Dimensional Affect Recognition from HRV: an Approach Based on Supervised SOM and ELM

Leandro A. Bugnon, Rafael A. Calvo, *Senior Member, IEEE*, Diego H. Milone, *Member, IEEE*

Abstract—Dimensional affect recognition is a challenging topic and current techniques do not yet provide the accuracy necessary for HCI applications. In this work we propose two new methods. The first is a novel self-organizing model that learns from similarity between features and affects. This method produces a graphical representation of the multidimensional data which may assist the expert analysis. The second method uses extreme learning machines, an emerging artificial neural network model. Aiming for minimum intrusiveness, we use only the heart rate variability, which can be recorded using a small set of sensors. The methods were validated with two datasets. The first is composed of 16 sessions with different participants and was used to evaluate the models in a classification task. The second one was the publicly available Remote Collaborative and Affective Interaction (RECOLA) dataset, which was used for dimensional affect estimation. The performance evaluation used the kappa score, unweighted average recall and the concordance correlation coefficient. The concordance coefficient on the RECOLA test partition was 0.421 in arousal and 0.321 in valence. Results shows that our models outperform state-of-the-art models on the same data and provides new ways to analyze affective states.

Index Terms—Physiological measures, affect sensing and analysis, supervised self-organization, extreme learning machines, dimensional affect estimation.

1 Introduction

Affective states, including emotions, moods, and feelings have a key role in the communication and decision-making process of a person. To improve human-computer interactions (HCI) and human-human computer mediated interactions (as in teleconferences), emotions, engagement and even psychological well-being should be taken into account [1].

The first step to improve interactions is the affect recognition, which can be of two types: categorical or continuous [2]. If the target labels are categories, the recognition task is known as classification. For example, the classes can be the basic emotions summarized by Ekman [3] or those more commonly used in HCI [4]. On the other hand, when labels take continuous values, the task is a regression or estimation of those values. In affective computing, this happens when dimensional models with arousal and valence as continuous variables are used [5]. The dimensional model of affect has also been frequently used in a classification context [6], [7]. In those cases, the labels were the result of a quantization over discrete values. For example, by defining the low, medium and high labels for arousal. These approaches will be referred in this work as classification tasks, leaving the term dimensional affect estimation only when the target affects take the original continuous values.

Several works have shown that physiology is correlated with mental states [8], thus it has been used for affect recognition [2]. One advantage of using physiological signals in real world HCI is that signals can be recorded continuously and may be more unconscious (due to the autonomic response) than traditional sources like voice and facial expressions [9]. Moreover, as one feels less noticed during the sensing of physiology, it may be less invasive in terms of privacy [10]. This is interesting in applications where the user does not want (or does not need) to be recorded by a camera or a mic, such as when playing games [11] or selecting a song playlist [12]. Another relevant case is when people have communicational impairments that makes difficult to analyze other sources [13]. Still, the sensing intrusiveness, the recording noise and the natural variations unrelated to emotions are challenging [7], [14]. The challenges have been addressed in studies using multiple physiological signals: electroencephalography (EEG) [15], [16], [17]; respiration patterns (RP) [18]; skin derived signals such as superficial temperature [19] or electrodermal activity (EDA) [20], [21], [22]; pupillary response [23]; and heart related signals such as electrocardiography (ECG) [24], [25] or photoplethysmography (PPG) [26]. Multimodal combinations of different sources have been addressed to improve recognition rates [2], [27], [28], [29]. Also, several efforts have been made to develop multimodal datasets; for example the DEAP dataset [30] combines EEG, PPG, EDA, RP, facial electromyography and skin temperature, along with audiovisual channels to analyze the impact of multimedia content on users.

The search for a minimally intrusive method is important for real-world applications. In this work, we use Heart Rate Variability (HRV) that has received attention for being related to the autonomic nervous system [14] and basic emotional processes [8]. HRV is the evolution of changes

- *Leandro A. Bugnon and Diego H. Milone are with the Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL/CONICET, Argentina. Ciudad Universitaria, Universidad Nacional del Litoral, 4to piso FICH, (S3000) Santa Fe, Argentina. Tel./Fax: +543424575233 ext 190; http://fich.unl.edu.ar/sinc.*
E-mail: lbugnon@sinc.unl.edu.ar, dmilone@sinc.unl.edu.ar
- *Rafael A. Calvo is with the School of Electrical and Information Engineering, The University of Sydney, Sydney NSW 2006, Australia.*
E-mail: Rafael.Calvo@sydney.edu.au.

in the beat-to-beat interval over time, and can be acquired using only one ECG lead (for example a chest strap) or it can be estimated from PPG using a specialized wristband. The advances of sensor engineering have lowered the costs and improved precision of HRV wearable devices, enabling to obtain this signal in a natural environment [31]. Studies have shown that the mean HR can be estimated from a smartphone accelerometer [32] and remotely from video [33], widening the possibility of a HCI system to include physiological analysis in their framework.

Current affective computing techniques can be improved in a number of areas. For example, most of the techniques require the use of multiple physiological signals, which necessitates more sensors [34] and intrusiveness [7] to the user. There is evidence that affect classification using a source with low intrusiveness like HRV is feasible [35], yet it is not accurate enough for real-world applications. Furthermore, unlike classification approaches, the dimensional estimation of affects has not been widely explored yet. Currently, improving the classifiers performance with novel approaches is an important challenge that should be addressed. Nevertheless, methods usually focus on performance estimation, but omit analyzing the hidden relations in the data. Novel methods for identification and visualization of the subjacent models of affect and its relation with the inputs should be evaluated.

In this work we approach physiological affect recognition with two different methods. For the first method, we propose a novel algorithm based on supervised self-organizing maps (sSOM) to improve recognition rates and also to provide a graphical representation of the underlying model. This representation can relate the features space with the target in a compact way. Opposed to a black-box, this type of models might allow an expert to find, in the trained model, new relations between physiology and affects. For the second method, we propose the use of extreme learning machines (ELM) [36]. ELM are emergent methods for pattern recognition which have shown improved recognition rates with low computational cost in different applications [37]. They have shown to be faster and more accurate than traditional multi-layer perceptrons and support vector machines (SVM) in several benchmarks [38]. ELM have been selected because of their theoretical capacity of dealing with the features non-linearity, the fast training algorithm and a simplistic computational framework.

