A Method for Daily Normalization in Emotion Recognition

Leandro A. Bugnon[1†], Rafael A. Calvo[2], and Diego H. Milone[1]

[1] sinc(i) - Research Center for Signals, Systems and Computational Intelligence, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral, CONICET, Santa Fe, Argentina
[2] Department of Electrical Engineering, University of Sydney, Sydney, NSW, 2006, Australia

† l.bugnon@conicet.gov.ar

Abstract. Affects carry important information in human communication and decision making, and their use in technology have grown in the past years. Particularly, emotions have a strong effect on physiology, which can be assessed by biomedical signals. This signals have the advantage that can be recorded continuously, but also can become intrusive. The present work introduce an emotion recognition scheme based only in photoplethysmography, aimed to lower invasiveness. The feature extraction method was developed for a realistic real-time context. Furthermore, a feature normalization procedure was proposed to reduce the daily variability. For classification, two well-known models were compared. The proposed algorithms were tested on a public database, which consist of 8 emotions expressed continuously by a single subject along different days. Recognition tasks were performed for several number of emotional categories and groupings. Preliminary results shows a promising performance with up to 3 emotion categories. Moreover, the recognition of arousal and emotional events was improved for larger emotion sets.

Keywords: Emotion recognition · Daily variability · Photoplethysmography · Biosignal pattern recognition

1 Introduction

Emotional states are relevant not only in a social context, but also influence directly on cognitive process and take a role in decision making [1]. For this reason, by including affect in human-computer interfaces, the communication performance may be improved. In the early theoretical developments, several discrete categorizations of emotions have been described, for instance the six basic emotions of Ekman [2]: *joy, anger, fear, boredom, sadness, disgust* and *neutral*. Later, particular affective states were proposed for certain research fields, for example *confusion, boredom* and *flow* in educational applications [3]. Furthermore, different continuous scales of emotions were described, such as the core affect theory of Russell [4], which support a model of basic neurophysiological reactions that one can feel like energized/not-energized (*arousal*) and pleasant/unpleasant (*valence*).

The ground in psychophysiology have demonstrated that the affects experienced by a subject have several implications in his physiology (or vice versa), modifying its behaviour at different levels [5]. In the central nervous system, the cortical and sub-cortical activity present electrical variations that can be measured [6]. More frequently, affect has been assessed through the influence of the autonomous nervous system; the sympathetic and parasympathetic branches control different systems, as blood circulation, respiration patterns and skin glands regulation. Additionally, the effects of emotions over the somatic nervous system is present on voluntary muscle activity and involuntary reflex responses [7]. Even though physiological recordings can be intrusive, those signals can be recorded continuously (unlike voice and facial expressions), are more difficult to mask and provides an alternative source in the case of communicational disorders. As more portable and less invasive biosignals acquisition systems are developed, it become more feasible to use them in real world applications.

Affective computing based on physiological variables have advanced over several applications. For example autism-disorders research [8], learning technologies [9], gamer experience [10], multimedia automatic tagging [11] and anti-stress therapy [12] are recent developments. In addition, major efforts have been done to collect representative data, as the experimental design to elicit and measure emotions is complex. Singularly, the *Eight Emotion Sentic Data*[1] was compiled to study the physiological variations on a single subject, acquiring 4 biosignals: facial electromyogram, photoplethysmography (PPG), electrodermal activity and respiration amplitude, over 20 daily sessions. For each one, the subject tried to pass through 8 emotional states: *neutral, anger, hate, grief, platonic love, romantic love, joy* and *reverence*. This dataset make possible to test how biosignals vary in short and long term for different emotions.

The first analysis on the mentioned dataset was performed with an offline classification scheme [13]. In that approach, the signals were previously segmented, obtaining a *recognition rate* (RR) of 46% for the whole emotion set. In their following research, the results were improved taking into account the heart rate and other physiologically relevant features, rising the RR to 81% [14,15]. In addition, a more recent publication report similar results employing different classifiers [16]. Unfortunately, segmented signals are not readily available in real applications. Furthermore, as segments length is about 3 minutes, an instantaneous emotion estimation cannot be obtained. As a first approach towards an online classifier, a feature extraction method with moving average window was employed on this data, achieving a RR of 48.98% for all the emotion set, using all the 4 signals [13]. In another interesting application, it is required to detect an emotional event (this is, a *wake-up call*) from a neutral state, using an online approach. This has been attempted employing an auto-associative neural network [17,18], training it with neutral patterns, and using the difference between new samples and the model estimates for classification.

