A Hybrid Approach at Emotional State Detection:

Merging Theoretical Models of Emotion with Data-Driven Statistical Classifiers

Pedro A. Nogueira
Artificial Intelligence and Computer Science Lab
University of Porto, Portugal
pedro.alves.nogueira@fe.up.pt

Rui Rodrigues
INESC-TEC
University of Porto
Portugal
rui.rodrigues@fe.up.pt

Eugénio Oliveira
Artificial Intelligence and Computer Science Lab
University of Porto, Portugal
eco@fe.up.pt

Lennart E. Nacke
GAMER Lab
Ontario Institute of Technology
Oshawa, Canada
lennart.nacke@acm.org

***Abstract*—With the rising popularity of affective computing techniques, there have been several advances in the field of emotion recognition systems. However, despite the several advances in the field, these systems still face scenario adaptability and practical implementation issues. In light of these issues, we developed a nonspecific method for emotional state classification in interactive environments. The proposed method employs a two-layer classification process to detect Arousal and Valence (the emotion's hedonic component), based on four psychophysiological metrics: Skin Conductance, Heart Rate and Electromyography measured at the corrugator supercilii and zygomaticus major muscles. The first classification layer applies multiple regression models to correctly scale the aforementioned metrics across participants and experimental conditions, while also correlating them to the Arousal or Valence dimensions. The second layer then explores several machine learning techniques to merge the regression outputs into one final rating. The obtained results indicate we are able to classify Arousal and Valence independently from participant and experimental conditions with satisfactory accuracy (97% for Arousal and 91% for Valence).**

Keywords—Psychophysiology, emotion recognition, skin conductance, heart rate, electromyography, games, regression analysis.

I. INTRODUCTION

In the past two decades, video games have pioneered breakthroughs in various computer science fields, such as computer graphics, artificial intelligence (AI) and interaction techniques. These achievements were propelled by their massive popularity and steady growth, which largely outgrew the film industry's total revenues in various regions [1]. Due to this popularity, the increasing availability of game design tools and their high emotion elicitation potential [2], games and digital media in general have become popular in affective experience studies, among other areas, such as AI. However, despite the consecutive advances in the aforementioned fields, there is still a distinct lack of affective experience evaluation tools. These tools are not only needed to perform usability tests on traditional applications, but as a crucial and necessary first step for more complex experiments or emotionally-enabled applications (e.g. emotionally-adaptive movies, or evaluation of therapeutic treatment's efficacy over time) [3], [4]. Thus, given both the current immediate need and complex design of future applications it means that these systems must not only be sufficiently accurate (which is dependant on each scenario), but also present a general approach that can easily be adapted to various situations, while requiring minimal calibration and implementation overheads.

Thus, we are interested in how to develop a physiologically based emotion detection system that offers a balanced trade-off between its complexity, versatility, interpretability and accuracy. Furthermore, we are also concerned with the aforementioned calibration and re-implementation issues. This paper is structured as follows: Sections II justifies the need for such a system, while describing the various approaches taken in the literature. Section III discusses how the aforementioned approaches relate to our needs and justifies our chosen approach. Section IV describes the applied experimental procedure, participants, data collection and feature extraction process. Section V describes the developed method and obtained results for each classification layer. Finally, Sections VI and VII present our discussion on the obtained results and also describe how our current results improve on our previous work [22], not only by exhibiting improved classification accuracies, but by also providing a higher classification granularity that is adequate for most real-time applications.

II. EMOTIONAL RECOGNITION TECHNIQUES

A. Types of Emotional Recognition

Various taxonomies exist for emotional detection systems in a wide range of applicational domains, each one with its own dedicated literature. Within our current context – video games and digital media – Yanakakis et al. [5] segments these taxonomies into three distinctive types of approach: Subjective, Objective and Gameplay-based. Although the authors refer to them as types of player experience modelling, the base concept is that these approaches attempt to define how each player affectively interprets the gaming experience and is thus mainly a matter of nomenclature.

B. Subjective Modelling

Subjective player experience modelling (SPEM) resorts to first-person reports on the player's experience. While the most descriptive of the three, it is very difficult to properly analyse since reports tend to be plagued by experimental noise derived from player self-deception effects, memory limitations and intrusiveness – e.g. if questionnaires are performed during the experience itself [5]. However, when properly timed and framed, data collected through these methods can provide powerful ground truth data for more data-driven techniques.

C. Objective Modelling

Objective player experience modelling techniques (OPEM) attempt to explore the possible correlations between game events and physiological alterations. This approach usually employs multiple input modalities for real-time applications [5]. Objective modelling techniques are further divided into two types: model-based and model-free. While model-based approaches link physiological changes to popular models derived from emotion theories such as, for example Russell's arousal and valence dimensions [6] or Plutchik's Emotion Wheel [7], model-free techniques build their mappings based solely on user annotated data. However, systems may not rely on either of these two types independently. Hybrid approaches assume some type of correlation exists between physiological measures and affect (i.e. assume a pre-existing model), but seek to define the structure of the correlation function by analysing the available data. In fact, many known systems use the latter approach; they assume a theoretical model of emotion as their structure and build the mappings via the annotated data [8-11].

