Exploring Physiological Multimodality for Emotional Assessment

Joana Pinto
joana.f.pinto@ist.utl.pt

Instituto Superior Técnico (IST), Lisboa, Portugal

May 2019

Abstract

Many emotion recognition schemes have been proposed in the state-of-the-art. They generally differ in terms of the emotion elicitation methods, target emotional states to recognize, data sources or modalities, and classification techniques. In this work several biosignals are explored for emotion assessment during immersive video visualization, collecting multimodal data from Electrocardiography (ECG), Electrodermal Activity (EDA), Blood Volume Pulse (BVP) and Respiration sensors. Participants reported their emotional state of the day (baseline), and provided self-assessment of the emotion experienced in each video through the Self-Assessment Manikin (SAM), in the valence-arousal space. Multiple physiological and statistical features extracted from the biosignals were used as input to an emotion recognition workflow, targeting both user-dependent and user-independent classifications with three and two classes per dimension, respectively. Support vector machines (SVM) were used, as it is considered one of the most promising classifiers in the field. The proposed approach led to accuracies of 51.07% for arousal and 67.68% for valence in the user-dependent approach, and 69.13% for arousal and 67.75% for valence in the user-independent approach, which are encouraging for further research with a larger training dataset and population.

Keywords: Emotion Recognition, Biosignals, Virtual Reality, Support Vector Machines

This document was written and made publically available as a legal academic requirement in the evaluation of the MSc thesis in Biomedical Engineering of the author. The work described herein was performed at the Institute for Bioengineering and Biosciences of IST (Lisbon, Portugal), supervised by Prof. Ana Fred.

1. Introduction

Emotions represent a valuable source of information in the daily interaction within the Human civilization. Communication widely relies on the interpretation of affective states [1], since the expression of emotions of individuals can considerably change the sense of their messages [2].

The modelling of emotion presents two major challenges, these being the vague definitions and boundaries of emotion, and the methodology followed [3]. Emotions can be referred as a mental state or feeling that occurs spontaneously rather than a conscious effort [4], having two main types of manifestation: a mental response of emotion, combining subjective feeling and cognitive processes; and a bodily expression, which includes motor and physiological responses [3].

Emotion manifestation through physiological signals, or biosignals, is determined by the autonomic nervous system (ANS), which can hardly be consciously controlled by the intention of the individual, thus enabling more objective and reliable results [4, 5] than comparing to external responses, such as facial expression, speech or gestures. Moreover, biosignals can be assessed with wearable and non-intrusive sensing techniques [5, 6] supported by the continuous miniaturization of their sensors [7].

Over the past decade, automatic emotion recognition systems have seen significant developments within academia and industry alike. Applications include psychology, healthcare, education, marketing, gaming or service robots [2].

Various emotion recognition schemes have been proposed in the literature. They generally differ in terms of the emotion elicitation methods, target emotional states to recognize, data sources or modalities, and classification techniques. Given the importance of research towards user-independent emotion assessment, both user-dependent and user-independent scenarios will be explored in this study. This work will thus explore a multimodal approach for emotion recognition, based on a virtual reality (VR) emotion elicitation protocol and on the usage of SVM towards that purpose. The emotion recognition system will be tested to classify emotional states in terms of the emotion dimensions of valence and arousal, for both user-dependent and user-independent scenarios, considering three and two classes per dimension, respectively.

2. Theoretical Background

2.1. Emotion Theory

Various theoretical models of emotion have been proposed over the years. The discrete model con-

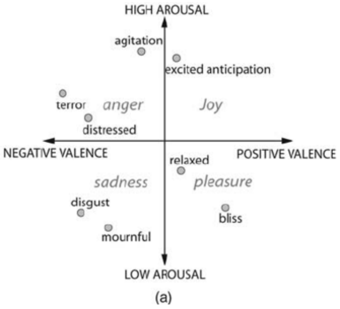

Figure 1: Two-dimensional model.

siders several basic emotions and claims their universality amongst cultures [8], existing considerable agreement in six emotions: happiness, sadness, surprise, anger, disgust, and fear [4]. The two-dimensional model introduced by Lang [9] is the most popular, characterizing emotions according to their valence and arousal [4], as illustrated in Figure 1. Valence represents pleasantness and ranges from negative to positive, while arousal indicates the activation level and ranges from low to high [9].

The definition of emotion is fairly subjective, influenced by cultural context, life experiences, and personality traits of each subject. However, according to the meta-analysis in [10], certain core components of emotions are universal and likely biological.

2.2. Data Sources

Emotion manifestations can occur both on an internal or external basis, being reflected by physiological activity or by physical responses. Several studies have thus explored various modalities, such as facial images and gestures [11, 2], speech [11, 2], or physiological signals [12, 13, 14, 15, 16, 17].

The group of methods that use facial images, gestures and speech can lack recognition accuracy, as they are not universal and depend on culture, gender and age [4]. Furthermore, when compared to physiological signals, those modalities are more susceptible to social masking, which can lead to wrong recognition of an emotional state. Regarding experimental settings, these modalities require special attention to lighting conditions and ambient noise for instance, which makes them challenging to be implemented in real time [4]. Their major advantages concern their feature extraction, believed to be easier compared to other modalities [4].

Alternatively, physiological signals can also be used, as they present patterns that are reflective of emotional expressions. In fact, the focus has shifted towards the usage of biosignals, from both the peripheral and central nervous systems, since they can provide continuous measurements and appear to be more efficient and reliable [11]. Furthermore, as those result from the activity of the Autonomous Nervous System (ANS), they cannot be easily triggered by any conscious or intentional control [4], which allows the researchers to overcome the social masking problem described for the previous group of modalities. When compared to visual data collection, e.g. facial expression, it is expected that the recording of biosignals is less disturbing than being "watched" by a camera [7]. Similarly, emotion recognition from speech has also been associated with critical difficulties in applications where users are listening to music or watching movies [7], as they are not expected to talk during these activities. Another advantage of using biosignals concerns the miniaturization of their sensors [7], happening to an extent that will soon enable their incorporation into everyday objects, accessories or clothes.