[1] Public access: http://affect.media.mit.edu/share-data.php

Previous works on affective computing supports the existence of correlation between biosignals and emotions, but the obtained RR still need to be improved significantly. Moreover, it is desirable to minimize the required sensors that allow an acceptable RR. In this direction, the goal of this work is to evaluate the PPG for affect estimation, extracting features related with psychophysiological regulation, and developing a classification scheme oriented to practical online applications.

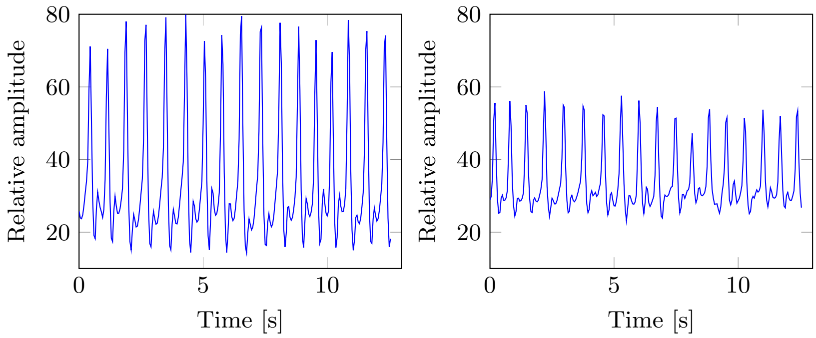
In the next section, the proposed feature extraction and post-processing methods are discussed, along with the employed classifiers. In Section 3, the experiment design to evaluate the models is presented, followed by relevant results and discussion. Ultimately, interesting conclusions and future work are mentioned in Section 4.

2 Feature Extraction and Classification

Cardiovascular measurements are highly available in public databases, presents low invasiveness and seems to be highly related with several emotional dimensions and categories [7]. Specifically, the heart rate reflect arousal by the influence of sympathetic and parasympathetic branches, in response to approach/withdrawal instincts [5]. One of the sources used to capture circulatory activity is the PPG, which measure the blood flow between an infrared led and a sensor, in particular on a finger (Fig. 1). While the heart is pumping, the blood flow depicts peaks in a quasi-periodical signal, and the distance between the peaks is found to be highly correlated with the heart rate [19]. Additionally, the pulse amplitude is related to vasoconstriction; when a subject is under stress, the vessels muscle is activated by sympathetic control [20] and the blood flow is reduced (compare Fig.1.a and Fig.1.b). Taking into account the low invasiveness of finger PPG, it results in a well suited source to estimate emotional information from physiology.

To obtain a heart rate estimate, the signal peaks were detected using a low pass filter and a windowed automatic threshold. The consequent distance between peaks was interpolated to the signal sample rate (20 Hz) using a piecewise cubic Hermite polynomial, which is less oscillating than spline polynomials for this data. Moreover, the PPG amplitude was estimated using the signal envelope, by interpolation of ascending and descending peaks. Over the obtained peak rate and amplitude, a moving window of width W was displaced with a fixed step of 1 s. For each window step, 4 simple features were extracted: the local mean and standard deviation of these two signals. The feature vector is associated with the central point of the window and the emotion label on that point. On the one hand, small values of W allow to detect short time events, but with a cost of more variance. On the other hand, longer lengths tends to provide a better estimate of mean values.