Various successful attempts have been made in the field of emotion recognition using the aforementioned types of objective modelling techniques – the hybrid ones being clearly the most popular ones. For instance, Chanel [8] was able to classify arousal using naïve Bayes classifiers and Fisher Discriminant Analysis (FDA), based on Electroencephalogram (EEG), skin conductance (SC), blood volume pressure (BVP), heart rate (HR), skin temperature (ST) and respiration rate measures. Complementary to this work, Leon [12] proposes the classification of valence in three intensity levels, using similar SC and HR measures and auto-associative neural networks. Similarly, Brown et al. propose a K-Nearest Neighbour approach at detecting valence using frontal alpha wave asymmetry indices in EEG readings [13]. In a more direct approach, Hazlett has found facial EMG can also be used to distinguish valence in gameplay experiences [14], having then further expanded these findings towards general software experiences [15].

In terms of simultaneous arousal and valence detection, Haag et al. [16] propose employing EMG, SC, ST, BVP, ECG and respiration rates to classify emotional states, reporting 89% accuracy for arousal and 63% for valence, with a 10% overall error margin. Under similar circumstances, Nasoz et al. have also successfully classified more complex emotional constructs, such as "happiness" and "fear" using a multi-modal approach [17].

Within the proposed line of low calibration approaches, the work by Vinhas et al. [9] proposes a system capable of measuring both arousal and valence in real-time, using the subject's SC response and HR derivative. A key factor of this work is that it introduced a continuous classification of arousal and valence. However, the method not only uses a limited set of psychophysiological measures, which limit its coping abilities to unseen scenarios, it also does not present sufficiently convincing results for valence classification using the HR derivative. Furthermore, as with the remaining literature, the issues arising from inter-subject physiological variations (see section V-A) are not effectively solved (although they are acknowledged).

Finally and similar to Vinhas, Mandryk presents an approach based on fuzzy logic that classifies EKG, EMG and SC measurements in terms of both arousal and valence [10]. Due to the ambiguous nature of physiological data, fuzzy logic presents itself as an elegant solution.

D. Gameplay-Based Modelling

According to Yanakakis et al., gameplay-based player experience modelling (GPEM) is driven by the assumption that player actions and/or preferences are also linked to his affective experience and as such can be used in detriment of more intrusive measures to identify his affective experience [5]. In essence, the rationale is virtually the same as in objective modelling. The annotated data obtained from the player's gameplay session is analysed and interpreted in light of some cognitive or behavioural model or theory. Since these approaches are not on our research's focus, we will limit their discussion. It is also important to note that while this type of modelling is the least intrusive of all three, it has been noted to result in low-resolution models of players' affective experience and models are often based on several strong assumptions between player behaviour and preferences [5].

In the field of gameplay-based modelling techniques Paiva et al. [4] has presented an empathic robot capable of reacting to the user's affective state, which is inferred through contextual information collected from the surrounding environment and interpreted according to an empathic behaviour model. Using the same type of approach, complex game aspects such as storyline have also been shown to be dynamically adaptable to individual players, in such a way that a pre-determined reaction is achieved [18]. Along the same research avenue, Pedersen et al. [19] have also shown the feasibility of constructing offline computational intelligence models capable of predicting optimal game parameter sets for the elicitation of certain affective states.

III. Employed Physiological Metrics

Although varied in their applicational focus, these approaches act as proofs-of-concept for the feasibility of real-time emotion detection systems in affective computing applications. However, when taking into consideration the previous discussion, it becomes apparent that building a generic or quasi-generic emotional state detection system that also requires minimal calibration is considerably harder for subjective and gameplay-based modelling techniques.

On one hand, SPEM techniques' noisy data collection and high subject and scenario dependency make it a very laborious approach. On the other hand, GPEM techniques are easier to generalise due to the re-utilization of extracted features but, in turn, require a high amount of domain specific knowledge, which is almost always coupled with strong assumptions and fairly complex adaptations in each new implementation. Furthermore, preferences must also re-learned for each new scenario, given they may not always be consistent.

Given these facts, it becomes clear why OPEM techniques are the most popular approach: data inputs and signal processing methods are fairly consistent between applications and, given the independent theoretical models, data interpretation is also maintained – i.e. minimal domain knowledge is required. In light of these conclusions, we have chosen a hybrid OPEM approach based on skin conductance, corrugator supercilii (brow) and zygomaticus major (cheek) facial electromyography and electrocardiographic metrics.

IV. Methods

In order to gather enough ground truth data to determine whether we could build an emotion recognition system for gameplay and multimedia content, we conducted a series of controlled experiments with a total of twenty-two healthy participants. In line with the good practices presented in the related literature [8], [9], [10], [12], [16], the applied protocol was initially tested and refined in an iterative prototypical cycle, using several pilot studies comprising a total of 10 participants. The results reported in this paper apply to the data collected and processed for the remaining twelve participants in the final iteration of the experimental procedure. Participants ranged from undergraduate students to more senior researchers and were aged 22 to 31 (M=24.83, SD=2.29), which constitute the demographic we limit our findings to. Seven of the participants reported playing video games at least monthly, while the remaining ones reported sporadic activity, whenever big titles came out.