Challenges in the physiological signals processing are related to their subjective and complex nature, sensitivity to movement artifacts and inability to visually perceive emotions from the data [13]. Moreover, artifacts are a common problem in physiological signals, which can be corrupted by power line interference, motion artifacts or electrode contact noise [12]. These artifacts can be encountered in laboratory experiments and might become even more significant when in real-life applications.

The fusion of multimodal physiological signals are sought to enhance the efficacy of the process of emotion recognition. Comparing to emotion recognition based on a single biosignal, the fusion of multimodal emotion-related biosignals provides robustness, by eliminating anomalous changes not caused by emotional elicitation that occasionally may appear in a specific biosignal [18]. Furthermore, this fusion can boost the emotion recognition accuracy, since each individual modality can provide complementary information [19]. Recognition reliability can also be enhanced when taking into account the complementarity between several classifiers [1]. Hence multimodality has been increasingly and widely implemented for emotion recognition [19].

Considering that emotion is mostly expressed by means of internal bodily manifestations, namely, that ECG, EDA, BVP, and respiration are some of the most commonly found modalities in emotion assessment, these were selected for this study.

2.3. Support Vector Machines

SVM correspond to a set of supervised learning methods and its applications can range within several learning tasks such as classification, regression and outlier detection [20]. SVM are based on statistical learning theory and intend to determine the location of decision boundaries that produce the optimal separation of classes, being firstly proposed by Vapnik and Chervonenkis [21]. A detailed formulation of the algorithm can be found in [22, 23].

3. State of the Art

3.1. Emotion Elicitation

The study of emotion has been consistently associated to the ANS activity [6, 1]. The ANS activity is the result of two interconnected components, the psychological and the physiological. Regarding the emotion elicitation process, most researchers agree upon the point that emotions usually occur as a response to internal or external stimuli or events that are significant to the organism. As the emotion processes are correlated with the activity of ANS, whose manifestation is translated into the physiological signals of the subject, any experimental setup for the elicitation of emotions must be as natural and as close as possible to a real-life scenario, in order to obtain reliable data [1, 7].

As stated, the elicitation of emotions can be performed as an external stimulus [17], for which several kinds of techniques have been tested, from visual [24, 7], audiovisual [16, 12], audio [25], personalized imagery [26], recall paradigm [4], to a multimodal approach [13] elicitation. Concrete methods include images, sounds, music, video-clips and immersive videos.

Many studies have already confirmed the emotion elicitation ability of films, TV programs and imagery techniques [27]. Moreover, a new approach for emotion elicitation has emerged in the last two decades, through the usage of VR [27, 16, 28], which can be understood as an extension of the audiovisual film clips, able to add key benefits. Considering its promising features, VR will be the emotion elicitation method used in this work.

3.2. Virtual Reality

VR environments have been defined as those that, including synthetic sensory information, lead to perception of environments and their contents as if they were not synthetic [29]. VR has stepped into the field of emotion recognition as a more reliable emotion elicitation agent, and to investigate human behavior in well controlled designs [27, 16, 28].

Besides the immersion ability [27], VR can benefit emotion elicitation procedures in several methodological aspects. The use of VR is prone to increase the engagement of the participants within well-controlled experimental situations, while providing more realist scenarios, and by facilitating the replication of the same methods and procedures amongst the researchers [29, 28, 16].

Li et al. [16] made available a database to help establish standardized content and allow public access to a set of immersive VR videos, constituting an emotion elicitation tool for new studies, with reliable valence and arousal ratings [16].

3.3. Emotion Assessment

3.3.1 Self-Assessment

An important part of the emotion modelling process is the collection of the self-reported emotional states of the user (labeled data). This self-assessment might be seen as a ground truth, providing the researchers with relevant objective information, since the emotion actually felt by an individual can strongly differ from the expected one (i.e. from the target emotion elicited) [30].

The Self-Assessment Manikin (SAM) [31], is an acknowledged technique and the most widely used scale for the measurement of the emotional states [24], for instance in terms of valence and arousal.

3.3.2 Computerized Assessment

Machine Learning Algorithms

A wide range of machine learning methods has been used to infer emotional states [26], and both supervised and unsupervised techniques have been explored. The reason behind such a wide range of machine learning algorithms is that the best predictor will not be the same for all the datasets, thus various studies [24, 32, 33] have been conducted to test classifiers from several families, in different contexts and types of datasets.

Rigas et al. [24] designed an emotion classification system for three emotions (happiness, disgust, and fear) using four biosignals (facial EMG, ECG, respiration and EDA), comparing the accuracy of two types of classifiers: the Random Forests and K-Nearest Neighbors (K-NN). Their results showed statistically similar performance, concluding that K-NN performed slightly better, with an accuracy of 62.70%, largely inferior to what is reported in the other studies hereafter explored.

In their work, using physiological signals (EDA, ECG, BVP, and temperature) for emotion recognition, Jang et al. [32] conducted an analysis on the performance of four machine learning algorithms: LDA (linear discriminant analysis), CART (classification and regression tree), SOM (self-organizing maps), and SVM, all well-known approaches used in emotion recognition. Aiming at identifying three single emotions (boredom, pain, and surprise), their results showed that SVM was the algorithm leading to the best performance. In another study, the same group [17] reinforced the previous result, this time, seven emotional states (happiness, sadness, anger, fear, disgust, surprise, and stress) were under classification, yielding to an accuracy of 99.04% in the emotion classification by SVM, the highest amongst the five algorithms tested. The authors thus claim that these results should help new studies and lead to better chance of recognizing various human emotions by physiological signals [32], pointing SVM as the optimal algorithm for the data used [17].