The regulation of the circulatory system is affected by humoral factors, circadian cycle and subject mood during the data registration, resulting in a significant data variance between days. Previous research accounted different nor-



(a) Neutral

(b) Anger

Fig. 1. Segments of PPG signal on a finger. In (a) a neutral state; and (b) an anger episode, recorded in the same day. This difference is not so evident in a wider scope of the signal. Data was extracted from *Eight-Emotion Sentics Dataset* [15].

malization techniques for segmented signals [15], but these approaches cannot be used in online applications. Hereby, a new feature normalization is proposed, which only uses statistical parameters of a neutral segment of length X. It is expected that this neutral sample can approximate the daily baseline, and it can be used to normalize the features in the same day. Furthermore, this approach can be seen as a daily calibration that does not require to acquire many emotion examples everyday. Thus, if $\mathbf{f}(t,d)$ is the feature vector for the time t of day d, and C is the set of the time points used for calibration, the proposed normalization method is applied element by element as:

$$\hat{f}_i(t,d) = \frac{f_i(t,d) - \frac{1}{|C|}\sum\limits_{t \in C} f_i(t,d)}{\max\limits_{t\in C}\{f_i(t,d)\} - \min\limits_{t \in C}\{f_i(t,d)\}} \quad . \tag{1}$$

For classification, two well-known models were tested. The MLP is an artificial neural network which structure is composed by forward full-connected perceptrons. Feature inputs feeds a first layer of perceptrons, and their outputs are connected with subsequent layers (hidden layers) until a final output layer, which have as many neurons as classes. This structure can resolve non-linearly separable problems, and were used with various architectures and training algorithms in categorical emotion problems [21,22,23] and arousal-valence [24] sets. In this work, the MLP was trained with the standard back-propagation algorithm [25].

SVMs are supervised learning methods which minimize the empirical classification error and maximizes the geometrical margin between the classes in the feature space. This margin maximization provides good generalization properties on high-dimension feature sets and few training samples. Even though SVMs are intrinsically binary classifiers, it can be extended to problems of several classes, for example, by *one-vs-all* method. SVMs have been applied to non-linear classification problems transforming the original feature space into a higher dimensional one, where is presumed that the problem becomes linearly separable. Moreover, it is possible to operate in the transformed feature space computing only the inner products between the projected versions of features pairs by using kernel functions. The Radial Basis Function (RBF) kernels are one of the most popular among them [8,23,26] and therefore the Gaussian kernel, a particular RBF, is used in the experiments.

In several datasets for emotion recognition, the classes are imbalanced. This may produce a significant bias in the mentioned classifiers, that advantage the majority class to minimize the overall fitting error. To reduce this effect, the training instances were resampled to equalize the classes occurrences, but leaving the test set as it was. Additionally, it is important to remark that the calibration segments are not used further in the classification task.

3 Results and Discussion

The experiments in this section were performed using signals from the *Eight Emotion Sentic Data*, particularly the *SetB*, which comprises one subject in sessions properly labeled, over 20 days. The 8 emotions elicited in each session are listed in Table 1, with the categorical, arousal and valence labels. Each emotional record have a mean length of 3 minutes, composing a continuous signal of around 25 minutes per session.

Table 1. Dataset labels in categorical, arousal and valence dimensions. The order of appearance is the same as the register.

Emotion	**Arousal**	**Valence**
Neutral	Low	Neutral
Anger	High	Negative
Hate	Low	Negative
Grief	High	Negative
Platonic Love	Low	Positive
Romantic Love	High	Positive
Joy	High	Positive
Veneration	Low	Neutral

For each of the following experiments, the performance of the models was estimated using 5-fold cross-validation on the complete features set. Specifically, the folds were arranged such as each day was contained in a separate fold. In fact, this approach is the most likely in a practical sense, as the model is trained with some days and tested in others. Additionally, in each cross-validation step the parameters of the model were selected according to the best performance using only the 16 training days.

There are several measurements for classifier performance. In this work was used the Cohen's Kappa statistic,

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}, \tag{2}$$

which compare the observed accuracy $\Pr(a)$ in relation with the estimated probability of obtaining the same result by-chance $\Pr(e)$, based on the available data in the confusion matrix [27]. The κ statistic takes the value 0 when the evidence of classification accuracy is the same as the by-chance probability, and 1 for the perfect accuracy. This provide a useful measure to compare sets with different number of classes. The algorithms for signal processing, feature extraction and result analysis were developed in Matlab, using the Weka library [28] to implement the classifiers.