Models were evaluated in two different datasets: one for classification and the other for dimensional affect estimation. The first dataset was recorded by Monkaresi et al. [39] and consists of multiple-subject recordings of emotions induced with pictures. The labels are binary self-reports in the four quadrants of the arousal-valence (AV) space. The other is the RECOLA dataset [40], composed of multimodal recorded interactions between pairs of subjects during a problem solving task. For each interaction, this dataset contains a dimensional rating in the AV space, which was performed by six external annotators. These datasets have different experimental protocols including: type of interactions, emotional elicitation, spontaneity and labeling methods. In all cases, we use ECG features as input, with special interest in the HRV component. This provides a rich evaluation set for proposed methods. A web-demo [41] interface to rapidly test the methods is available[1]. The source-code of proposed methods is also freely available for academic purposes[2].

In the next section, related works on affect recognition using HRV are reviewed. In Section 3 the datasets used in this work, the feature extraction stage and the experimental setup are presented. In Section 4.1 a sSOM for affect recognition is presented. In Section 4.2 different ELM classifiers are introduced. In Section 5 the most relevant results are presented and discussed. Finally, the conclusions of this work are presented in Section 6.

2 Related works

Several works used the HRV for AV classification. Valenza et al. [42] proposed a nonlinear method for feature extraction from HRV, along with EDA and RP, followed by principal component analysis (PCA) and a quadratic discriminant classifier. Authors obtained promising results using the standard International Affective Picture System (IAPS) as stimuli. Following a similar methodology, relevant improvements have been achieved with sound elicitation for classification in five classes in arousal and valence, using only HRV features [43]. Monkaresi et al. explored the binary classification in the AV space [44] and engagement recognition during a writing-reviewing process [45]. In these works, the authors combined remote HR sensing and facial expressions using a voting classifier composed by SVM, k-nearest neighbor (KNN), decision trees and logistic regression. These works have shown that affect classification using a source with low intrusiveness like HRV is feasible. Currently, improving the classifiers performance is an important challenge that can be addressed by research on novel methods.

Although the categorical approach to affect recognition has been employed successfully in several applications, many human states or traits vary continuously rather than in the rigid classes used in categorical approaches. In such cases the quantization into a few categorical labels might lead to a loss in model representativeness [7]. In comparison with the categorical problem, only a few publications have addressed the dimensional recognition challenges, yet it has become a trend in the affective computing community [7], [40], [46], [47], [48], [49]. Some works approximated dimensional affect indicators with fine-grained quantization scales on segmented data, as in [42]. Haag et al. [50] proposed an assessment of IAPS ratings using a multi-layer perceptron as regressor with multimodal inputs. Later, Bailenson et al. [51] used the same method to estimate levels of sadness and amusement with external raters. However, true dimensional affect estimation with physiological signals is quite recent. Ringeval et al. [52] used HRV, along with other physiological and audio-visual sources, to estimate arousal and valence levels during spontaneous interaction between humans. Several multimodal recognition systems have been tested with the dataset of this work, using for example PCA and linear regression (LR) [53], SVM for regression (SVR) [35], [54], [55], deep neural networks (DNN) [56],

[1] http://fich.unl.edu.ar/sinc/web-demo/dimensional-affect-recognition/

[2] https://sourceforge.net/projects/sourcesinc/files/emoHR

variations of the long-short-term memory recurrent neural network (LSTM) [57], [58], [59], [60], [61], relevance vector machines (RVM) [62], ensembles of random forests (RF) and neural networks [63], and more recently, an end-to-end approach using convolutional and recurrent networks [64]. The accuracy with physiological signals, particularly for HRV, is promising but should be improved to aim for naturalistic interaction applications.

3 Materials and evaluation setup

Datasets, preprocessing and further postprocessing of the output of classifiers are presented in this section. Then, the experimental setup is detailed for both classification and dimensional estimation tasks.

3.1 Classification

The dataset used for classification consisted of 16 laboratory sessions recorded by Monkaresi et al. [39]. Affects were elicited for different subjects using the IAPS and emotions were recorded as self-reports. Each session consisted of approximately 75 images, while one-lead ECG signal was being registered. Images were displayed sequentially in blocks with similar AV score. Each image was shown for 10 s, after which the subjects reported their affective state from 1 to 9 in the Self-Assessment Manikin (SAM) scale. Valence label was binarized in negative and positive classes. For arousal, low and high classes were defined. Our experimental setup with this dataset followed the same procedure of the authors. Sessions were segmented in chunks of one IAPS image. We used the same features provided by the authors: 84 classical features from ECG, including the distances between fiducial points of the PQRS complex, mean heart rate value and first order statistics. The RELIEF-F method was used for feature selection [65]. Each feature was normalized as $\check{x}_{n,j} = (x_{n,j} - \mu_j)/a_j$, where μ_j is the mean feature value and a_j one measure of deviation. The features of the validation partition were normalized using the training partition parameters.

The classifiers were tested for the classes defined above, using one classifier for each subject, one for arousal and another for valence (single-dimension approach) as in [39]. A nested cross-validation was used, including partitions for training, parameter optimization and validation. The hyper-parameters optimization was performed without using the validation partition to get an unbiased performance estimation, including the model optimization. We used two performance measures. The first one is Cohen's Kappa [66],

$$\kappa = \frac{A_0 - A_c}{1 - A_c}, \tag{1}$$

where A_0 is the classification accuracy and A_c the by-chance probability observed in the confusion matrix. It takes into account a baseline reference for classification, being $\kappa = 0$ when there is no evidence that the classifier performs better than a random guess, and $\kappa = 1$ for perfect classification. The second measure is the unweighted average recall (UAR),

$$\mathrm{UAR} = \frac{\sum_i^{n_c} A_i}{n_c}, \tag{2}$$

where A_i is the accuracy for class i data and n_c the number of classes. These coefficients are more robust to class imbalance than the simple accuracy.