Changchun et al. [33] have reached to the same conclusion. Their empirical study comparing four machine learning techniques (K-NN, Regression Tree, Bayesian Network and SVM) using physiological features (from ECG and facial EMG) found an advantage for SVM over the others [33], reaching a classification accuracy of 85.81%.

Most of the works only performed user-dependent emotion recognition [1, 7], even though a user-independent system would be more significant and with wider application [16]. Haag et.al [7] used a Neural Network to classify user-dependent emotions, split into arousal and valence, using ECG, EDA, respiration, BVP, EMG and temperature. Their results showed that the estimation of the valence was a harder task than the estimation of arousal, despite the overall results were considered good for both, accomplishing 89.7% of correct classification for arousal and 63.8% for valence. A similar result was found by Wagner et al. [14], whom also accomplished better results for arousal than for valence classification.

Nevertheless, the user-independent approach proposed by Li and Chen [16] yielded encouraging results, using four physiological signals (ECG, EDA, temperature and respiration) and adopting a canonical correlation analysis, achieving an accuracy of 85.3%. Similarly, Kim et al. [13] developed a user-independent system using short-term monitoring of physiological signals (ECG, EDA and temperature), by classifying the patterns with an SVM, yielding to the correct-classification ratios of 78.4% and 61.% (for 50 subjects), for the recognition of three (sadness, anger, stress) and four (sadness, anger, stress, surprise) emotions, respectively. These results suggest the feasibility and importance of a user-independent emotion recognition system.

Considering the evidence of classification accuracy of the several algorithms, SVM was the one selected in this work. Moreover, due to the importance of further research towards user-independent systems, this work will explore both user-dependent and user-independent classification scenarios.

3.3.3 Feature Extraction

Several physiologic-based features have been suggested for each biosignal. The most common for ECG and BVP are the heart rate (HR) and heart rate variability (HRV) [15, 17, 24, 3, 6]. In the frequency domain, the power in low frequency (LF) and in high frequency (HF) are commonly extracted from ECG [34, 35], with the LF/HF ratio and the sum LF+RF also being found [17]. Concerning the EDA, a wide range of features has been explored, although not being completely clear which of those provide more valuable information. Examples include the zero crossing rate [15, 36], average of the absolute derivative [15], Skin Conductance Response (SCR) [15, 17, 6] and Non-Specic Skin Conductance Response [15, 17], power spectral density, rise time (r_time), recovery time (rec_time) [36], initial skin conductance level (SCi), final skin conductance level (SCf) and their difference (SCi-SCf, a_var) [3, 6], and time and amplitude difference between the onset and peak of the SCR (amp_PO) [6]. For respiration signals, the respiration rate is the most common physiological feature [15, 6].

According to Martinez et al. [3], the features most typically extracted from biosignals are common statistical features, calculated on the time or frequency domains. Namely, the signal mean amplitude [24, 3, 34, 15, 7], the standard deviation [15, 3, 7], and the median [24] have been widely used in these biosignals. The root mean square of differences between RR intervals has been used for ECG [15], and the mean of absolute values of first differences has been also proposed for several biosignals [24]. Higher Order Statistics (HOS), including skewness and kurtosis, have been previously explored in ECG signals [37].

4. Methodology

4.1. Experimental Setup

Our setup, depicted in Figure 2, comprises the sensor modules for multimodal biosignal acquisition, a computer application that records data from these modules, and the VR setup for stimuli presentation. Biosignal data was collected using two devices based on the BITalino system [38], one placed on the arm and the other on the chest of the participants. The participants used the HP Windows Mixed Reality Headset and headphones, so as to fully immerse on the VR videos. Only the visual stimulus was included in the VR experimental setup.

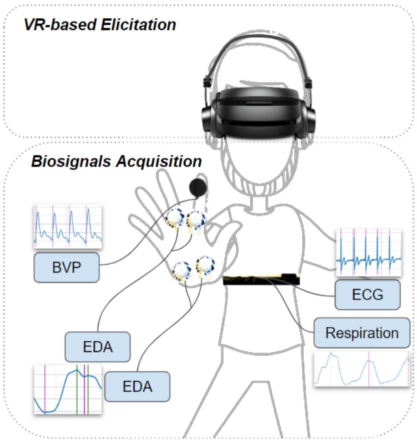

Figure 2: Experimental setup used in this work.

4.2. Experimental Protocol

Elicitation videos were selected with the goal of obtaining a representative range of emotions, while presenting good quality in VR. Content was selected from the database provided in [16], which has a mapping of the target emotion for each video, in the valence-arousal space. However, the database lacks videos for panic/fear and anger; these (3^{rd} & 5^{th}) were selected by independent research on YouTube. Table 1 depicts the complete experimental protocol.

Table 1: Experimental protocol summary, including: preparation; calibration; and elicitation period, with target emotions and corresponding ID in the VR video database [16]. Duration entries in parenthesis highlight the steps in which biosignals were acquired. A video sequence comprises the visualization of a neutral (black) screen, of the immersive video, and the SAM annotation.

	Target Emotion	**Duration (s)**	**ID in the Database [16]**
Informed consent and objectives of the study.	-	60	-
Annotation of the emotional state of the day.	-	30	-
Wearing the acquisition system and VR setup.	-	120	-
Adaptation time and final recommendations.	-	60	-
Calibration 1	Sadness	(30)	
Calibration 2	Anger	(30)	
Calibration 3	Happiness	(30)	
Calibration 4	Relaxation	(30)	
Video sequence 1	Boredom	5 + (43) +30	1
Video sequence 2	Joyfulness	5 + (250) +30	70
Video sequence 3	Panic/Fear	5 + (160) +30	-
Video sequence 4	Interest	5 + (65) +30	42
Video sequence 5	Anger	5 + (75) +30	-
Video sequence 6	Sadness	5 + (120) +30	6
Video sequence 7	Relaxation	5 + (210) +30	32

4.3. Signal Processing and Feature Extraction

This work uses a feature-based approach. Figure 3 illustrates the complete workflow, from the raw biosignal data to the prediction of the emotional state, and Table 2 provides an overview of the features used, selected building upon the related work.