3.1 Categorical Emotion Discrimination

In the first place, the proposed methods were tested in the task of recognizing individual emotions, using sets of 2, 3, 5 and 8 emotions. For each group, the proposed feature normalization method was tested, comparing the performance using MLP and SVM classifiers. The best model parameters were selected between two values of W (30 and 60 s), the length of the calibration segments $X \in \{10, 20, 30\}$ s, and basic classifier parameters: the number of hidden neurons {2,4,8,12} for MLP, and the soft-margin coefficient C and kernel exponential $\gamma \in \{2, 4\}$ for SVM. In the case of W, it was found that smaller values than 30 s worsen the results significantly for the current classification scheme, agreeing with other research [13]. For MLP, the best performance was obtained for 2 to 4 neurons in all groups.

Results for categorical emotion recognition are summarized in Figure 2. In general, all models performed significantly better than the baseline (the by-chance probability). It can be seen that the effect of feature normalization is positive for 2 and 3 emotions. Regarding the classifier model, it was found that MLP and SVM had a similar performance. Comparing with previous research, the present results does not overtake the ones accounted for 8 emotions in an online scheme [13]. However, there is no report of results using only one signal (versus all the four in the dataset), neither on sets smaller than the 8 emotions.

Going further in the evaluation of the normalization method, its effect on several consecutive emotion groupings was tested. It was found that the impact of normalization in RR seems to degrade in relation with the distance between the calibration and the evaluated segments. This explain why normalization does not improve the κ value for bigger set of emotions. Moreover, repeating the experiment with the whole neutral state as calibration segment (and therefore excluding it from the emotion groups) brings a better κ for those groups. This suggest that the size of the calibration segment have a significant effect to estimate a reference value for each day.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. A. Bugnon, R. A. Calvo & D. H. Milone; "A Method for Daily Normalization in Emotion Recognition"
15th Argentine Symposium on Technology, AST 2014, 43º JAIIO, pp. 48-59, 2014.

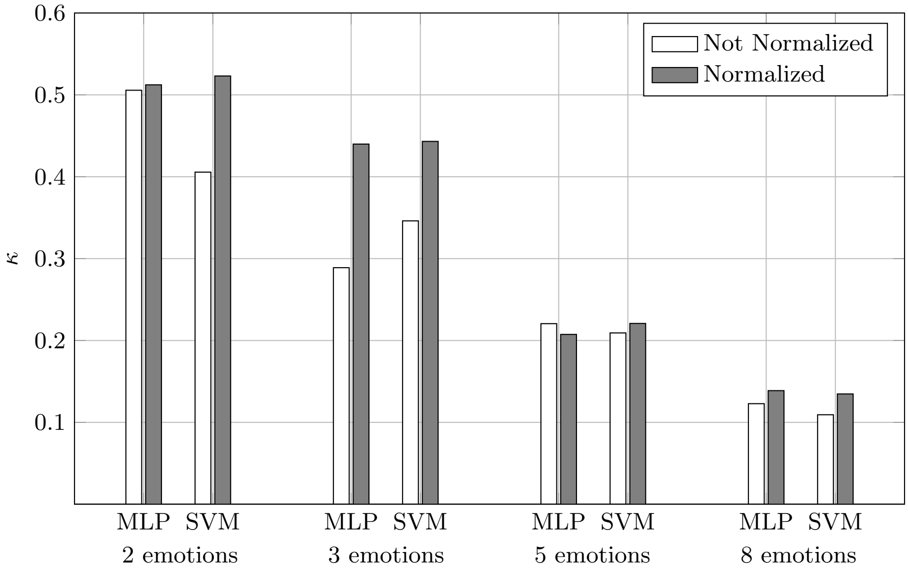


Fig. 2. Mean κ score obtained from cross-validation for individual emotion recognition. Results are shown for emotion sets of various sizes: 2 classes (Neutral-Anger), 3 classes (adding Hate), 5 classes (adding Grief and Platonic Love) and the whole 8 emotions set. The impact of the proposed feature normalization is depicted, along with the comparison of MLP and SVM classifiers.

