological measurement: The player's physiological signals were collected during the two missions and during the replays (data from the replay were not used in this study). Electrodermal, cardiac, electromyographic and respiratory activities were recorded using a MP150 Biopac system (Biopac System Inc., Santa Barbara, CA). Electrodermal activity was measured using two pre-gelled electrodes placed on the palm of the left hand (the site was chosen in order not to have interferences with the controller). Cardiac activity was measured with three thoracic electrocardiogram electrodes placed in a Lead II configuration. The electromyographic signal was detected from the long abductor muscle of the right thumb with three pre-gelled electrodes placed on the right forearm. A respiration belt transducer placed around the player's chest measured respiratory activity (see Fig. 3 for the system's design). Cardiac activity was sampled at 1000 Hz whereas respiratory and electrodermal activities were sampled at 125 Hz using the Acqknowledge 4.3 data acquisition software. All psychophysiological signals were up-sampled to 1000 Hz (for synchronization purpose) and bandpass filtered (EDA 0-1 Hz, ECG 1-20 Hz, EMG 10-500 Hz and RESP 0-0.7 Hz). No further preprocessing has been conducted on the physiological signals since this study was designed as a prequel to a real time application. The signals and the self-assessment measure of fun were divided into epochs. The epochs were designed to last five seconds and an overlapping window of 2.5 seconds.

C. Analysis of fun

Three different situations of the self-reported measure of fun were identified. The first two corresponded to an increasing and a decreasing trend in the fun rating, respectively. For classification purposes, they were identified as the first two classes and labeled fun 'increasing' or 'decreasing' (+1 and -1 respectively). The third possible situation arose when studying a stable segment of fun rating, which provided two additional classes. We considered as high-stable fun the ratings that were stable but at a level above the average fun value of the whole mission, and low-stable fun at levels below the average level. These two other classes were labeled as fun 'above-average' or 'below-average' (+0 and -0 respectively). To summarize, out of each epoch, the classification analysis will output one of the four classes described above (see Sec. II-F below). Depending of the sign of the class (positive or negative) the software will then apply the modification in the game. In particular, with a positive output (+1 and +0), fun was either increasing or stable but over the average, thus in a adaptive scenario the game would not need any adjustment because the player is categorized as satisfied. However, in the case of negative classes (-1 and -0), a real time adjustment of the game would become essential to increase the fun.

D. Feature Extraction

A total number of 488 features were extracted from the four physiological signals as follows:

ECG The largest number of features is obtained from each electrocardiographic signal, such as 90 extracted features categorized into four different groups: spectral power with the fast Fourier transform (FFT) in multiple subbands, statistical components (average, min, max etc.), statistical features extracted from the analysis of separated parts of the QRS complex [25], and, finally, from the analysis of the Heart Rate Variability (HRV). Moreover, normalized versions of these 90 features were also computed. Normalization was performed based on the three minute baseline resting period, using:

$$ratio\xi_{norm,i} = 10\log\left(\xi_i^{epoch}/\xi_i^{baseline}\right), \quad (1)$$

with $\xi_{norm,i}$ being the 'i'-th feature. As such, a total of 180 ECG features were extracted.

EMG The electromyographic signal provided 53 features, that can be divided into three groups containing spectral, statistical evaluation and the sensitivity to change (first and second derivative) features. As with the ECG features, normalization was performed using baseline data, thus totaling 106 EMG features.

RESP From the Respiration signal, 74 features were obtained. These can be grouped into 3 different classes: rate of change, statistical and spectral analysis. Seventy-four additional features derived from the baseline normalization technique yielded a total of 148 features.

EDA Lastly, EDA provided 27 features (statistical, spectral and rate of change groups) from the band-passed signal. A total of 54 features were extracted including the baseline-normalized versions.

Fig. 3: Experiment set-up for dataset collection.