4.3.1 Filtering

The most common noise sources affecting biosignal data are motion artifacts and electromagnetic interference [13, 4]. For the pre-processing of ECG signals, a finite impulse response (FIR) filter was used, with a passing band of 3-45 Hz [39]. In EDA signals, we followed the approach in [40], using a forward and reverse Butterworth lowpass filter with 5 Hz cutoff frequency [41]. Even though BVP sensors are usually highly susceptible to noise and artifacts, we used a BVP [1] whose plastic clip-on housing for placement on the finger to house the light emitter and detector allowed the minimization of interferences from external light sources. Nevertheless, the method used in [42] was applied and the signal filtered using a 4th order Butterworth bandpass filter with 1-8 Hz passing band. Concerning respiration signals, the approach used in [43] was followed, and the signals filtered using a 30th order FIR lowpass filter with cutoff frequency of 0.15 Hz, which revealed to be efficient in reducing the noise of the signal.

Table 2: Set of features extracted from each of the biosignals, split into (i) Physiological and (ii) Statistical features.

Biosignal	**(i) Physiological Features**	**(ii) Statistical Features**
ECG	heart rate, HRV, LF, HF, LF/HF, LF+HF	mean, std, ad, kurtosis
EDA	SCR, r_time, ampPO, rec_time, a_var	mean, std
BVP	IBI, heart rate	mean, median, maxAmp, kurtosis, skewness
Respiration	respiratory rate	mean, median

4.3.2 Segmentation and Outlier Removal

Given that several features depend on fiducial points within the signals, segmentation is an important part of feature extraction, for which the BioSPPy[2] library were used. Despite the filtering step, the segmented data may still be influenced by artifacts, resulting mainly from motion either from the subjects movement when browsing the immersive video VR space by looking from one side to the other, or from the accidental contact of their limbs with the chest mounted device. To mitigate these issues, outlier removal has also been considered.

Considering the periodicity of ECG, BVP and respiration signals, segmentation was performed at the cycle level. In the case of the EDA, there was detection of skin conductance response events.

Segmentation of ECG signals was performed as proposed by Hamilton [44]. A final step was performed towards removing abnormal ECG templates, by applying an exclusion algorithm based on the physiology of ECG waves proposed in [39].

For EDA signals, segmentation was performed using the method found in [13] to isolate SCR events.

In BVP signals, onset detection was based on the approach proposed in [45]. In this case, an outlier removal step was performed by computing the av-

[1] https://store.plux.info/bitalino-sensors/42-pulsesensor.html

[2] https://github.com/PIA-Group/BioSPPy

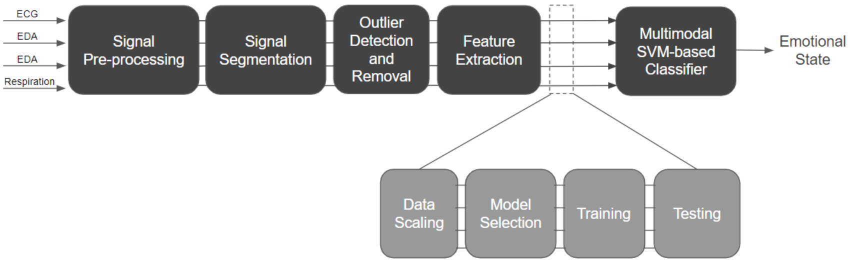

Figure 3: Overall structure of the emotion recognition system proposed.

erage inter-beat intervals (IBI) across the database and excluding the templates with IBI differing in more than 0.1 seconds from that value.

Respiration signals were subjected to an outlier removal step before segmentation, at the peaks level, following the approach proposed in [46]. This method relies on the use of quartiles for both effective outlier detection and segmentation, as those are less sensitive to spikes that may appear in noisy respiration measurements. The upper quartile was used as a lower threshold for a point to be considered a peak, and no more than one peak could be detected in a window of 1.5 seconds [46].

4.4. Classification

In this workflow SVMs were adopted using the LIBSVM[3] implementation, and following the methodology guidelines proposed in [47]:

1. Scaling of the data
2. Kernel selection
3. Cross-validation for model selection
4. Training with the optimal parameters
5. Performance evaluation

In the user-dependent approach, the goal is to classify the emotional state of one individual taking into account only their data as training, whilst in the user-independent scenario the goal is to classify the emotional state of a given individual within the whole population. The self-assessments collected during the experimental protocol were used as ground truth. Although the ratings were assessed through a 9-points scale, due to the reduced size of the database and variability in the reporting, the 9 classes were mapped into a coarser scale, grouping the ratings into negative (classes 1-3), intermediate (classes 4-6) and positive (classes 7-9) in the user-dependent case; and negative (classes 1-5) and positive (classes 6-9) (the median of the initial scale, i.e. 5, was included in the negative class to ensure more balance in the number the samples in each class).

Linear scaling is computed for each feature into the $[-1, +1]$ range. This step is important to avoid attributes in greater numeric ranges that could dominate those in smaller numeric ranges [47].

Nested cross-validation (CV) was used for the tuning of the SVM hyperparameters, aiming a robust model with strong generalization performance [23], comprising the inner (model fitting/training) and outer (model selection) loops. CV was computed using 4-folds, over 30 random trials[4]. By finding the highest score index after the nested CV, the optimal parameters were determined. A range of kernels and respective hyperparameters were determined selecting the optimal model by testing different combinations over the following set:

- *Kernel*: [RBF (radial basis function), Linear]
- C (regularization parameter):[1, 5, 10, 50, 100]
- γ parameter (of RBF): [0.01, 0.001, 0.0001]

The model selection was computed and further implemented for the testing step, by defining the model hyperparameters accordingly.