3.2 Arousal-Valence Discrimination

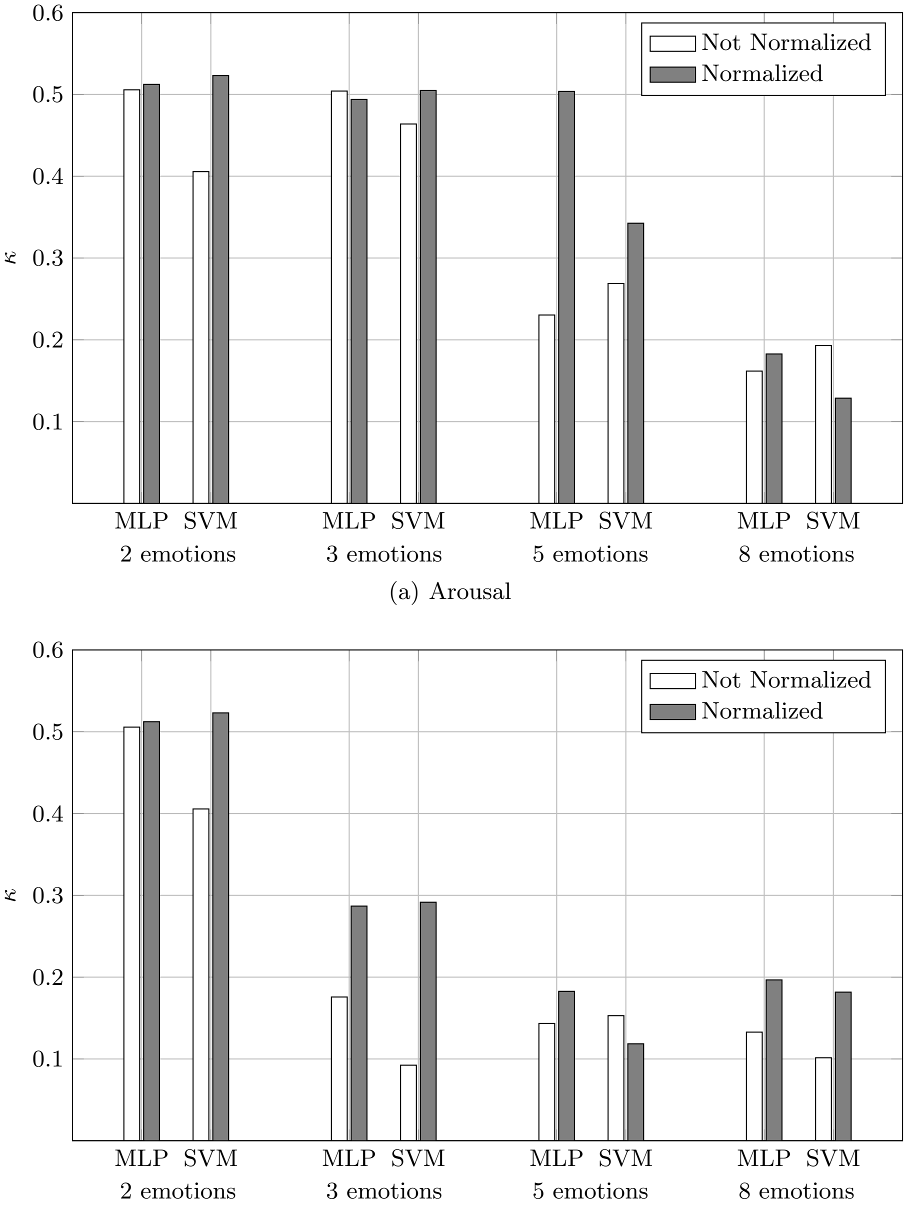
Emotions can be mapped in *arousal* (in this case discretized as low and high) and *valence* (negative, neutral and positive), as shown in Table 1. This mapping allow to group emotions that shares a similar nature, which have a special meaning for different applications. Thereby, it is interesting to evaluate the performance of the PPG and the presented methods over arousal and valence dimensions.