A. Experimental Conditions

Each session was divided into three conditions designed for obtaining the necessary data samples to train our system. The first two conditions were aimed at eliciting extreme arousal and valence ratings: the first one being a 10-minute long session of relaxing music and the second one playing the horror video game Slenderman, by Parsec Productions. The third condition was aimed at eliciting neutral to mild reactions using a set of 36 images from the IAPS library [20], representative of its full gamut (i.e. low to high elicitation potential).

In each of the experimental conditions, the participant's SC, facial EMG and BVP readings were recorded. SC was measured at the subject's index and middle fingers using two Ag/AgCL surface sensors snapped to two Velcro straps. BVP was measured at the thumb using a clip-on sensor. Both these readings were made at the non-dominant hand [21]. Facial EMG was measured at the zygomaticus major (cheek) and the corrugator supercilii (brow) muscles and correlated with positive and negative valence, respectively [21]. Sessions were timed and conducted in a room with acoustic insulation and controlled lighting and temperature conditions. Participants were isolated in the room at the beginning of each section in order to limit contamination effects. The only human interaction was during the relaxation/briefing periods in between conditions.

B. Experimental Procedure

All participants were exposed to each condition in a fixed order. During our pilot studies, we found it was necessary expose the participants to the music and the horror videogame conditions before presenting the IAPS images; otherwise they

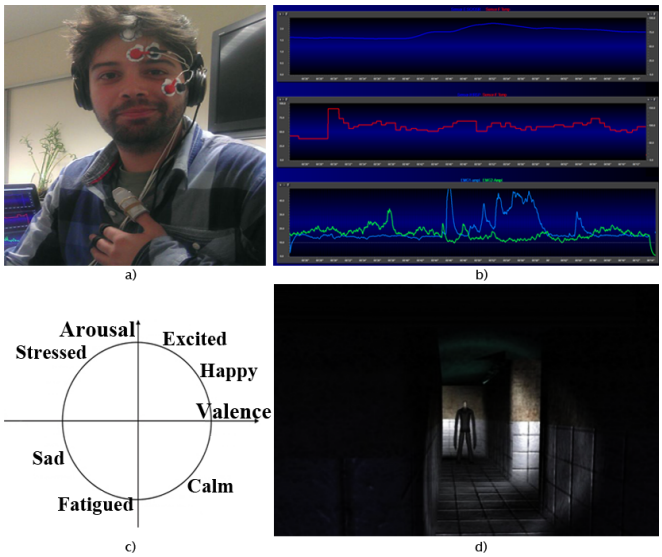

Fig. 1. *Experimental material: a) Participant sensor placement, b) Screen-capture of the physiological data, c) AV Space, d) Slenderman screen-capture.*

tended to rate the images relatively to one another, instead of on an absolute scale. By using the relaxing music and video game to delimit their responses, participants were able to rate the IAPS images in more absolute terms and a drastic reduction (~40%) in the ratings' standard deviation was indeed observed.

After signing an informed consent form, participants were briefed and underwent the full experimental protocol. Each condition was preceded by a relaxation period of approximately 5 minutes, through which baseline (averaged) values for each channel were extracted. The participants then underwent each of the experimental conditions, whilst reporting their affect ratings.

Experimental conditions were sub-divided into its constituent training samples, each with the same length as the event, plus a buffer length of 5 seconds added to both its boundaries. Regarding training samples, since we were interested in the participant's lowest emotional activation values for the relaxing music condition, in this case the training sample was equal to the whole condition. On the remaining two conditions, each event – images for the IAPS condition and gameplay events for the Slenderman condition – was isolated and independently rated by the participants.

Regarding the annotation procedure, participants rated the training samples immediately after their presentation. The exception to this was the Slenderman condition, since interrupting the gameplay activity to rate each event was not only intrusive; it also implied our physical presence, which could contaminate the experience. As such, for this condition, the gameplay session was recorded by a commercial frame grabber (Fraps) and analysed in conjunction with the participant in a post-gameplay interview. Participants reported their absolute maximum arousal and valence ratings in a 21-point Likert scale questionnaire ranging from -5 to 5 in 0.5 increments. We chose to ask participants to rate each event according to their absolute maximum because it introduced the least noise in the annotation process since it is harder for

individuals to rate their mean affect over a 10 or 20-second time window than to isolate a more intense emotional peak.

Each condition had an average duration of 10 to 12 minutes, with the exception of the terror videogame, which usually took 15 to 20 minutes. Overall, from setup to debriefing, the experiment had an approximate duration of 2 hours.

C. Apparatus

All of the sessions were performed on a laptop computer running Windows 7. The monitor was a 17" LCD display running at a resolution of 1920x1200 pixels. The gaming condition was recorded and synched with the physiological data at 30 Hz, using their starting timestamps (see Fig. 1). Physiological data was collected and exported using the Nexus-10 hardware by Mind Media.