The multimodal fusion of the independent modalities is performed by weighing each individual decision, with the weights selected based on the accuracy obtained in the nested CV process, leading to a final multimodal classification.

5. Results and Discussion

5.1. Sample Characteristics

A total of 23 participants were enrolled in this study (43.5% female). Due to previous evidence of differentiated emotional processing in old age, the age was limited to 18-40 years old (23±3.7 years old) to minimize the difference in the perceiving of emotion. Only subjects with no history of psychological or neurological conditions were admitted, and none of the participants were reportedly taking any medication that would affect the cardiovascular, respiratory, or central nervous system.

Subjects participated as volunteers in the experiment, and consented to the use of the collected data for the scientific purpose of this work.

Preliminary assessment of the data revealed the presence of artifacts in EDA recordings, caused by physical motion, environmental factors, and electrical noise [12, 41]. Only artifact-free signals, from 18 participants were considered into the analysis.

[3] https://www.csie.ntu.edu.tw/ cjlin/libsvm

[4] https://scikit-learn.org/

5.2. Expected vs. Self-assessed Emotions

The average ratings given by the participants for each video, and respective standard deviation, are plotted in Figure 4. The grey dots represent the expected ratings, based on those determined by [16], except for videos 3 and 5, that are with respect to the conceptual expected values [48].

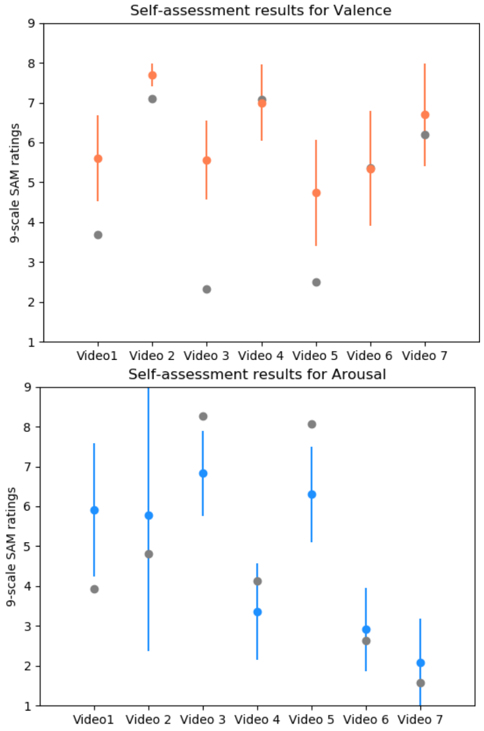

Figure 4: Average of the SAM ratings for valence and arousal per video, along with the respective standard deviation.

Due to the subjective nature of emotions, the reported ratings were compared with those found in previous studies. The target emotions of the videos were assessed with respect to their conceptual labellings as set forth in [48] (100 participants (50% female), with 37±3.14 years-old). Furthermore, we compared the VR video ratings of our population with those of the original study that created the database, by Li et al. [16] (95 participants (56% female), with age within 18-24 years-old). These comparisons are summarized in Table 3.

Average ratings obtained in our work were similar to those found in the VR videos database [16], with less than 1-point (out of 9) difference, except for video 1 that presented both higher valence and arousal than expected. Comparing the ratings with those conceptually expected [48], one can observe larger average differences.

The videos that deviated the most from their conceptual target emotions were 3, 5 and 6, all expected to elicitate lower valence than what was verified. Video 6 elicitated a more melancholic state rather than sadness, while videos 3 and 6, even though with some suspense and violent events, respectively, were not successful in elicitating fear and anger.

5.3. User-Dependent Emotion Classification

The model selection performed in this scenario yielded to several different combinations of the various kernels and model parameters tested for each participant. Considering the performance obtained (cf. Table 4), data fusion was done by weighting each modality accordingly.

Overall performance was quantified as the percentage of correctly classified emotional states per participant, summarized in Figure 5.

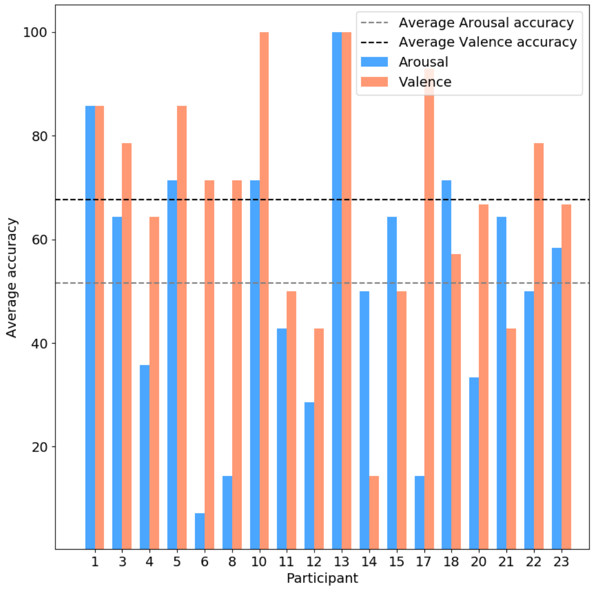

Figure 5: Accuracy obtained in the user-dependent approach, for each participant.

Taking into account the small amount of data available within each participant, it was decided to calculate the recognition performances accepting a penalty for some misclassified emotions. Thus, it was given half of the punctuation if the classes obtained were ”neighbors” of the expected ones (e.g. if the class 1 was expected for a given emotion/video but it would be classified with class 2, it would score half of the punctuation). Under these assumptions, the average recognition performance was 67,68% for valence and 51,07% for arousal.