Figure 3 shows the results of testing again the proposed normalization method and the classifiers for the same emotion sets, but grouped into arousal and valence dimensions. In the first case (Fig. 3.a), the normalization rise the score for several number of emotions, significantly better than the baseline in all cases. Moreover, the κ value for 3 and 5 emotions, using MLP and normalization, is close to the 2 emotion group, suggesting that the emotions can be effectively grouped in the arousal sets. In the case of valence (Fig. 3.b), the first 3 emotions have only two classes (neutral and negative), so this should be a similar problem as for arousal. Nevertheless, there is an imbalance towards the negative class in 3 and 5 emotions sets, given by the elicitation ordering. However, the proposed normalization method improves the results significantly for the set of 3 emotions. Finally, the results obtained with the selected methods suggests that arousal discrimination is more feasible than valence with the PPG signal, probably explained by the strong effect of arousal in the circulatory system.

3.3 Emotional Event Detection

The ability to detect an emotional event can be useful, for example, when users have imperative requests. In particular, the dyad neutral-anger is one of the simpler problems, because anger (or rage) has a very active effect in heart rate and vasoconstriction. However, it is not evident if the emotions of different nature, like anger-joy-reverence could be grouped in opposition to the neutral state. To evaluate this, the binary problem of emotion/non-emotion was set by grouping all states but *neutral* as an *emotion*, using a set of the first 5 emotions. Because the dataset is imbalanced for this task (neutral samples are about 1/5 of total, and even lower after removing the calibration segment), two considerations were made. In the first place, the training examples were resampled to equate the occurrences of the classes in training, as used in the valence case. Secondly, as the RR and κ can be misleading in imbalanced data, results were analyzed using the sensitivity (true positive rate) of neutral and emotion classes.

The comparison between the methods shown that, without normalization, MLP achieved a mean sensitivity of 61.7% ($\sigma = 8.2\%$) and SVM scored 58.0% ($\sigma = 7.4\%$). However, the feature normalization rose the mean sensitivity to 73.8% for MLP ($\sigma = 5.8\%$) and 74.0% for SVM ($\sigma = 6.8\%$). Besides the importance of a sufficiently large calibration segment, this factor is trade-off in this task, as higher values overly reduce the neutral instances.



(a) Arousal

(b) Valence

Fig. 3. Mean κ score from cross-validation over different emotion sets. The normalization method and classifiers are compared with the same sets as the categorical emotions experiment, but labeled in: (a) Arousal and (b) Valence dimensions.

4 Conclusions and Future Work

In this work was proposed a low invasive emotion recognition method using only PPG. A simple method for online feature extraction and a classification scheme were tested in different affective computing tasks, involving several categorical and clustered emotion sets. By taking a small segment of the source signal for calibration, a realistic feature normalization procedure was proposed, which reduce the effect of day variance for moderate periods of time.

It was shown that the employed methods performed significantly better than the baseline in all cases. Moreover, the proposed feature normalization improved the results by reducing daily variability. In particular, the categorical emotion recognition have reached acceptable results for small set sizes. Besides the smaller number of classes, the selected methods obtained a good result for arousal recognition, which is interesting for several applications. The current methods also had promising results in the emotional event detection task, despite the drawbacks of the imbalanced dataset. Additionally, it was found that the length of the calibration segment had an important role to reduce the high variability through different day recordings.

In future research, better feature extraction methods will be addressed, aiming to find more discriminant features. On the other hand, classifiers that properly models the time dynamics of biological signals will be developed. Finally, different ways to combine the information of several labels, for example categorical and arousal sets, will be pursued.

References

1. Calvo, R.A., D'Mello, S.: Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. IEEE Transactions on Affective Computing **1**(1) (2010) 18–37
2. Ekman, P., Friesen, W.V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W.A., Pitcairn, T., Ricci-Bitti, P.E.: Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion. Journal of Personality and Social Psychology **53**(4) (1987) 712–7
3. Alzoubi, O.: Automatic Affect Detection From Physiological Signals: Practical Issues. PhD thesis, University of Sydney (2012)
4. Russell, J.: Core Affect and the Psychological Construction of Emotion. Psychological Review **110**(1) (2003) 145–172
5. Cacioppo, J., Tassinary, L.G., Berntson, G.G.: The Handbook of Phychophysiology. 3rd edn. Cambridge University Press (2007)
6. Petrantonakis, P.C., Hadjileontiadis, L.J.: Emotion recognition from brain signals using hybrid adaptive filtering and higher order crossings analysis. IEEE Transactions on Affective Computing **1**(2) (2010) 81–97
7. Kreibig, S.D.: Autonomic Nervous System Activity in Emotion: a Review. Biological Psychology **84**(3) (2010) 394–421
8. Liu, C., Conn, K., Sarkar, N., Stone, W.: Physiology-based Affect Recognition For Computer-Assisted Intervention of Children with Autism Spectrum Disorder. International Journal of Human-Computer Studies **66** (2008) 662–677