D. Data Analysis & Feature Extraction

Regarding data analysis and feature extraction, the raw HR, heart rate variability (HRV) and SC readings were collected at 32Hz, while facial EMG was collected at 2048 Hz. HR and HRV (R-R intervals) readings were computed from the raw BVP readings. All of the physiological signals were then exported to a tab-delimited text file sampled at 32 Hz using the BioTrace+ software suite for future analysis.

Due to past hardware failures, the exported raw data was filtered for anomalies. Since no corrupt data was observed all of the collected data was retained for analysis. Regarding data pre-processing, HR readings were filtered using an irregularity detection method similar to [16]. This method only allowed a 25% variation of the HR signal at each new reading, based on its past variation over the previous 5 seconds. HRV was then recomputed from the corrected HR values. EMG amplitude values were extracted from the raw EMG readings using the Root-Mean Square procedure. Raw readings were corrected by subtracting their corresponding baseline values. Subsequent signal analysis revealed no additional filtering was necessary. Finally, sensor readings were smoothed over a moving window. Initially we employed an approximated Gaussian kernel for this smoothing process. However, this introduced unwanted signal distortions and was thus replaced with a traditional moving average kernel. Window sizes for HR and HRV were 2 seconds, 5 seconds for SC and 0.125 seconds for EMG, as suggested by Stern et al. in [21].

As previously mentioned, each condition was segmented into several training samples that were independently rated by participants using a 21-point Likert scale that ranged from -5 to 5. These were later converted to the 0-10 range (the same range as the IAPS ratings). Since sometimes participants had difficulty in differentiating similar stimulus using only the Likert scale, they were also allowed to provide a numeric answer, using the Likert scale as a reference. This implies that the gathered ground truth data can be treated as numeric, rather than nominal. Also, since participants were asked to rate their absolute emotional peaks, so were these values extracted from each channel's training sample. Overall, an average of 230 data points were collected per participant: the minimum values for the relaxing music condition (5 data points, one per channel), 36 samples for the IAPS images condition (180 data points), an average of 6 samples (30 data points) in the terror videogame condition – number of gameplay events varied across participants – and 3 neutral baseline samples (15 data points).

V. Detecting AV States

This section details how the annotated ground truth gathered in the three experimental conditions was used to detect the participants' emotional states. The developed method detects participants' arousal and valence ratings through a two-layer classification process (Fig. 2). The first classification layer simultaneously scales and correlates each input to an AV (arousal-valence) dimension by applying participant-specific regression models (i.e. each individual has his own set of regression models – one per physiological metric/AV correlation). This regression process generates a numeric output that is then fed to the second classification layer, which combines these predictions into one final rating by minimising their intrinsic error margins.

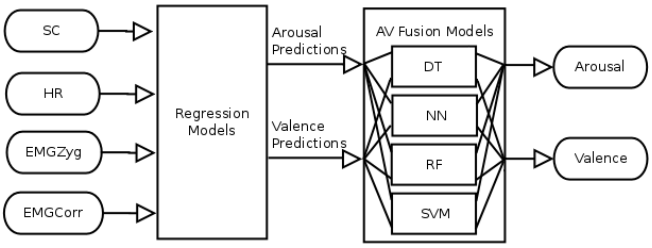

Fig. 2. *A high-level overview of the proposed system's architecture. Each of the physiological metrics is used to create two distinct sets of arousal and valence ratings. These ratings are then fed in parallel to one or more machine learning classifiers, which combine these rating sets into one final rating, for either arousal and valence, depending on the set provided.*

A. Physiological Input Regression Models

One of the most common issues with emotional recognition systems is the difficulty in obtaining an absolute scaling for the reported measures, an open issue that has been well documented in the literature [9], [10], [22]. This usually occurs due to either the used material's inability to elicit meaningful emotional alterations or by failing to annotate how each recorded event actually impacted the participant (i.e. assuming that a full emotional range elicitation occurred throughout the process). This issue is further aggravated by the considerable physiological differences (or physiological activation functions) displayed across individuals.

We tackled the first cause of the scaling problem (insufficient emotional elicitation) by exposing participants to a wide gamut of emotional content; ranging from relaxing music sessions, to a large representative sample from the International Affective Picture System (IAPS) library [20], and the psychological terror videogame, Slenderman. The second aspect of the scaling issue (unduly annotation) was addressed with a carefully planed experimental protocol – as discussed in detail throughout section IV.

The most common method of addressing both aspects of this issue – insufficient emotional elicitation and inter-participant physiological activation function variation – is by normalising the recorded values. However, this process

introduces a fairly high amount of noise, since it not only assumes that all participants experienced the same emotional ranges but, more importantly, that all of them experienced these ranges to their fullest extent. Thus, we approached this problem from a different perspective. Instead of a simple normalisation, we explored the correlation functions between each of the physiological metrics and the subjective ratings using regression models. Regression analysis has also been successfully used in emotional reaction identification to musical [23] and audio-visual stimuli in PET scans [24], thus further motivating our choice.