5.4. User-Independent Emotion Classification

In this approach, the RBF kernel and $\gamma = 0.01$ were optimal for all the cases. The regularization parameter C was either 1 or 100, the former having a softer margin and less error penalty on the training data.

As stated, data fusion was done by weighting each modality according to the performance, in Table 5.

Overall performance was quantified as the percentage of correctly classified emotional states per

Table 3: Summary of the ratings obtained for the stimuli used in our work, and comparison (differences of mean ratings) with the ratings reported in two previous studies [16, 48].

		Results from this Work				**Reported in Previous Studies**				**Differences of the mean Ratings**			
	Target	Valence		Arousal		Valence	Arousal	Valence	Arousal	Study by [16]		Study by [48]	
	Emotion	Mean	SD	Mean	SD	Mean	Mean [16]	Mean	Mean [48]	Valence	Arousal	Valence	Arousal
Video 1	Boredom	5.61	1.08	5.91	1.68	3.69	3.94	4.17	4.04	1.92	1.97	1.44	1.87
Video 2	Joy	7.74	0.24	5.78	3.41	7.10	4.80	8	4.55	0.64	0.98	-0.26	1.23
Video 3	Fear	5.57	0.99	6.83	1.08			2.33	8.26			3.24	-1.43
Video 4	Interest	7.00	0.96	3.35	1.21	7.07	4.13	7.46	3.96	-0.07	-0.78	-0.46	-0.61
Video 5	Anger	4.74	1.33	6.30	1.20			2.5	8.06			2.24	-1.76
Video 6	Sadness	5.35	1.44	2.91	1.04	5.36	2.64	3.4	5.91	-0.01	0.27	1.95	-3.00
Video 7	Relaxation	6.70	1.29	2.09	1.09	6.19	1.57	7.63	1.72	0.51	0.52	-0.93	0.37

Table 4: Average accuracy scores of the models selected in the user-dependent approach, and resulting weighting assigned to each biosignal.

	ECG		**EDA**		**BVP**		**Respiration**	
	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal
Accuracy	0,769	0,735	0,695	0,678	0,726	0,665	0,692	0,626
Weighting	**0,27**	**0,27**	**0,24**	**0,25**	**0,25**	**0,25**	**0,24**	**0,23**

Table 5: Average accuracy scores of the models selected in the user-independent approach, and resulting weighting assigned to each biosignal.

	ECG		**EDA**		**BVP**		**Respiration**	
	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal
Accuracy	0.700	0.656	0.721	0.572	0.695	0.660	0.629	0.585
Weighting	**0.26**	**0.27**	**0.26**	**0.23**	**0.25**	**0.27**	**0.23**	**0.24**
Model Selected	C=100 RBF γ=0.01	C=100 RBF γ=0.01	C=100 RBF γ=0.01	C=100 RBF γ=0.01	C=1 RBF γ=0.01	C=100 RBF γ=0.01	C=1 RBF γ=0.01	C=1 RBF γ=0.01

video, depicted in Figure 6; the performance was slightly better for arousal than valence, in accordance with the reported in related work [7, 14].

When considering all the videos, the system yielded recognition rates of 58.11% for arousal and 57.12% for valence. Results have shown a noteworthy variability, leading us to further analyse the performance when certain videos were excluded, since some yielded to considerably worse accuracy (e.g. videos 2 & 5 for arousal, and 1 & 5 for valence). The complex and subjective nature of self-assessing emotion may explain this variability. Also, the elicitation capabilities of some videos may lead to ambiguity in the emotion felt.For example, as shown in Figure 4, the variability in the arousal ratings of video 2 may be indicative of an ambiguous perception of the content across participants. Therefore, the emotions classified by the system are not necessarily wrong, but instead discrepant from the self-assessed ones. Considering the results obtained for the other three videos mentioned, one can observe that they did not present such large variability, but all presented a considerable difference between the average self-reported and the expected ratings (larger than the standard deviation).

Although not fully conclusive, one might suggest that the performance assessment may be negatively biased by the difficulty of the participants to rate their real emotions, cultural factors, desensitizing or other variability sources that make the content perception differ from what was expected. The best recognition rate was accomplished for the valence classification of video 2, which was the one with the smallest standard deviation from all, somehow corroborating this analysis. As such, we computed the recognition rate excluding videos 2 & 5 in arousal, and 1 & 5 in valence, leading to 69.13% recognition rate for arousal and 67.75% for valence ((*) in Figure 6). It was also assessed excluding all the videos in which the absolute difference between the average and expected ratings was larger than the standard deviation, yielding performances of 61.98% for arousal and 65.74% for valence ((**) in Figure 6).

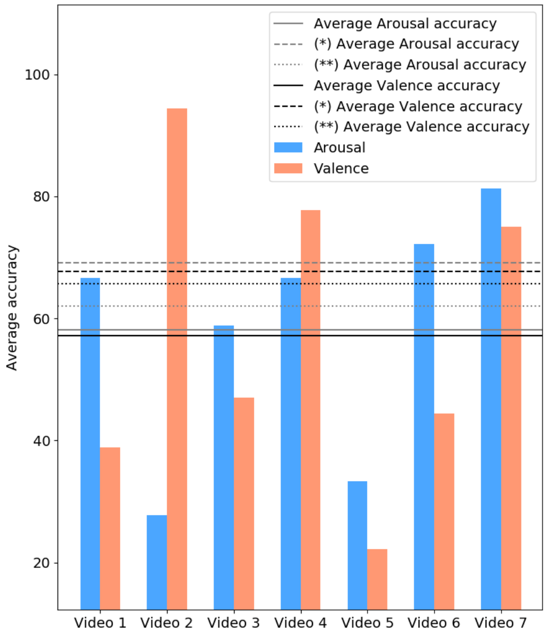

Figure 6: Accuracy obtained for each video. Solid line: considering all videos; (*) excluding videos 2 & 5 in arousal, and 1 & 5 in valence; (**) considering only the videos whose absolute difference between the average and expected ratings was smaller than the standard deviation.