3.2 Dimensional estimation

The RECOLA dataset [40] is a multimodal corpus that consists of recordings from dyadic interactions of subjects through an online communication channel (i.e. teleconference). Audio, video, ECG and EDA were registered while the participants were discussing how to solve a survival task. During the interactions affects were expressed spontaneously by the participants. Affects were tagged by six external raters as they perceived them. Their rating was based in a dimensional model of affects, using continuous values in arousal and valence. Also, the rating was annotated frame by frame for the first 5 minutes, thus all the variations in the AV space (according to the raters) are represented. A gold standard target was proposed by the dataset authors to convert the information of the six raters into a unique frame-by-frame rating. To do so, the target was defined as a weighted average of the raters based on their mutual agreement [67]. From the total of 46 subjects, 27 have a complete record of physiological signals. This set was divided by the authors into training, development and test partitions, containing 9 subjects each. In this work we use the 18 subjects that are publicly available (training and development partitions). Proposed methods with optimal hyper-parameters were also evaluated in the test partition.

The HRV signal was estimated from the ECG recording. First the R peaks were identified using the Pan-Tompkin method [68]. Then, the HR was estimated by taking the inverse of R-R distance and interpolating at ECG sampling frequency. Well-known HR features were obtained with a Hamming window of 20 s and a step of 0.5 s. This window length makes possible to have enough data for feature extraction without losing time resolution [49]. In time domain, the HR mean and standard deviations were calculated. From spectral domain, low frequency band (0.04-0.15 Hz), high frequency band (0.15-0.4 Hz) and their ratio were used to estimate autonomous regulations. The spectral decay slope, modeled with a quadratic regression, provided more information of these regulations [69]. Additionally, the total spectral power, 5 fixed bands from 0.04-1 Hz and high order statistics (skewness and kurtosis) were included. The window length permits nearly continuous estimation with sufficient sample length for low frequency features [70]. The first and second derivatives of the features were computed to get information on how they changed in time. Contextual information was also considered by using frame stacking. Given a features vector, $\mathbf{x}_n$, a new set was constructed by adding the m frames before and after each frame, $\check{\mathbf{x}}_n = [\mathbf{x}_{n-m}, \ldots, \mathbf{x}_n, \ldots, \mathbf{x}_{n+m}]$. The features were normalized with the methods detailed above in a session-basis, as described in [67]. Two postprocessing methods were applied to the models outputs. First, a filter was applied to reduce the outputs noise. It was optimized from two common methods in time series processing, the moving average and the exponential smoothing. Then, an output correction factor was tested to adjust the output amplitude. This factor was defined as the mean ratio between filtered outputs and target amplitudes in the training set.

The set of sessions was divided in a 3-folds nested cross-validation scheme, so each session was used either for training or testing at a time. To compare the estimations with the targets, we used the Lin's concordance correlation coefficient ρ_c [71], which is the scoring metric used by other works on the RECOLA dataset and it is the official metric of the Audio/Visual Emotion Challenge and Workshop (AVEC) since 2015. This metric is defined as

$$\rho_c(\mathbf{y}, \hat{\mathbf{y}}) = \frac{2\rho\sigma_{\mathbf{y}}\sigma_{\hat{\mathbf{y}}}}{\sigma_{\mathbf{y}}^2 + \sigma_{\hat{\mathbf{y}}}^2 + (\mu_{\mathbf{y}} - \mu_{\hat{\mathbf{y}}})^2}, \tag{3}$$

where ρ is the Pearson's correlation coefficient, μ_y is the mean and σ_y is the standard deviation. The range of ρ_c is [-1,1], taking values around 0 if there is no concordance evidence, and 1 for a perfect concordance. This coefficient is an improvement of ρ, as it considers the correlation of the signals in time along with the mean square error.

The proposed methods were tested in different scenarios. In all cases, the gold standard rating was used as the estimation target. In first place, the methods were faced to the single-dimension estimation, training models for arousal and valence independently. However, it has already been shown that arousal and valence are dependent during emotional elicitation [72]. Thus, the two-dimensions AV estimation was conducted for comparison. In this experiment, the two outputs were estimated simultaneously by one model, whose parameters were optimized to maximize the mean ρ_c of both targets. As a baseline, two standard classifiers were evaluated. One is a SVR with Gaussian kernel[3], the regularization factor (C) optimized in the range $[2^{-5}, 2^{25}]$ and the Gaussian exponential (γ) in the range $[2^{-30}, 2^{-5}]$. The other classifier is a RF[4], in which the number of trees was optimized in the range $[5, 150]$ and the size of the features subsets from 5 to the whole feature set for each tree. To compare the results with previous works, additional experiments were conducted. The best models were trained with the training partition, and validated on the development partition. In this case, the best hyperparameters were taken from cross-validation experiments as in [63], thus minimizing model overfitting. A final evaluation was made using the test partition. The optimal models from cross-validation experiments were trained with the whole public set (training and development partition) and labels were estimated on the test set of features. These estimations were made only once and sent to the authors of the RECOLA dataset for evaluation.

4 Methods

4.1 Supervised self-organizing maps

A self-organizing map is a neural network generally composed by one bi-dimensional layer of units. This model has been proposed for dimensionality reduction, clustering and classification [73]. In this section, we propose a novel method to train a sSOM for dimensional affect estimation. To this end, the inputs in the training stage will be the features extended with the target affects. Rather than minimizing an error function between the model output and the expected targets, an unsupervised method creates a map based on the similarity between the extended input vectors. Different regions are conformed on this map, associating the values of features and targets. When new unlabeled data is presented, the features are compared with the learned weights of all units in the map and the closest unit is chosen as output unit. Then, the affect learned by this unit is the estimated target, which was chosen based on the spatial structure of the map previously defined by the training data. An important advantage of this method is that the high-dimensional input space is mapped into a 2D representation. Therefore, new relations between the features and the affective space can be discovered by simple inspection.

Formally, given a set of N samples with F-dimensional features and P-dimensional continuous targets, lets define the input matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^T$, with $\mathbf{x}_n \in \mathbb{R}^F, n = 1, \dots, N$, and the target matrix $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]^T$, with $\mathbf{y}_n \in \mathbb{R}^P$. The features and targets in the training set are concatenated as a single input matrix. The new input matrix is given by $\check{\mathbf{X}} = [\check{\mathbf{x}}_1, \dots, \check{\mathbf{x}}_N]$, with $\check{\mathbf{x}}_n = [\mathbf{x}_n, \lambda\mathbf{y}_n] \in \mathbb{R}^{F+P}$, where the scaling factor λ must be set to balance the influence of the targets in the map topology.