By using the participants' own subjective ratings (see Section IV.D) as our ground truth we were able to simultaneously take into account each participant's individual physiological activation functions and link them to the AV dimensions. We proceeded to explore the correlations of each annotated physiological channel to the AV dimensions and – apart from the HRV channel – confirmed each of the correlations referred in the literature. However, despite HR negatively correlating with valence, this was not observed for all three experimental conditions, since HR did not significantly fluctuate in the IAPS condition. As such, the results reported in Table I for the HR-valence correlation refer only to the training samples extracted in the first two experimental conditions, not all three as per the remaining described correlations. This finding is discussed in greater detail in [25] and [26], to which we refer the reader for further information. In this exploratory phase we used both linear and non-linear (third degree polynomial) models. Model complexity was kept in check using bidirectional stepwise regression. The procedure was based on their adjusted-R^2 values in order to minimise the effects of a large number of predictors on the polynomial models. The selected models were evaluated for correctness using 3-fold and 10-fold cross-validation, as can be seen in Table I.

TABLE I. FITNESS VALUES FOR THE CREATED REGRESSION MODELS USING 10-FOLD CROSS VALIDATION. MODEL COMPLEXITY WAS CONTROLLED USING STEPWISE-REGRESSION.

Physiological Channel	**AV Space Dimension**	**Adjusted-R2 Model Values (μ, σ)**	
		Linear	***Polynomial***
SC	*Arousal*	0.90 ± 3.8^{-2}	0.95 ± 3.0^{-2}
HR	*Arousal*	0.68 ± 7.1^{-2}	0.74 ± 8.9^{-2}
EMG_{Zyg}	*Valence*	0.84 ± 1.4^{-2}	0.92 ± 1.6^{-1}
EMG_{Corr}	*Valence*	0.83 ± 7.9^{-2}	0.95 ± 7.5^{-2}
HR	*Valence*	0.88 ± 1.0^{-1}	0.96 ± 6.4^{-2}

[a] *Regression models were obtained through the least-squares regression technique present in the R statistical package software.*

[b] *Presented results refer to the cases where statistical significance for the regression model was found ($p<0.05$). Furthermore, the stepwise regression process selected the third-order polynomial models for all presented correlations. As such, results refer to these models alone.*

Although there are multiple accounts of a linear correlation between SC and arousal [9], [10], [12], [16], [21], there is no common evidence that the remaining collected metrics correlate linearly with any of the AV dimensions. Thus, a statistical analysis between the generated linear and polynomial models was conducted and revealed that non-linear correlations are indeed supported by our data. One-tailed paired t-tests using the models' adjusted-R^2 values as within-subject conditions revealed statistically significant ($p<0.05$) differences between the linear and polynomial models for the following correlations: SC-arousal (t = -2.397, p = 0.035), HR-arousal (t = -2.393, p = 0.036), EMG_{Zyg}-valence (t = -2.396, p = 0.038), EMG_{Corr}-valence (t = -2.825, p = 0.018) and HR-valence (t = -3.297, p = 0.007). Closer inspection also revealed that although the polynomial SC-arousal models presented a significant improvement over the linear ones, they were marginally different from the latter and only presented a very small fitness improvement (5%), while the remaining models presented improved fitness values ranging from 9 to 14%. We thus decided to maintain the linear model for SC-arousal and opt for the polynomial models in the remaining ones.

B. A/V Rating Fusion Models

Having obtained each of the ratings from the regression models in the previous step it became necessary to fuse them to obtain a final AV rating. Given that we were aiming at a hybrid approach between a theoretically-grounded and ML-based method, and had already established the sensor mappings to our theoretical model's dimensions, we chose to leverage the benefits of a more data-driven technique: the ability to find more complex relationships in the data and better generalization capabilities across multiple subjects without prior strong assumptions on the data's structure or properties [5]. The basis behind this step was that the fusion models would be able to leverage each channel's error margins to decide how much weight each one should be assigned at each given time in the final rating (i.e. create an optimal combination policy for the regression models' ratings, according to their individual error functions).

Multiple ML classifiers were trained using the regressed data. Classifier diversity was promoted in the selection process with the intention of exploring various classification logics. Selected classifiers were: decision trees, single-layer perceptron neural networks, random forests and support vector machines. Regarding decision trees, the splitting criterion was determined using the normalised information gain at each node and trees were pruned to avoid over-fitting using the expected reduction in the Laplace error estimate. Neural networks were parameterised with a hidden layer comprised of 10 neurons and trained using back-propagation with a stopping threshold of 0.1 units. It has been suggested that while random forests do not require cross-validation techniques to obtain unbiased results, since they are a "random" statistical method, it is advisable to test their stability by training them with an increasing number of trees until convergence is met [27]. Thus, we trained models for each AV dimension with a number of trees ranging from 50 to 5000, following a regularly sampled, linearly increasing function and chose the models where convergence was attained. This was at 500 trees for arousal and 2000 trees for valence. The number of randomly preselected predictor variables used at each split was the square root of the number of predictor variables, as suggested by Strobl et al. [27]. Finally, the support vector machine classifiers were trained using a linear kernel type from the 'e1071' R library. Gamma and epsilon values were maintained at their default values (0.3 and 0.1, respectively). All classifiers were taught to rate the provided inputs (the regressed physiological data) according to the ground truth provided by the participants.