E. Feature Selection

Due to the large number of features extracted (i.e., 488), and particularly in cases of feature fusion techniques, such large number of features may result in classifier overfitting. As such, the so-called mRMR [26] feature selection algorithm is used. mRMR is a mutual information based algorithm that finds near-optimal features using forward selection with the chosen features maximizing the combined max-min criteria. Two criteria are applied as one: the maximum-relevance criterion (maximization of the average of mutual information between features and labels) and the minimum-redundancy criterion (minimization of the average mutual information between two chosen features). In the present work, 20% of the available data was set aside for feature ranking. The remaining 80% was used for classifier training and testing in a cross-validation scheme. Such partitioning corresponds to having 10 samples per class for testing and 40 samples per class for training. More details can be found in Section II-F.

The features were grouped into three separate sets. The first one included the non-normalized features coming from the four physiological signals, the second set consisted in the normalized ones, and the last group was constituted from a fusion of the first two. Feature ranking was conducted for the non-normalized feature set alone, for the normalized feature set alone and the combined feature set on a per-subject basis. Then, the first ten selected features were further ranked based on the number of times they were selected across all participants.

F. Classification

Support Vector Machine (SVM) classifiers have been used in the present work. Given its widespread use, a description

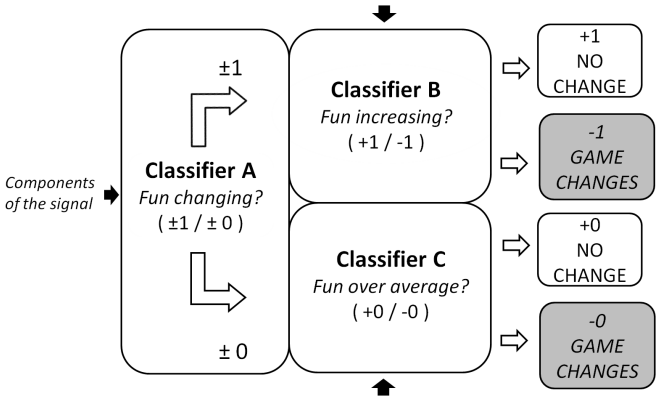

Fig. 4: Classification scheme designed to detect the level of fun and decide if the game needs real-time adjustment.

of the support vector machine approach is not included here and the interest reader is referred to [27] for more details. SVM classifiers are trained on three different (cascaded) binary classification problems, as depicted by Fig. 4, namely i.e. detecting fun changing/non-changing (classifier A), increasing/decreasing (classifier B) and above/below average (classifier C). The first, termed classifier A, discriminates between fun 'changing' or 'non-changing'. Based on this output, the second classifier to be used is decided. If the output of classifier A is fun 'changing', classifier B will successively discriminate between fun 'increasing' or 'decreasing'. If the output of classifier A is fun 'non-changing', then classifier C will discern between fun 'above-average' or 'below-average'.

In order to discover the best classification modality, two feature- and one decision-level fusion strategies are tested with the remaining 80% of the data. Regarding feature fusion strategies, the first one aims at testing the capacity to predict the output using the entire dataset within a 10-fold cross-validation scheme, whereas in the second case a per-subject classification with a Leave-one-sample-out (LOSO) cross validation scheme has been tested for comparison. In both cases, default SVM parameters have been used throughout our analyses (i.e., $\lambda = 1$ and $\gamma_{RBF} = 0.01$); moreover, a Radial Basis Function kernel was used and implemented with the Scikit-learn library in Python [28]. Lastly an optimally weighted decision fusion scheme has been tested [29]. First, the training data of the normalized feature sets for each physiological signal modality have been treated separately for both feature ranking and per-subject classification. Next, based on the performance achieved, a weight has been determined for the four signal modalities. According to the decision fusion technique used, the parameter t_i is the achieved performance for a particular modality, on the training dataset, such that the sum across all modalities equals unity [29]. The t_i parameter is calculated as follows:

$$t_i = \frac{A_i}{\sum_{i=1}^{N} \alpha_i A_i} \tag{2}$$

where, A_i is the accuracy obtained training the dataset belonging to a particular modality, N is the number of modalities and α_i are the weights corresponding to each modality ($\sum_{i=1}^{N} \alpha_i = 1$). Optimally weighted decision fusion relies on optimal weights for each of the four modalities which are obtained calculating the α_i values that result in the best performance on the training set.