5.4.1 General Considerations

The ambiguous and subjective nature of emotion self-assessment, as well as of the scaling process, are a general concern one must take into account in the scope of emotion recognition and in the analysis of the results of the present work.

Dissociations between subjective and objective measures are referred as very often, being caused by various sources. For instance, context effects have been widely recognized as a common source of bias in subjective judgement [49], e.g. the emotional state of the user in the day of the experiment. In fact, the appropriate scaling structure in psychological attributes is unclear. This way, although emotion can be assigned values on a numerical scale, it lacks an identifiable, completely objective, unit of measurement [50]. Furthermore, Alvarado [50] claims that despite there is evidence that justifies the assumption of an ordinal scale type during data analysis of the emotional response of an individual, there is no evidence that the subjective distances between adjacent numbers on every portion of the scale are equal. Thus comparisons amongst the ratings of individuals are problematic because it is unclear how individual differences in emotional response are related to individual differences in the use of rating scales, and the distances between numbers have not been shown to correspond to the same subjective differences in response for each individual in a study [50].

One should also address the downscaling process that was performed in this work to convert the 1-9 ratings into 1-3 and 1-2 ratings, for the respective scenarios. On one hand, it is legitimate to identify the drawbacks of this approach; considering the above-mentioned problems concerning self-assessment through scales, the computing of this linear scaling might have yielded some conceptual errors with respect to the subjective interpretation of those numbers by the participant. On the other hand, the usage of the 9-points scale was advantageous in terms of conformity and comparability with the ratings in another studies, as it is the most commonly used scale. Future work should perhaps collect, besides the 9-points self-assessments, ratings in scales with the number of points of classes one aims to classify, in order to avoid that subjective scaling and consecutive generalization issues.

Regarding the usage of VR for elicitation stimuli, future studies should assess possible bias induced by this tool, in the sense that its novelty can arguably tend to positively influence the emotions reported.

6. Conclusions

Overall, the user-dependent results have shown an underperformance with respect to related work, which is explained by the fewer training data used in this study, whilst the user-independent approach results are in conformity with the state-of-the-art [13, 14], being considered promising.

As argued in [49, 50], the appropriate scaling structure in psychological and emotional attributes is not fully understood, being unclear how individual differences in emotional response are related to individual differences in the use of rating scales.

Future work will focus on increasing the database size and take into account the subjective factors involved in emotion interpretation.

Acknowledgements

I would like to thank my supervisor, Professor Ana Fred, and Professor Hugo Silva, for all the support and the enthusiastic ideas throughout this work.

References

[1] M. Paleari, R. Benmokhtar, and B. Huet. Evidence theory-based multimodal emotion recognition. In Benoit Huet, Alan Smeaton, Ketan Mayer-Patel, and Yannis Avrithis, editors, *Advances in Multimedia Modeling*, pages 435–446, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[2] S. Thushara and S. Veni. A multimodal emotion recognition system from video. pages 1–5, March 2016.

[3] H. P. Martinez, Y. Bengio, and G. N. Yannakakis. Learning deep physiological models of affect. *IEEE Computational Intelligence Magazine*, 8(2):20–33, May 2013.

[4] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan. Physiological signals based human emotion recognition: a review. In *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*, pages 410–415, March 2011.

[5] Q. Li, Z. Yang, S. Liu, Z. Dai, and Y. Liu. The study of emotion recognition from physiological signals. In *2015 Seventh International Conference on Advanced Computational Intelligence (ICACI)*, pages 378–382, March 2015.

[6] H. Silva, A. Fred, S. Eusebio, M. Torrado, and S. Ouakinin. Feature Extraction for Psychophysiological Load Assessment in Unconstrained Scenarios. IEEE Engineering in Medicine and Biology Society (EMBC) Proceedings, 2012.

[7] A. Haag, S. Goronzy, P. Schaich, and J. Williams. Emotion recognition using bio-sensors: First steps towards an automatic system. In *Affective Dialogue Systems*, pages 36–48, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[8] P. Ekman. Basic emotions. In Tim Dalgleish and M. J. Powers, editors, *Handbook of Cognition and Emotion*, pages 4–5. Wiley, 1999.

[9] P. J. Lang. The emotion probe: Studies of motivation and attention. *The American psychologist.*, 50:372–85, 1995.

[10] H. A. Elfenbein and N. Ambady. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2):203235, March 2002.

[11] A. Alhargan, N. Cooke, and T. Binjammaz. Multimodal affect recognition in an interactive gaming environment using eye tracking and speech signals. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI 2017, pages 479–486, New York, NY, USA, 2017. ACM.

[12] J. Lorenz C. Arnrich B. Trster G. Setz, C. Schumm. Combining worthless sensor data.

[13] S. Bang K. Kim and S. Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, 42:419–427, 2004.

[14] J. Wagner, J. Kim, and E. Andre. From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In *2005 IEEE International Conference on Multimedia and Expo*, pages 940–943, July 2005.

[15] M. Ali, A. H. Mosa, F. A. Machot, and K. Kyamakya. *Emotion Recognition Involving Physiological and Speech Signals: A Comprehensive Review*, pages 287–302. Springer International Publishing, Cham, 2018.

[16] B. J. Li, J. N. Bailenson, A. Pines, W. J. Greenleaf, and L. M. Williams. A public database of immersive vr videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures. *Frontiers in Psychology*, 8:2116, 2017.

[17] E. H. Jang, B. J. Park, S. H. Kim, Y. Eum, and J. H. Sohn. Identification of the optimal emotion recognition algorithm using physiological signals. In *2011 2nd International Conference on Engineering and Industries (ICEI)*, pages 1–6, Nov 2011.