The sSOM have a rectangular array of units s_j, with $j = 1, \dots, J$. For a given input $\check{\mathbf{x}}_n$, the output of s_j is given by

$$h_j = \varphi(\check{\mathbf{x}}_n, \check{\mathbf{w}}_j), \tag{4}$$

where $\check{\mathbf{w}}_j = [\mathbf{w}_j^x, \mathbf{w}_j^y] \in \mathbb{R}^{F+P}$ is the synaptic weight vector, composed by the feature weights $\mathbf{w}_j^x$ and the target weights $\mathbf{w}_j^y$. The operator φ is a similarity function, usually based in the euclidean distance. Weights are traditionally instanced at random [73]. However, the data distribution can be used to avoid local minima and speed up the training. Thus, an alternative to random initialization is to use PCA in the input space. First, the method finds the two greater eigenvalues and eigenvectors from the training set. Then, the weights of the map are generated by linear spanning in the two dimensions. In this way, the main data variability is initially arranged along the main axes of the map.

The sSOM training is an iterative procedure. At each time $t = 1, \dots, T$, a sample $\check{\mathbf{x}}(t)$ is presented to the map. The best matching unit is the one with higher similarity with the input pattern. It is found by solving

$$s^*(t) = \arg\min_j ||\check{\mathbf{x}}(t) - \check{\mathbf{w}}_j||_2. \tag{5}$$

The method rewards the neuron s^* by adjusting $\check{\mathbf{w}}_{s^*}$ for a better matching with the sample. To induce a topological ordering in the map, the rewarding effect is scattered through the neighbouring units. The neighbours are defined in hexagonal shape, thus a 1-unit neighbourhood is a set of 6 units plus the central unit. Then, the weights are updated using the steepest-descent gradient optimization

$$\check{\mathbf{w}}_j(t+1) = \begin{cases} \check{\mathbf{w}}_j(t) + \alpha[\check{\mathbf{x}}(t) - \check{\mathbf{w}}_j(t)], & \text{if } s_j \in \mathcal{N}_{s^*} \\ \check{\mathbf{w}}_j(t), & \text{if } s_j \notin \mathcal{N}_{s^*} \end{cases}, \tag{6}$$

where $0 < \alpha < 1$ is the learning factor, and $\mathcal{N}_{s^*}$ is a neighbourhood function around s^*. For the early iterations, the $\mathcal{N}_{s^*}$ radius and α take large values. This configuration leads to a rough ordering of the map, defining the main topographic zones. In the later iterations, $\mathcal{N}_{s^*}$ and α are

[3] Implemented with the quadratic programming functions of Matlab. It is included in the provided source code.

[4] Standard Matlab implementation: TreeBagger class.

reduced until only s^* is affected by the optimization algorithm. This results in a fine-tuning of the map while the main topological structure is conserved. In addition to the traditional planar sSOM, a toroidal form can be defined by linking opposed border units of the plane in a same neighborhood, thus every unit will have the same amount of neighbours. Preliminary evaluations of toroidal model have not provided better estimations than the planar one, and graphical analysis of a toroid is more complicated. Thus, only the planar map will be used.

Once the sSOM training is complete, similar inputs will lead to closer winner units. As training inputs contains the features and targets, each region of the map will model the spatial relations of both spaces. Then, in the recognition stage only the feature weights $\mathbf{w}_j^x$ are used. Thus, the best matching unit is obtained with

$$s^* = \arg\min_j ||\mathbf{x} - \mathbf{w}_j^x||_2, \tag{7}$$

and the output estimation is given by

$$\tilde{\mathbf{Y}}_{s^*} = \frac{\mathbf{w}_{s^*}^y}{\lambda}. \tag{8}$$

The smoothness of the outputs may depend on both the input data and the size of the map. Therefore, a spatial interpolation method is incorporated as last step. Given the input $\mathbf{x}$, the K closest units are determined, $[s_1^*, \dots, s_K^*]$, with s_1^* the best matching unit as in (7). Then, the smoothed output is obtained with the weighted average

$$\bar{\mathbf{Y}} = \sum_{k=1}^{K} \gamma_k \tilde{\mathbf{Y}}_{s_k^*}, \tag{9}$$

where

$$\gamma_k = \frac{||\mathbf{x} - \mathbf{w}_{s_k^*}^x||_2^{-1}}{\sum_{j=1}^{K} ||\mathbf{x} - \mathbf{w}_{s_j^*}^x||_2^{-1}} \tag{10}$$

is the normalized inverse distance for each s_k^* in the feature space. With this expression, the closest s_k^* in the feature space receives the higher weight.

4.2 Extreme learning machines

The theoretical context of ELM includes several related methods [74]. In this section, two different ELM approaches are introduced: the original model as a neural network, and a later derivation based in kernels.

In the first conception of ELM, the classifier can be seen as a neural network with one hidden layer (nELM). The central paradigm is that the hidden units are randomly generated, thus the tuning of their parameters is avoided. As a direct consequence, the training time is dramatically reduced compared with other training methods. For a formal derivation, consider J hidden units with F inputs and P output units. The output of the hidden layer is given by

$$h_j = \Phi(\mathbf{v}_j^T \mathbf{x} + b_j), \tag{11}$$

where Φ is the activation function, $\mathbf{v}_j$ the input weights and b_j the bias for the j-th hidden unit. This can be expressed in a matrix form by defining the hidden-layer output matrix $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_N]^T$, also called the feature projection matrix, and $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_P]$ as the output layer weights, with $\mathbf{w}_p \in \mathbb{R}^J$ and $p = 1, \dots, P$. Then, the nELM output can be written as

$$\tilde{\mathbf{Y}} = \mathbf{HW}. \tag{12}$$

If Φ is an infinitely differentiable function, and $(\mathbf{v}_j, b_j)$ are randomly selected, it can be demonstrated that for any pair $(\mathbf{X}, \mathbf{Y})$ there exist a number $J < N$ such $||\tilde{\mathbf{Y}} - \mathbf{Y}|| < \epsilon$ for any small ϵ [74]. This means that the ELM can approximate the target $\mathbf{Y}$ of a given input $\mathbf{X}$ by adjusting only the number of hidden units and the output weights. To find $\mathbf{W}$, the problem can be stated as

$$\underset{\mathbf{W}}{\text{minimize}} \quad ||\mathbf{HW} - \mathbf{Y}||_2, \tag{13}$$

which is a least square optimization problem. The smallest norm solution is given by

$$\hat{\mathbf{W}} = \mathbf{H}^\dagger \mathbf{Y}, \tag{14}$$

where $\mathbf{H}^\dagger$ is the Moore-Penrose pseudo-inverse [75].