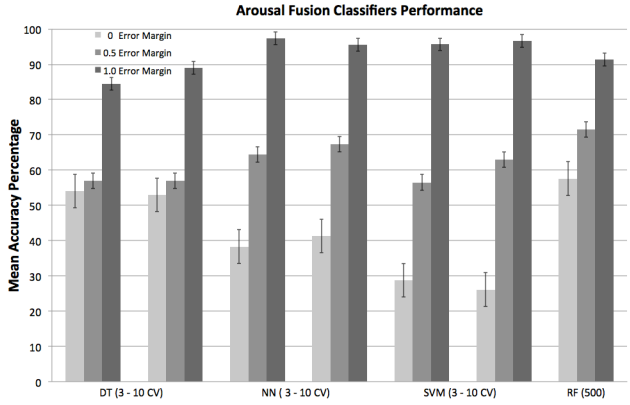

Fig. 3. *Means (± SE) of achieved predictive accuracy for arousal. Values are represented in percentage and separated by classifier type and cross-validation scheme (except, as previously mentioned, for the RF classifier).*

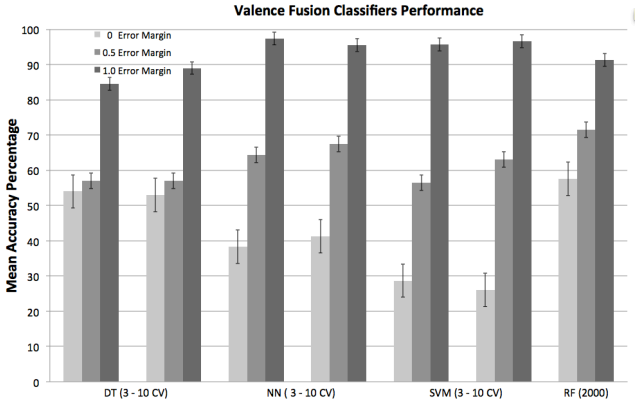

Fig. 4. *Means (± SE) of achieved predictive accuracy for valence. Values are represented in percentage and separated by classifier type and cross-validation scheme (except, as previously mentioned, for the RF classifier).*

TABLE II. AROUSAL FUSION MODELS ACCURACY RATINGS (%)

Error Margin	**Employed Classifiers (3-fold CV, 10-fold CV)**						
	DT		**NN**		**SVM**		**RF**
0.1	*54*	*22*	*38.3*	*41.3*	*28.69*	*26.1*	*57.6*
0.2	*57*	*57*	*64.4*	*67.4*	*56.5*	*63*	*71.52*
0.5	*84.5*	*89*	*97.4*	*95.5*	*95.65*	*95.7*	*91.3*
Avg Error:	*0.26*	*0.26*	*0.38*	*0.17*	*0.22*	*0.19*	*0.177*

TABLE III. VALENCE FUSION MODELS ACCURACY RATINGS (%)

Error Margin	**Employed Classifiers (3-fold CV, 10-fold CV)**						
	DT		**NN**		**SVM**		**RF**
0.1	*42*	*61.9*	*38.5*	*42.9*	*41.3*	*38.1*	*59.8*
0.2	*69*	*64.3*	*63.5*	*52.4*	*56.7*	*50.0*	*72.5*
0.5	*84.6*	*76.2*	*91.3*	*71.4*	*88.5*	*76.2*	*84.8*
Avg Error:	*0.35*	*0.61*	*0.34*	*0.46*	*0.35*	*0.41*	*0.263*

One final contemplation regarding these models is that, as previously mentioned, since the regression models have already accounted for the participants' own physiological traits, the models employed in the second classification layer are not participant-dependent. Thus, the amount of training data available for these methods was substantially higher. To test the accuracy of the built models, they were validated using 3 fold and 10 fold cross validation techniques. Since folds were already pre-calculated individually for each participant in the first layer, the folds for each classifier were computed by randomly merging one fold from each participant to generate the "population" folds. All of the presented classifiers were trained and validated using these same folds. While this served no significant computational purpose or gain, it avoided injecting unseen data into the second layer's training samples. Care was also taken to, as much as possible; equally divide the training samples across classes, so as to not bias the classifiers.

The obtained results are presented in Tables II and III and show that we were able to identify arousal with as much as 97% accuracy and valence as precisely as 91%, using neural networks, with an acceptable error margin of 0.5 points in the AV scale. Further considerations on the presented results can be found in the last paragraphs of the following section. Each predicted AV rating was evaluated through the following binary thresholding function:

$$T(c) = \begin{cases} 1, & if\ |c - \tilde{x}| \leq t \\ 0, & if\ |c - \tilde{x}| > t \end{cases}$$

Where c is the predicted AV rating by the classification function, $\tilde{x}$ is the ground truth data for the data sample under evaluation and t is the maximum acceptable error threshold for c to be considered correct. Following the observed error margins in the literature, t was set at values between 0.1 and 1.0. The average classification error (final line of Tables II and III) was computed as the average absolute difference between each of the predicted values $C = c_1, .., c_n$ and their corresponding ground truth annotations $\tilde{X} = \tilde{x}_1, .., \tilde{x}_n$. We consider that this range of threshold values provide a good overview of how the method performs with varying levels of granularity and, as such, represent its adequacy for the considered scenarios.