9. Hussain, M.S., AlZoubi, O., Calvo, R.A., D'Mello, S.: Affect Detection from Multichannel Physiology during Learning Sessions with AutoTutor. In: International Conference on Artificial Intelligence in Education, Auckland, New Zealand (2011) 131–138
10. Nacke, L.E., Kalyn, M., Lough, C., Mandryk, R.L.: Biofeedback Game Design: Using Direct and Indirect Physiological Control to Enhance Game Interaction. In: ACM Conference on Human Factors in Computing Systems. Volume 11. (2011) 103–112
11. Soleymani, M., Lichtenauer, J.: A Multimodal Database For Affect Recognition and Implicit Tagging. IEEE Transactions on Affective Computing **3**(1) (2012) 42–55
12. van der Zwaag, M.D., Janssen, J.H., Westerink, J.H.: Directing Physiology and Mood through Music: Validation of an Affective Music Player. IEEE Transactions on Affective Computing **4**(1) (2013) 57–68
13. Vyzas, E.: Recognition of Emotional and Cognitive States Using Physiological Data. PhD thesis, Massachusetts Institute of Technology (1999)
14. Healey, J.A.: Weareable and Automotive Systems for Affect Recognition from Physiology. PhD thesis, Massachusetts Institute of Technology (2000)
15. Picard, R., Vyzas, E., Healey, J.: Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(10) (2001) 1175–1191
16. Meftah, I.T., Ben Amar, C.: Multimodal Approach for Emotion Recognition Using an Algebraic Representation of Emotional States. In: International Conference on Signal Image Technology and Internet Based Systems. (2012) 541–546
17. Leon, E., Clarke, G., Callaghan, V., Sepulveda, F.: Real-Time Detection of Emotional Changes For Inhabited Environments. Computers & Graphics **28**(5) (2004) 635–642
18. Leon, E., Montalban, I., Schlatter, S.: Computer-Mediated Emotional Regulation: Detection of Emotional Changes Using Nonparametric Cumulative Sum. In: International Conference of the IEEE Engineering in Medicine and Biology Society. (2010) 1109–1112
19. Giardino, N.D., Lehrer, P.M., Edelberg, R.: Comparison of Finger Plethysmograph to ECG In the Measurement of Heart Rate Variability. Psychophysiology **39**(2) (2002) 246–53
20. Kahneman, D.: Attention and Effort. Prentice-Hall (1973)
21. Nasoz, F., Alvarez, K., Lisetti, C.L., Finkelstein, N.: Emotion Recognition From Physiological Signals Using Wireless Sensors For Presence Technologies. Cognition, Technology & Work **6**(1) (2004) 4–14
22. Wagner, J., Kim, J., André, E.: From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification. In: IEEE International Conference on Multimedia and Expo. (2005) 940–943
23. Yang, S., Yang, G.: Emotion Recognition of EMG Based on Improved L-M BP Neural Network and SVM. Journal of Software **6**(8) (2011) 1529–1536
24. Haag, A., Goronzy, S., Schaich, P., Williams, J.: Emotion Recognition Using Biosensors: First Steps Towards an Automatic System. In: Affective Dialogue Systems. (2004) 36–48
25. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning Representations by Backpropagating Errors. Nature **323**(6088) (1986) 533–536
26. Soleymani, M., Pantic, M., Pun, T.: Multimodal Emotion Recognition in Response to Videos. IEEE Transactions on Affective Computing **3**(2) (2012) 211–223

27. Cohen, J.: A Coefficiente of Agreement for Nominal Scales. Educational and Psychological Measurement (1) (1960) 37–46
28. Hall, M., Frank, E., Holmes, G.: The WEKA Data Mining Software: An Update. SIGKDD Explorations **11**(1) (2009) 10–18