A generalized ELM method based on kernels (kELM) can be derived from this theory [38]. To improve the solution stability and generalization, a regularized optimization problem was proposed as

$$\begin{aligned} \underset{\mathbf{W}}{\text{minimize}} \quad & \frac{1}{2}||\mathbf{W}||_2 + C\frac{1}{2}\sum_{n=1}^{N} ||\epsilon_n||_2 \\ \text{subject to} \quad & \mathbf{h}_n \mathbf{W} = \mathbf{y}_n^T - \epsilon_n^T, \end{aligned} \tag{15}$$

where ϵ_n is the training error vector for the sample $\mathbf{x}_n$ and C a regularization factor. The solution is given by

$$\hat{\mathbf{W}} = \mathbf{H}^T \left(\frac{1}{C}\mathbf{I} + \mathbf{HH}^T \right)^{-1} \mathbf{Y}. \tag{16}$$

Let be $\breve{\mathbf{H}}$ the feature projection of the training set $(\breve{\mathbf{X}}, \breve{\mathbf{Y}})$, and $\mathbf{H}$ the projection of any other set $(\mathbf{X}, \mathbf{Y})$. The estimation of $\mathbf{Y}$ is given by (12) and (16)

$$\tilde{\mathbf{Y}} = \mathbf{H}\hat{\mathbf{W}} = \mathbf{H}\breve{\mathbf{H}}^T \left(\frac{1}{C}\mathbf{I} + \breve{\mathbf{H}}\breve{\mathbf{H}}^T \right)^{-1} \breve{\mathbf{Y}}. \tag{17}$$

With the selection of a kernel function $\mathcal{K} : (\mathbb{R}^F, \mathbb{R}^F) \to \mathbb{R}$, the kernel matrix for the inputs $(\mathbf{X}, \mathbf{X}')$ is defined as

$$\mathbf{\Omega}(\mathbf{X}, \mathbf{X}') = \mathbf{HH}'^T : \Omega_{i,j} = \mathbf{h}_i \cdot \mathbf{h}_j = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j'). \tag{18}$$

Thus, (17) becomes

$$\tilde{\mathbf{Y}} = \mathbf{\Omega}(\mathbf{X}, \breve{\mathbf{X}}) \left(\frac{1}{C}\mathbf{I} + \breve{\mathbf{\Omega}} \right)^{-1} \breve{\mathbf{Y}}, \tag{19}$$

where $\breve{\mathbf{\Omega}}$ is the training kernel matrix. From (19), it can be seen that the kernel function replaces the projection matrices.

Several hyper-parameters detailed in Section 3 and 4 were optimized with a grid search. The order of feature derivatives, frame stacking size and post processing parameters were optimized for each case. The feature normalization factor a was the standard deviation for sSOM and the maximum amplitude of the data for ELM. For sSOM, we explored different map architectures (size and shape), the scaling factor λ, the spatial interpolation method and training length. For nELM, a hidden layer of variable size and standard activation functions were used, including sigmoid, hard-limit, and sinusoidal functions. The kELM was implemented with a radial basis function, with the exponential coefficient γ and associated regularization factor C being tuned in a grid with logarithmic scale.

5 Results and discussion

In the first part of this section we show the classification performance of the proposed methods in comparison with baseline classifiers and previous works. In the second part we show experimental results for dimensional affect estimation on the RECOLA database. We show the distinctive use of sSOM as a qualitative model to explore affects and their relations with the physiological features. Then we compare the proposed models and the baselines in a quantitative way with cross-validation results. In the last part, we make a comparison with state-of-the-art works by using the same dataset partitions.

5.1 Categorical affect classification

The results shown in Table 1 are the average of 10 randomized cross-validation repetitions on the categorical dataset. The columns are the single-dimension classification tasks, while in the rows are listed our models, baseline classifiers (SVM with radial basis function kernel and RF) and the reference for comparison. In [39], a vote classifier was used to combine the decision of four standard classifiers: SVM, KNN, decision trees and logistic regression. From Table 1 it can be seen the most effective classifiers are kELM for arousal and SVM for valence.

Our methods outperformed previous results on the classification task. Although that classes were balanced and selected to be contrasting (high versus low, and positive versus negative), the score was considerably higher for valence. This may suggest that the selected features from ECG had a better discriminability or the elicitation process was more effective in valence. This was consistent with the results reported by the authors of the database [39]. When comparing arousal results with the reference, our methods show a higher difference in the kappa score. By using a single classification model as proposed here, instead of several classifiers, the number of tuning parameters has been reduced, making a simpler model for the problem. As it was shown, by using the same features and experimental setup, the classifiers proposed here can improve the results for binary categorization of arousal and valence.

Comparing now the models with higher scores, SVM and kELM show similar results. Both methods share the theoretical objective of projecting data to a higher dimension where data may be easier to separate, in this case using the same kernel function. However, kELM is faster and uses less memory during the optimization. Differences between kELM, nELM and RF are significant ($p < 0.01$, one-way ANOVA) for arousal and valence in both performance measures. However, kELM required more resources as the algorithm uses the training data to provide estimations. The trained sSOM is represented in a small set of parameters, thus the memory usage for training is considerably low. The sSOM seems to be effective as well, which may be explained by the unsupervised association of features and affects, providing robustness to outliers. This model also shows a significant difference with nELM and RF for valence and nELM for arousal ($p < 0.01$). The better scores provided by kELM can probably be explained by its capacity for non-linear modeling of the feature space.

TABLE 1
Mean κ and UAR for binary affect classification on Monkaresi et al. dataset.