VI. DISCUSSION

Having presented the results for each component of our system, we now discuss our findings, along with some of the system's conceptual features. Our first conclusion is related to the observed correlation indexes between recorded physiological metrics and AV dimensions. Although HR did significantly correlate with valence in the relaxing music and Slenderman conditions, it did not for the IAPS images condition. While this does not impact on the system's performance for our case scenario, it leads us to the conclusion that whereas it is possible to estimate valence from cardiovascular measures, the same correlation may not always apply and thus should be confirmed prior to the system's calibration in new scenarios. We again refer the reader to [26], where we analyse and discuss this issue in considerably greater detail.

A detailed analysis of Tables II and III also reveals some interesting conclusions regarding each classifier's performance. Naturally, the best classification accuracies occur when the acceptable error margin is increased – in this case, 0.5 points in the AV scale. Based on the results presented in the literature, we considered this was an acceptable error margin that maintained an acceptable significance for our results.

Concerning arousal, the highest classification accuracy (~97.5%) is obtained using the single-layer neural networks, closely followed by the SVM classifier (95.6%). Also, there is a clear division between the performance of the NN and SVM classifiers and the DT and RF classifiers. This perhaps hints that the underlying concepts for the former ones are better suited at classifying this type of data. Methods exhibited overall consistent average classification error values, with the lower ones belonging to the NN and RF classifiers, respectively. Regarding valence, results are lower than for arousal, but the same general trends apply with neural networks presenting the highest classification accuracy (~91%). However, the previous divide between the NN / SVM and DT / RF classifiers does not hold. In fact, random forests present themselves as, in our opinion, the *de facto* best choice given their average classification error being the lowest between all classifiers (24% to 57% lower, in comparison). In sum, we consider that for the tested scenarios, since arousal and valence are distinct concepts, the underlying classification logic may differ and thus, the best choices for AV recognition appear to be neural networks and random forests, respectively.

Furthermore, a considerable advantage in using the NN, RF or SVM classifiers is that we are able to achieve a considerably fine-grained classification output, thus improving on our past results [25] and greatly contributing towards the method's real-world applicability. The most important drawback in employing these models is that they are created in a black-box fashion and are thus difficult to interpret and evaluate whether the constructed model is not overly complex due to the classifiers' expressive capabilities. This means a direct comparison between the theoretical models of emotion requires an indirect parallel validation approach, which may require multiple validation scenarios and a high additional workload.

Finally, two main factors contribute towards the system's adequacy in real-time scenarios: *a*) the relatively low computational cost involved in the NN and RF classification process for which they are known for, and *b*) the aforementioned high-grained classification output these same models are capable of producing. However, the latter factor implies that while our method is able to accurately measure participants' emotional states in terms of the AV space, it is also possible that it may not be adequate for all types of emotional recognition – e.g. evaluation of continuous time series, where a smooth evolution of the emotional signal is desirable. Tweaking the method to allow a continuous, smooth classification could be done using, at least, two alternatives: *a)* forcing participants to always state nominal ratings and increasing the number of classification classes – which, in our opinion, is largely unfeasible due to the ambiguous nature of the rating process and high amount of required sample data – or, *b)* replacing the ML classifiers with a simple weighting mechanism – perhaps one based on the normalised residual error functions of each regression model, as we have shown to be feasible in [26], albeit with a predictable and slight loss in accuracy.

As it stands, this approach can be used to accurately identify emotional reactions to specific events over large time windows, using a considerably fine-grained scale. Thus, it represents a contribution not only towards affect detection, but also towards affective experience analysis and evaluation practices.

VII. Conclusions

Emotional detection is not only a critical component of a large majority of affective computing applications, but also a highly complex task with no major consensus by the scientific community and largely lacks (complete) development guidelines.

In this paper we have presented a theoretical/data-driven hybrid, multi-layered method to interpret selected psychophysiological measures in terms of the arousal and valence affect dimensions. The exhibited results show that we are able to successfully address the recurring emotional scaling and participant physiological activation function normalisation issues present in the literature through combination of careful experimental design and a mixture of linear and low-degree polynomial regression models. Furthermore, this regression process allows us to keep the system complexity relatively low and humanly interpretable, while also generalising the second classification layer, which means that upon an initial calibration of the regression models, the method is participant-independent. Finally, the current version of our method has also further improved on our past results by performing the classification process with higher accuracy and a larger number of classes [25], all the while maintaining real-time classification capabilities and low sensor calibration requirements. Since we were able to build our system in a participant-independent fashion, it has shown to adequately generalise within the considered affective experiences and population, although subsequent tests on a larger population are needed for a strong generalisation proof outside these controlled experimental conditions and demographic.

We consider our method has the potential to contribute to the recently growing standardization of continuous emotional recognition development guidelines. We expect this to translate into a quickening of their implementation cycle, ultimately allowing for a higher percentage of the time to be allocated to the creation of more complex affective computing systems and affect-related experimental studies.

Acknowledgments

This research was partially supported by the Portuguese Foundation for Science and Technology (FCT) through the SFRH/BD/77688/2011 scholarship.