TABLE I: Percentage of participation of each physiological signal (ECG, EMG, EDA and RESP) for the classification schemes (classifiers A, B and C), as well as for the three methods together.

	Classifier A	Classifier B	Classifier C	Total
ECG	39 %	33 %	23 %	32 %
EMG	14 %	21 %	25 %	20 %
RESP	28 %	31 %	36 %	31 %
EDA	19 %	15 %	16 %	17 %

III. Results and Discussion

A. Feature Ranking

Here, only the feature ranking analysis conducted on the normalized feature set is reported as it resulted in the highest accuracy. Table I shows the percentage of features used to reach the best accuracy from each of the four signals (EDA, ECG, EMG and RESP) for each of the three classifiers separately and for the total. As can be seen, ECG and RESP play a relevant role, representing almost two thirds of the total amount of top-ranked features.

An in-depth analysis on the features ranked by the mRMR algorithm has been conducted in order to understand the most relevant signals and the top-ranked features. Across all physiological signals, two thirds of the contribution comes from ECG and RESP, with a peak of 67% in the case of classifier A. Moderate importance can be attributed to EMG

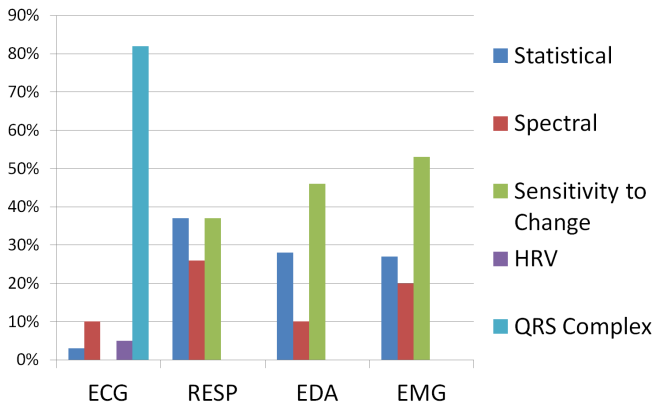

Fig. 5: Different contribution of the groups of features extracted (see Sec. II-D) for each signal: ECG (spectral, statistical, HRV and QRS complex), EMG, EDA and RESP (spectral, statistical and sensitivity to change).

and EDA, contributing on average one third of the top features. One potential reason for EMG playing a secondary role could be due to arm movement artifacts, as well as fatigue, which has been shown to affect EMG characteristics [30]. The low temporal resolution of EDA could also be a factor limiting the contribution of this signal. On the contrary, ECG and RESP have been previously connected to emotional expressivity, and their contribution is relevant to understanding inter-individual differences regarding players' physiological responses [31].

Fig. 5 shows the type of features more frequently selected by the feature ranking algorithm, for each of the four signal modalities. As previously explained, three groups of features can be differentiated for RESP, EDA and EMG (statistical, spectral and sensitivity to change), whereas for ECG four groups are identified (statistical, spectral, QRS analysis and HRV test). As can be seen, the analysis conducted on the first and second derivative of the signal is predominant for EMG and EDA. Whereas for the ECG, the study of QRS complex prevails over the other three groups. RESP, in turn, had the same number of features coming from the rate of change and statistical measures. This situation could be due to its low temporal resolution and its relationship with the ECG signal [32].

In order to give a better sense of the most used features, for each signal and for each classifier, the top-10 components have been analyzed. Within the ECG class, one third of the investigated top-features represent information about S and T waves, whereas another 20% is attributed to information measured from the segment between the P and Q waves. Moreover, two out of three classifiers had as their top-10 features the root-mean-square of the band-passed signal. Regarding EMG, 40% of the first 10 selected features were chosen from the analysis of the first derivative of the signal. Furthermore, by analyzing the repetition of the components, the presence of the power spectrum between 50 and 100 Hz has been ranked as top-feature for all the three classifiers. For RESP and EDA the most significant features (respectively 40% and 37%) were spectral related features. In particular, the power spectrum in the range of 0-0.7 Hz for RESP and 0-0.4 Hz for EDA.