	Arousal		Valence	
Classifier	κ	UAR	κ	UAR
Vote classifier [39]	.071	-	.191	-
SVM	.143	.570	**.213**	**.607**
RF	.109	.558	.163	.582
sSOM	.137	.566	.202	.597
nELM	.068	.541	.119	.559
kELM	**.148**	**.576**	.208	.603

Performance for the classification task is lower than the dimensional estimation (as will be detailed in the next section). The general differences between our framework in classification and dimensional estimation tasks could be explained by the effect of several factors. In the first place, emotion expression is different in the datasets. Emotions in both datasets are naturally expressed, this is not acted. However, the dataset used for classification involves induced emotions (using IAPS) and the dimensional estimation dataset involves spontaneous emotions. The report is also different; the use of affect reports of several raters in RECOLA may involve a better estimation. Last but not least, categories in the classification case are binarized from a dimensional model. This may restrain emotion expression, as extreme emotions are in the same category than near-neutral values. In this way, continuous labels may be more difficult to register but they have a richer expression that the classifiers can use.

5.2 Dimensional affect estimation

We analyze first the interesting visual information obtained from sSOM trained on the RECOLA dataset. The high-dimensional space of the features and targets can be reduced to intuitive bi-dimensional representations. Some of them are shown in Fig. 1. Each one represents the distribution of a coefficient of the synaptic weights $\check{w}_j$ in the map. Let us take for instance the *Arousal* plane. As explained in Section 4.1, arousal is part of an extended input of the model during training. The hexagonal cells represent the sSOM units, placed with their neighbours as they are in the model. The value of the input, in this case the arousal level, is modeled with the $\check{w}_j$ of each unit. This value is indicated in the image with a color scale that ranges from blue at the minimum to red at the maximum value. Upon visual inspection, one consequence of the training algorithm is that similar values are arranged in neighbouring units. In the *Arousal* plane it can be seen that the low-arousal zone was ordered in the upper-left of the image, increasing approximately along with the vertical axis to the bottom, which is the high-arousal zone. In the same way, looking now at the *Valence* plane, we can see a smooth value progression in an almost perpendicular direction to the *Arousal* plane.

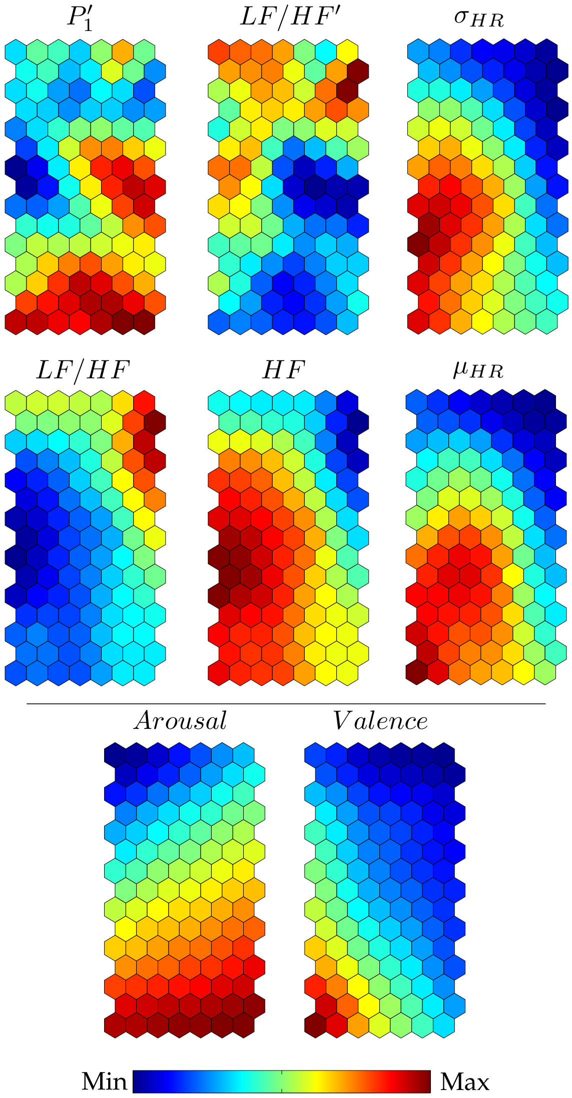

Fig. 1. Graphical representation of the inputs and targets in a trained sSOM. On the top, six features layers are shown: the first derivative of P_1 frequency band (P_1'), the low and high frequency bands ratio (LF/HF) and its first derivative (LF/HF'), the high frequency band (HF), the amplitude of HR (σ_{HR}) and its mean (μ_{HR}). On the bottom, the target layers for arousal and valence are displayed, along with the color-bar for reference.

Comparing now the targets (arousal and valence) with the other input planes (the actual features), relations between them can be assessed in a qualitative form. Although all features are considered in a multidimensional way to obtain an accurate estimation, this analysis can contribute to finding relevant features. From Fig. 1, it seems that P_1' and LF/HF' are strongly related with arousal. This can be seen by observing that the high P_1' zone and the low LF/HF' zone are overlaid with the high-arousal zone. In a similar way, LF/HF, HF and the HR statistics (σ_{HR} and μ_{HR}) can be associated with the valence distribution. It has been argued that the HRV is related with valence and well-being, specifically HF being directly correlated with valence [76]. It can be seen in Fig. 1 that low valence has a coincident area with low and medium HF, as well as with σ_{HR} and μ_{HR}, thus adding empirical support to the argument. This type of analysis provides a tool to visualize the relationship between affects and important features from the data.

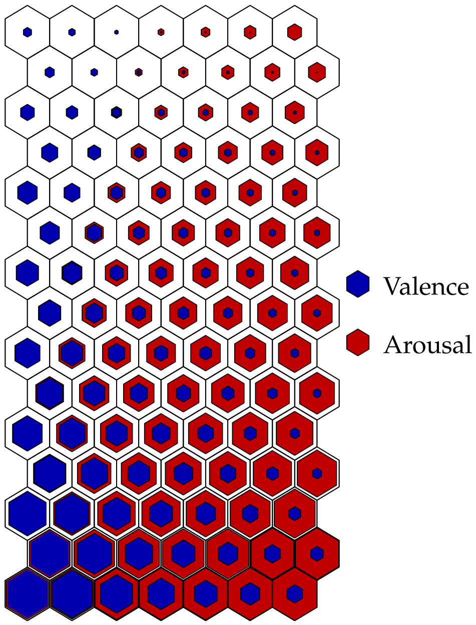

Fig. 2. Graphical representation of the AV space using sSOM. Output components of the trained map are superposed in the same plane. Arousal is represented in red and valence in blue. The level of these variables in each sSOM unit is coded by the size of each hexagon. Notice that this is the same information displayed by the trained model, now summarized in one single image.