References

[1] "Board, Entertainment Software Rating: Video Game Industry Statistics," 2012. [Online]. Available: 0www.esrb.org/about/video-game-industry-statistics.jsp. [Accessed: 10-Mar-2013].

[2] L. Ermi and F. Mäyrä, "Fundamental components of the gameplay experience: Analysing immersion," in *Digital Games Research Association Conference: Changing Views - Worlds in Play*, 2005.

[3] M. Cavazza, D. Pizzi, F. Charles, T. Vogt, and E. André, "Emotional input for character-based interactive storytelling," in *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, 2009, pp. 313–320.

[4] I. Leite, A. Pereira, S. Mascarenhas, G. Castellano, C. Martinho, R. Prada, and A. Paiva, "Closing the Loop: From Affect Recognition to Empathic Interaction," in *3rd Int. Workshop on Affect Interaction in Natural Environments*, 2010.

[5] G. N. Yannakakis and J. Togelius, "Experience-driven Procedural Content Generation," *Transactions on Affective Computing*, vol. 2, no. 3, pp. 147–161, 2011.

[6] J. A. Russel, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[7] R. Plutchik, "A General Psychoevolutionary Theory of Emotion," *Emotion: Theory, research, and experience*, vol. 1, no. 1, pp. 3–33, 1980.

[8] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun, "Emotion Assessment : Arousal Evaluation Using EEG's and Peripheral Physiological Signals," in *Proc. Int. Workshop on Multimedia Content Representation, Classification and Security*, 2006, pp. 530–537.

[9] V. H. V. G. Moreira, "BioStories Geração de Conteúdos Multimédia Dinâmicos Mediante Informação Biométrica da Audiência," 2010.

[10] 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, Apr. 2007.

[11] A. Drachen, L. E. Nacke, G. Yannakakis, and A. L. Pedersen, "Correlation between Heart Rate, Electrodermal Activity and Player Experience in First-Person Shooter Games," in *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games*, 2010, pp. 49–54.

[12] E. Leon, G. Clarke, V. Callaghan, and F. Sepulveda, "A user-independent real-time emotion recognition system for software agents in domestic environments," *Engineering Applications of Artificial Intelligence*, vol. 20, no. 3, pp. 337–345, Apr. 2007.

[13] L. Brown, B. Grundlehner, and J. Penders, "Towards wireless emotional valence detection from EEG," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society*, vol. 2011, pp. 2188–91, Aug. 2011.

[14] R. Hazlett, "Measuring Emotional Valence during Interactive Experiences : Boys at Video Game Play," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 2006, pp. 1023–1026.

[15] R. Hazlett and J. Benedek, "Measuring emotional valence to understand the user's experience of software," *International Journal of Human-Computer Studies*, vol. 65, no. 4, pp. 306–314, Apr. 2007.

[16] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion recognition using bio-sensors: First steps towards an automatic system," *Affective Dialogue Systems*, 2004.

[17] F. Nasoz, C. L. Lisetti, K. Alvarez, and N. Finkelstein, "Emotion Recognition from Physiological Signals for User Modeling of Affect," in *Proceedings of the 3rd Workshop on Affective and Attitude User Modelling*, 2003.

[18] R. Figueiredo and A. Paiva, "'I want to slay that dragon' - Influencing Choice in Interactive Storytelling," in *Digital Interactive Storytelling*, 2010.

[19] C. Pedersen, J. Togelius, and G. N. Yannakakis, "Modeling Player Experience for Content Creation," *Computational Intelligence and AI in Games*, vol. 2, no. 1, pp. 121–133, 2009.

[20] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS)," 2008.

[21] R. M. Stern, W. J. Ray, and K. S. Quigley, *Psychophysiological recording*, 2nd ed. New York: Oxford University Press, 2001.

[22] F. Levillain, J. O. Orero, M. Rifqi, and B. Bouchon-Meunier, "Characterizing Player's Experience From Physiological Signals Using Fuzzy Decision Trees," in *IEEE Symposium on Computational Intelligence and Games (CIG)*, 2010, pp. 75–82.

[23] Y.-H. Yang, "A Regression Approach to Music Emotion Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448–457, 2008.

[24] A. S, W. E, N. P, H. J, M. L, S. H, and K. H., "Regression analysis utilizing subjective evaluation of emotional experience in PET studies on emotions," *Brain Res Brain Res Protoc.*, vol. 15, no. 3, pp. 142–154, 2005.

[25] P. A. Nogueira, R. Rodrigues, E. Oliveira, and L. E. Nacke, "A Regression-based Method for Lightweight Emotional State Detection in Interactive Environments," in *XVI Portuguese Conference on Artificial Intelligence (EPIA)*, 2013, p. (to appear).

[26] P. A. Nogueira, R. Rodrigues, and E. Oliveira, "Real-Time Psychophysiological Emotional State Estimation in Digital Gameplay Scenarios," in *14th Conference on Engineering Applications of Neural Networks (EANN)*, 2013, p. (To appear).

[27] C. Strobl, J. Malley, and G. Tutz, "An Introduction to Recursive Partitioning," *Psychological Methods*, vol. 4, no. 14, pp. 323–348, 2009.