B. Classification

Table II reports the highest accuracies and F1 scores [33] achieved with the individual and fused feature sets, for both all-dataset (10-fold validation) and per-subject (LOSO) classification schemes, as well as the number of features required to achieve such results. Values followed by an asterisk indicate significantly higher than chance according to an independent one-sample t-test ($p < 0.01$). As can be seen, the normalized feature sets achieves the best accuracy and F1 score results for the three classifiers. The best performance is obtained with a per-subject LOSO classification and in particular with the classifier C, reaching 70% discrimination accuracy.

Three main concepts can be inferred from feature-fusion classification results. First, fun is a dimension that can be detected by physiological signals, and in particular it is easier to measure the level of fun than its trend. When we compare the performance between the three classifiers, the best performance in accuracy is obtained when classifying the level of fun (fun 'above-average' and 'below-average' for classifier C), whereas performance is lower when using discriminating tendencies (fun 'increasing/decreasing' in classifiers A and B). Furthermore, a per-subject classification surpassed the one conducted using the full dataset. This effect can be related to the fun conceived as an individual factor [12]. A support to this idea also comes from the evaluation of the performances of the three feature sets. In fact, the normalized set outperforms or balances the level reached by the fusion set, thus showing the importance of per-subject normalization for automated fun assessment. Additionally, when comparing the number of features in a per-subject classification, two out of three classifiers (A and C) reached the best results with the normalized feature sets while using the minimum number of features. Classifier B with the fused feature set, on the other hand, required only one third of the components needed by the normalized set. For what concerns the all-dataset 10-fold cross validation classification, the best accuracy result was always reached with fewer features compared to the other classification schemes, but at the same time resulting in lower classification performance.

In turn, Table III shows the performances achieved with the decision level fusion scheme for the three classifiers. While decision level fusion of classifiers A and C did not lead to gains over simple feature fusion, decision level fusion of classifier B did improve the accuracy with a gain over the feature level fusion of 6 % for the normalized feature set and 4 % for the non-normalized set.

While decision level fusion with classifiers trained on these four separate modalities (ECG/EMG/EDA/RESP) resulted in an improvement only for classifier B, decision level fusion did result in further improvements, particularly when discriminating the increasing/decreasing dimension, thus suggesting the complementarity of the four physiological responses. Higher contribution rate is attributed to ECG (classifiers B and C). RESP holds more decision fusion weight in classifier A, whereas EDA contributed with the third highest weight. Decision level fusion was previously shown to be a useful tool for affective state recognition [33].

IV. Conclusion and Future Work

In this work, a triple-classifier system was tested for automated fun-level state recognition during video game sessions. Experimental results showed relevant performances in terms of accuracy. Feature level fusion has been proved to work better when detecting the level of fun, whereas decision level fusion when discriminating trends (70% and 57% respectively). Moreover, the importance of attributing an individual component to the players' fun-factor was demonstrated and essential physiological features are detected. Such findings suggest the importance of a robust adaptive video game based on personal characteristics of player's physiological signals, and capable of maintaining a high level of fun.

TABLE II: Performance comparison of SVM classifiers for different feature sets and feature-level fusion along with the required number of features needed to achieve such results. Asterisks indicate whether the accuracy or the F1-score distribution over subjects is significantly higher than chance according to an independent one-sample t-test ($p < 0.01$). 'Per-subj' corresponds to per-subject LOSO results, whereas 'All-dataset' to 10-fold cross-validation on the entire dataset.