For a practical and compact representation of the emotion model learnt by sSOM, we can merge the arousal and valence maps from Fig. 1 in a unique map as in Fig. 2. The targets are shown in colors, red for arousal and blue for valence. Now their values are coded in the size of the hexagons instead of a color scale. This map provides an idea at a glance of the structure and relation of arousal and valence in the model. The sSOM units are represented in the same positions as in Fig. 1, so the topological relations between features can be related with this new target map. Even more, if it is used to estimate dimensional affects in real-time, the AV map could serve as a display to show graphical interactions between the variables, highlighting the winner unit at each moment.

The relationship between the traditional AV plane and the new data-driven representation is schematized in Fig. 3. The gray dots in the AV plane are theoretical affects that may be reported by a subject during an affect elicitation experiment. Some idealized cases (stressed, excited, relaxed and sad) are displayed in both representations. Thus, Fig. 3 illustrates how affects can be mapped from the theoretical AV plane to the sSOM by looking to the arousal and valence levels. Using this relationship, the sSOM graphical representation can be discussed using the following example case. It has been reported that affective stimuli (like audiovisual resources) are not equally effective to induce affects all over the AV plane [30], [72], [77]. In fact, in those works it was found that with low arousal levels there is very little possible variation in valence, which stays near the neutral point. However, for high arousal it can be reached the full spectrum of valence expression. That is, if the affects are

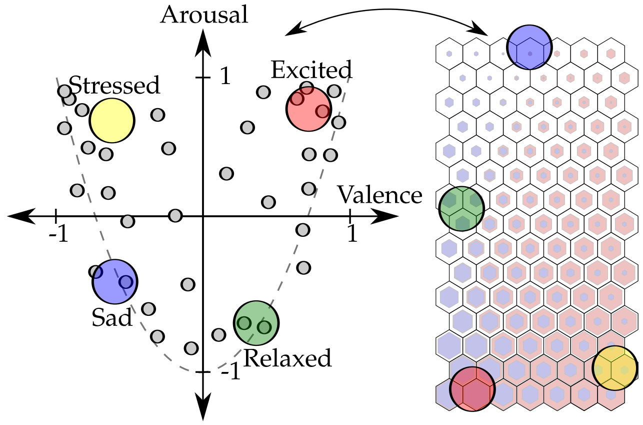

Fig. 3. Representation of the mapping between the theoretical AV space and the sSOM graphical model. The pictured affects (stressed, excited, sad and relaxed) are illustrated to exemplify the data-driven mapping.

TABLE 2
Mean ρ_c and standard deviation from 3-fold cross-validation on the RECOLA dataset. Each session features were normalized independently. Proposed models were trained using a single-dimension (one model for arousal and another for valence) and using the two dimensions togheter.

	Classifier	Arousal	Valence
	SVR	.378 (.030)	.243 (.064)
	RF	.369 (.018)	.282 (.028)
Single-dimension output	sSOM	.362 (.032)	.313 (.059)
	nELM	.366 (.039)	**.322** (.054)
	kELM	**.388** (.009)	.320 (.049)
Two-dimensions output	sSOM	.364 (.033)	.312 (.059)
	nELM	.379 (.039)	.313 (.060)
	kELM	.388 (.010)	.321 (.044)

plotted in a Cartesian AV space, the dots are inscribed in a parabolic shape, as seen in Fig. 3. If we now compare this results and the map of Fig. 2, we can see a similar behaviour in the affect representation of our experiment. The complete range of valence is only observable in high arousal zone (the bottom part of the map), while in low arousal zones the valence is between neutral and negative. This way, the theoretical relations between arousal and valence have been warped to the sSOM, using only the training data. This mapping provides an alternative representation of the AV space given the mutual closeness between the feature and target samples, as a direct property of the sSOM training algorithm. Moreover, this example case of data-driven representation provides empirical evidence to support theoretical models described previously. This is indeed an advantage of the sSOM over black-box models, in that it provides graphical representations of data isles and may provide support for theoretical assumptions, based only on the training data.

The first cross-validation results on the RECOLA dataset are shown in Table 2. In the columns are the estimated targets, and in the rows are the proposed methods along with standard models as SVR and RF for comparison. Methods are detailed as single-dimension when trained with either arousal or valence, and two-dimensions when trained with both outputs simultaneously. It can be seen that ELM achieve the higher concordance rates. On the one hand, nELM has a fast and low memory implementation. On the other, kELM provided the lowest variances, denoting a more stable model across the tested sessions. However, sSOM follows these results closely with the advantage of providing an explicit model for visual analysis, as mentioned above. In agreement with [67], arousal estimation was shown to be more accurate, with higher ρ_c and lesser variance. It is also interesting to note that nELM works very well with the RECOLA dataset but not so in the classification dataset (Section 5.1). A possible explanation is that the generalization capacity of nELM improve with the amount of data available.

Estimating both targets with the same model seems to slightly improve sSOM and nELM performance in arousal. However, the general performance is similar compared to the single-dimension approach and it did not yield significant differences. The results show that the combination of arousal and valence with the current models does not seem to provide better recognition capabilities than individual target models. Measuring the concordance between arousal and valence targets for all sessions, it yields a mean of 0.29. This suggests concordance between the targets and may explain the lack of improvement, as the additional information provided when estimating one target along the other is not leveraging the results. However, the bi-dimensional outputs makes possible the analysis presented with sSOM about Fig. 2.