[18] Y. Dai, X. Wang, X. Li, and P. Zhang. Reputation-driven multimodal emotion recognition in wearable biosensor network. In *2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings*, pages 1747–1752, May 2015.

[19] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schller. Enhanced semi-supervised learning for multimodal emotion recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5185–5189, March 2016.

[20] A. Shmilovici. *Support Vector Machines*, pages 257–276. Springer US, Boston, MA, 2005.

[21] V. N. Vapnik. An overview of statistical learning theory. *Trans. Neur. Netw.*, 10(5):988–999, September 1999.

[22] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152. ACM Press, 1992.

[23] C. J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, Jun 1998.

[24] G. Rigas, C. D. Katsis, G. Ganiatsas, and D. I. Fotiadis. A user independent, biosignal based, emotion recognition method. In Cristina Conati, Kathleen McCoy, and Georgios Paliouras, editors, *User Modeling 2007*, pages 314–318, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[25] T. Eerola and J. K. Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49, 2011.

[26] G. Chanel, J. J.M. Kierkels, M. Soleymani, and T. Pun. Short-term emotion assessment in a recall paradigm. *International Journal of Human-Computer Studies*, 67(8):607 – 627, 2009.

[27] G. Riva, F. Mantovani, C. S. Capideville, A. Preziosa, F. Morganti, D. Villani, A. Gaggioli, C. Botella, and M. Alcaiz. Affective interactions using virtual reality: The link between presence and emotions. *CyberPsychology & Behavior*, 10(1):45–56, 2007. PMID: 17305448.

[28] J. Diemer, G. W. Alpers, H. M. Peperkorn, Y. Shiban, and A. Mühlberger. The impact of perception and presence on emotional reactions: a review of research in virtual reality. In *Front. Psychol.*, 2015.

[29] J. Blascovich, J. Loomis, A. C. Beall, K. R. Swinth, C. L. Hoyt, and J. N. Bailenson. Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry*, 13(2):103–124, 2002.

[30] S. A. Hosseini. Classification of brain activity in emotional states using hos analysis. 4, 02 2012.

[31] P. J. Lang. Behavioral treatment and bio-behavioral assessment: Computer applications. In J. B. Sidowski, J. H. Johnson, and T. A. Williams, editors, *Technology in mental health care delivery systems*, pages 119 – 137. Norwood, NJ: Ablex, 1980.

[32] E. Jang, B. Park, S. Kim, and Jin-Hun Sohn. Emotion classification by machine learning algorithm using physiological signals. 2012.

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

[34] E. Jang, B. Park, M. Park, S. Kim, and J. Sohn. Analysis of physiological signals for recognition of boredom, pain, and surprise emotions. *Journal of Physiological Anthropology*, 34(1):25, Jun 2015.

[35] M. Murugappan, S. Murugappan, and B. S. Zheng. Frequency band analysis of electrocardiogram (ecg) signals for human emotional state classification using discrete wavelet transform (dwt). In *Journal of physical therapy science*, 2013.

[36] R. Gupta, M. K. Abadi, J. A. Cárdenes Cabré, F. Morreale, T. H. Falk, and N. Sebe. A quality adaptive multimodal affect recognition system for user-centric multimedia indexing. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ICMR '16, pages 317–320, New York, NY, USA, 2016. ACM.

[37] J. Selvaraj, M. Murugappan, K. Wan, and S. Yaacob. Classification of emotional states from electrocardiogram signals: a non-linear approach based on hurst. *BioMedical Engineering OnLine*, 12(1):44, May 2013.

[38] H. P. da Silva, A. Fred, and R. Martins. Biosignals for everyone. *IEEE Pervasive Computing*, 13(4):64–71, Oct 2014.

[39] A. Lourenço, H. Silva, C. Carreiras, and A. Fred. Outlier detection in non-intrusive ecg biometric system. In Mohamed Kamel and Aurélio Campilho, editors, *Image Analysis and Recognition*, pages 43–52, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[40] A. Greco, G. Valenza, and E. Scilingo. Modeling for the analysis of the eda, 11 2016.

[41] M. Kelsey, R. V. Palumbo, A. Urbaneja, M. Akakaya, J. Huang, I. R. Kleckner, L. F. Barrett, K. S. Quigley, E. Sejdic, and M. S. Goodwin. Artifact detection in electrodermal activity using sparse recovery. 2017.

[42] S. Ouakinin M. Santos, A. Fred. Biometrical and psychophysiological assessment through biosensors, 2012.

[43] Kemalasari and P. S. Wardana. Processing of respiration signals using fir filter for analyze the condition of lung. In *2017 International Electronics Symposium on Engineering Technology and Applications (IES-ETA)*, pages 229–233, Sept 2017.

[44] P. Hamilton. Open source ecg analysis. In *Computers in Cardiology*, pages 101–104, Sep. 2002.

[45] W. Zong, T. Heldt, G. B. Moody, and R. G. Mark. An open-source algorithm to detect onset of arterial blood pressure pulses. In *Computers in Cardiology, 2003*, pages 259–262, Sep. 2003.

[46] Md. M. Rahman, A. A. Ali, K. Plarre, M. al'Absi, E. Ertin, and S. Kumar. mconverse: inferring conversation episodes from respiratory measurements collected in the field. In *Wireless Health*, 2011.

[47] C. Hsu, C. Chang, and C. Lin. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2003.

[48] R. Hepach, D. Kliemann, S. Grneisen, H. Heekeren, and I. Dziobek. Conceptualizing emotions along the dimensions of valence, arousal, and communicative frequency implications for social-cognitive tests and training tools. 2:266, 10 2011.

[49] J. Annett. Subjective rating scales: science or art? *Ergonomics*, 45(14):966–987, 2002. PMID: 12569049.

[50] N. Alvarado. Arousal and valence in the direct scaling of emotional response to film clips. *Motivation and Emotion*, 21(4):323–348, Dec 1997.