	Classifier A								
	Non-Normalized Features			Normalized Features			Feature Fusion		
	Accuracy	F1 Score	No. Features	Accuracy	F1 Score	No. Features	Accuracy	F1 Score	No. Features
Per-Subj	0.59*	0.59*	87	0.60*	0.60*	69	0.59*	0.60*	70
All-dataset	0.55	0.55	35	0.54	0.54	35	0.55	0.55	32
	Classifier B								
	Non-Normalized Features			Normalized Features			Feature Fusion		
	Accuracy	F1 Score	No. Features	Accuracy	F1 Score	No. Features	Accuracy	F1 Score	No. Features
Per-Subj	0.54*	0.53*	53	0.54*	0.53*	91	0.54*	0.53*	36
All-dataset	0.51	0.50	18	0.50	0.53	22	0.50	0.51	4
	Classifier C								
	Non-Normalized Features			Normalized Features			Feature Fusion		
	Accuracy	F1 Score	No. Features	Accuracy	F1 Score	No. Features	Accuracy	F1 Score	No. Features
Per-Subj	0.69*	0.68*	54	0.70*	0.70*	69	0.70*	0.69*	101
All-dataset	0.55	0.52	30	0.55	0.52	40	0.54	0.52	20

TABLE III: Performance comparison of SVM classifiers for different decision-level fusion schemes. Asterisks indicate whether the accuracy or the F1-score distribution over subjects is significantly higher than chance according to an independent one-sample t-test ($p < 0.01$).

	Non-Norm. Feats		Norm. Feats		Feature Fusion	
	Acc.	F1 Score	Acc.	F1 Score	Acc.	F1 Score
Class. A	0.57*	0.57 *	0.56 *	0.56*	0.57 *	0.57*
Class. B	0.56*	0.55 *	0.57*	0.57*	0.56*	0.55*
Class. C	0.68*	0.69*	0.70*	0.70*	0.69*	0.69*

Game design can be significantly improved when conducting analysis of cognitive and affective neuro-ergonomics. To improve the performance of the proposed classifier system and automated game affective tasks, supplementary steps could be undertaken. First, the binary classification tasks performed here could be replaced by a regression task where the actual continuous value of the fun could be predicted. A second improvement would be to introduce a personalized calibration before starting the game, in order to train the classifiers based on personal traits of the subject. Third, classification performance could be improved by selecting optimal classification models by tuning hyperparameters (based on the individual signature) and explore different fusion strategies. Despite being innovative, the continuous rating of fun performed after the videogame session does have its drawbacks: it might not represent the actual fun that was perceived during the play and it relies on participants memory of their enjoyment which could have been forgotten and/or biased. Furthermore, participants did not have the possibility to rate fun on more than one dimension. Still, the outcomes of the present work can be applied to the development of a real-time adaptable intelligent game with application in console as well as online gaming. As a matter of fact, the present work is part of the FUNii (interactive intelligent) project [31] that aims at the development of an intelligent and interactive system capable of predicting the player's level of fun and adjusting the game to maximize that value.

ACKNOWLEDGMENT

This research was supported by a collaborative research and development grant to Philip L. Jackson and Sébastien Tremblay from the National Sciences and Engineering Research Council of Canada (NSERC) in collaboration with Ubisoft Québec. The authors would like to thank Jérémy Bergeron-Boucher, Marc-André Bouchard, Roxanne Poulin, Eric Arsenault, Sophie Regueiro for assistance in data collection and organization, as well as Ludovic Lefebvre for the support provided at the Ubisoft Québec studio.

REFERENCES

[1] D. E. S. A. Washington, Ed., *2015 sales, demographics and usage data: Essential facts about the computer and video game industry.*, Entertainment Software Association, 2015.

[2] Newzoo, Ed., *Global games market*, Games Market Research, 2015.

[3] R. W. Picard, *Affective computing*. MIT press, 2000.

[4] F. Portnoy, R. Aseron, M. Harrington, K. Kremer, T. Nichols, and V. Zammitto, "Facing the human factors challenges in game design a discussion panel," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 55, no. 1. SAGE Publications, 2011, pp. 520–524.

[5] N. Lazzaro, "Why we play games: Four keys to more emotion without story," 2004.

[6] K. Poels, Y. De Kort, and W. Ijsselsteijn, "It is always a lot of fun!: exploring dimensions of digital game experience using focus group methodology," in *Proceedings of the 2007 conference on Future Play*. ACM, 2007, pp. 83–89.

[7] C. Bateman and L. E. Nacke, "The neurobiology of play," in *Proceedings of the International Academic Conference on the Future of Game Design and Technology*. ACM, 2010, pp. 1–8.

[8] P. Sweetser and P. Wyeth, "Gameflow: a model for evaluating player enjoyment in games," *Computers in Entertainment (CIE)*, vol. 3, no. 3, pp. 3–3, 2005.

[9] M. Csikszentmihalyi, "Flow: The psychology of optimal experience," 1990.

[10] G. Calleja, "Revising immersion: A conceptual model for the analysis of digital game involvement," *Situated Play*, pp. 24–28, 2007.

[11] P. M. Desmet. (2003) Measuring emotions:development and application of an instrument to measure emotional responses to products.

[12] R. J. Pagulayan, K. Keeker, D. Wixon, R. L. Romero, and T. Fuller, "User-centered design in games," 2002.

[13] L. E. Nacke, "Games user research and physiological game evaluation," in *Game User Experience Evaluation*. Springer, 2015, pp. 63–86.

[14] R. L. Mandryk, K. M. Inkpen, and T. W. Calvert, "Using psychophysiological techniques to measure user experience with entertainment technologies," *Behaviour & information technology*, vol. 25, no. 2, pp. 141–158, 2006.

[15] R. L. Mandryk and M. S. 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.

[16] C. Pedersen, J. Togelius, and G. N. Yannakakis, "Modeling player experience in super mario bros," in *Computational Intelligence and Games, 2009. CIG 2009. IEEE Symposium on*. IEEE, 2009, pp. 132–139.

[17] N. Shaker, G. N. Yannakakis, and J. Togelius, "Towards automatic personalized content generation for platform games." in *AIIDE*, 2010.

[18] S. Tognetti, M. Garbarino, A. Bonarini, and M. Matteucci, "Modeling enjoyment preference from physiological responses in a car racing game," in *Computational Intelligence and Games (CIG), 2010 IEEE Symposium on*. IEEE, 2010, pp. 321–328.

[19] G. N. 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.

[20] J. M. Kivikangas, G. Chanel, B. Cowley, I. Ekman, M. Salminen, S. Järvelä, and N. Ravaja, "A review of the use of psychophysiological methods in game research," *Journal of Gaming & Virtual Worlds*, vol. 3, no. 3, pp. 181–199, 2011.

[21] C. Liu, P. Rani, and N. Sarkar, "An empirical study of machine learning techniques for affect recognition in human-robot interaction," in *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*. IEEE, 2005, pp. 2662–2667.

[22] 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*, vol. 58, 2005.

[23] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, "Boredom, engagement and anxiety as indicators for adaptation to difficulty in games," in *Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era*. ACM, 2008, pp. 13–17.

[24] H. P. Martinez, G. N. Yannakakis, and J. Hallam, "Dont classify ratings of affect; rank them!" *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 314–326, 2014.

[25] Y. Xu, G. Liu, M. Hao, W. Wen, and X. Huang, "Analysis of affective ECG signals toward emotion recognition," *Journal of Electronics (China)*, vol. 27, no. 1, pp. 8–14, 2010.

[26] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[27] B. Schölkopf and A. J. Smola, *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[28] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[29] S. Koelstra and I. Patras, "Fusion of facial expressions and EEG for implicit affective tagging," *Image and Vision Computing*, vol. 31, no. 2, pp. 164–174, 2013.

[30] N. Vuillerme, V. Nougier, and N. Teasdale, "Effects of lower limbs muscular fatigue on anticipatory postural adjustments during arm motins in humans," *Journal of sports medicine and physical fitness*, vol. 42, no. 3, p. 289, 2002.

[31] C. Chamberland, M. Grégoire, P.-E. Michon, J.-C. Gagnon, P. L. Jackson, and S. Tremblay, "A cognitive and affective neuroergonomics approach to game design," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 59, no. 1. SAGE Publications, 2015, pp. 1075–1079.

[32] J. Boyle, N. Bidargaddi, A. Sarela, and M. Karunanithi, "Automatic detection of respiration rate from ambulatory single-lead ECG," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 13, no. 6, pp. 890–896, 2009.

[33] S. Koelstra *et al.*, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.