Although the proposed models are able to perform the recognition in real-time, an additional factor that should be considered with the recognition performance is the computational cost of training the models. Both sSOM and nELM models are compressed in low quantity of parameters and low training time. On the contrary, the training data is needed to compute every estimation with kELM, as seen in Section 4.2. The training time for kELM was significantly higher than nELM and sSOM models, but these are much lower than SVR. This difference may be important for future applications in limited hardware, as wearable devices, or when using bigger datasets.

As described in [67], the gold-standard rating is composed by six human raters. Let us consider a rater as either one of these humans or one of the proposed models. It is interesting to evaluate the behaviour of the models in comparison to the humans. The agreement between a pair of raters can be measured with the mean ρ_c across the sessions. The agreement of each rater with the other human raters is shown in Table 3. In the rows are listed the six human raters, the mean inter-rater agreement and the proposed models. The columns are arousal and valence as independent targets. These results show that the proposed models have a mean agreement superior to some raters (the rater 6 in arousal and rater 5 in valence). In the case of arousal, the ELM models even approximate the second least agreeing rater (the number 5). Although valence estimation is more challenging, consistent with the results discussed above, humans have a higher inter-rater agreement for valence. This may be explained by the natural human ability to identify valence states from face expressions [78]. Comparing the mean inter-rater concordance from Table 3 and results from

TABLE 3
Mean ρ_c between a rating (either from a human rater using audio-visual cues or a model using physiology) and all the human raters in the database.

Raters	Arousal	Valence
Rater 1	.305	.386
Rater 2	.296	.361
Rater 3	.349	.327
Rater 4	.231	.259
Rater 5	.223	.181
Rater 6	.113	.298
Mean inter-rater	.253	.302
sSOM	.203	.192
nELM	.216	.214
kELM	.215	.208

Table 2, it can be seen that the models have competitive performance. Therefore, it can be said that the proposed models could be playing the role of an external rater with a consistent agreement with other raters and good rating towards the gold standard label.

An example of the estimated ratings for a validation session is shown in Fig. 4. The arousal and valence targets correspond to the session tagged as *dev-3* in RECOLA. These are compared with the sSOM, nELM and kELM outputs. The shaded area is the standard deviation of ratings given by the human raters. It can be seen that models yield close estimations and follows the main events in the target signal, like the peaks around 45 s and 150 s in arousal. Also, the estimations mainly remain in the shaded zone. For this session, the percentage of the arousal estimations inside the human rater deviation is 83%, 80% and 81% for nELM, kELM and sSOM respectively, while for valence is 71%, 67% and 78%. These relations are similar for the whole database, with 73%, 75% and 76% for arousal, 67%, 68% and 71% for valence. It can be seen that the sSOM outputs follow the reference more closely in general. However, Table 2 shows that ELM models achieve higher ρ_c, which may be explained by its higher sensibility to small variations. This can be seen, for example, around the 260 s point in Fig. 4. Another aspect that have been discussed in previous works is the asynchrony between emotional expression and the emotional labels provided by the external raters [52]. It was reported [35] that when using the HRV signal, a delay on the training labels does not improve the performance of the classifier. As HRV responses are slower than audio-visual cues, it is possible that the rater delay may have been partially compensated with the physiological delay.

As detailed in the previous section, the experiments described here involve features that were normalized for each session independently, as in [67]. However, a more challenging case is the task of estimating dimensional affects on a totally new subject in real-time, without any prior sensing on this new subject. This case can be evaluated with a small change in the feature normalization stage. Instead of normalizing the features in a session basis, the normalization parameters (features mean and deviation) are obtained from the training sessions only. Results for single and two-dimensions models are shown in Table 4. It can be observed that ELM models can provide a better estimation and also seems more robust than sSOM to the feature normalization challenge. Differences between ELM methods and baseline classifiers are significant in all experiments ($p < 0.05$). Moreover, kELM achieved significant improvement against the other methods as well ($p < 0.05$). This suggest that feature complexity of multiple subject experiments is better handled with proposed methods.

TABLE 4
Mean ρ_c and standard deviation from 3-fold cross-validation on the RECOLA dataset. The normalization parameters (features mean and deviation) are obtained from the average of training sessions.

	Classifier	Arousal	Valence
	SVR	.155 (.019)	.104 (.046)
	RF	.119 (.048)	.126 (.018)
Single-dimension output	sSOM	.165 (.051)	.141 (.060)
	nELM	.217 (.025)	.230 (.034)
	kELM	.260 (.002)	.223 (.047)
Two-dimensions output	sSOM	.181 (.038)	.148 (.056)
	nELM	.262 (.008)	**.253** (.029)
	kELM	**.263** (.007)	.227 (.046)

TABLE 5
Mean ρ_c of the proposed models and other works on the RECOLA *development* partition.

Classifier	Reference	Arousal	Valence
Using ECG features			
Vote classifier	Ringeval et al. 2015 [67]	.275	.183
LSTM	Chao et al. 2015 [59]	.222	.182
LSTM	Chen et al. 2015 [58]	.333	.314
DNN-LSTM	He et al. 2015 [57]	.297	.293
DNN	Cardinal et al. 2015 [56]	.262	.124
NN ensemble	Kachele et al. 2015 [63]	.344	.256
oaRVM	Manandhar et al. 2016 [62]	.293	.274
S-fusion SVR	Weber et al. 2016 [55]	**.468**	.221
Using HRV features			
PCA + LR	Povolny et al. 2016 [53]	.391	.388
SVR	Valstar et al. 2016 [35]	.379	.293
SVR	Sun et al. 2016 [54]	.392	.264
S-fusion SVR	Weber et al. 2016 [55]	.424	.413
LSTM	Brady et al. 2016 [61]	.357	.364
End-to-end	Keren et al. 2017 [64]	.426	**.419**
sSOM		**.402**	.354
nELM		.399	**.375**
kELM		.388	.338

5.3 Comparisons with state-of-the-art methods

Previous works on the RECOLA dataset reported their results on the partition called *development* (detailed in [67]). State-of-the-art models that used only the ECG recordings for the affect recognition are listed in Table 5. We show these results together with our results in the same partition for comparison. Our results were achieved with the proposed methods optimized by cross-validation, thus minimizing the overfitting on development partition. As expected, our results are near the cross-validation results from Table 2, with the sSOM performing better for arousal and nELM for valence.

The state-of-the-art methods were reported in two groups. The first group of publications used general features

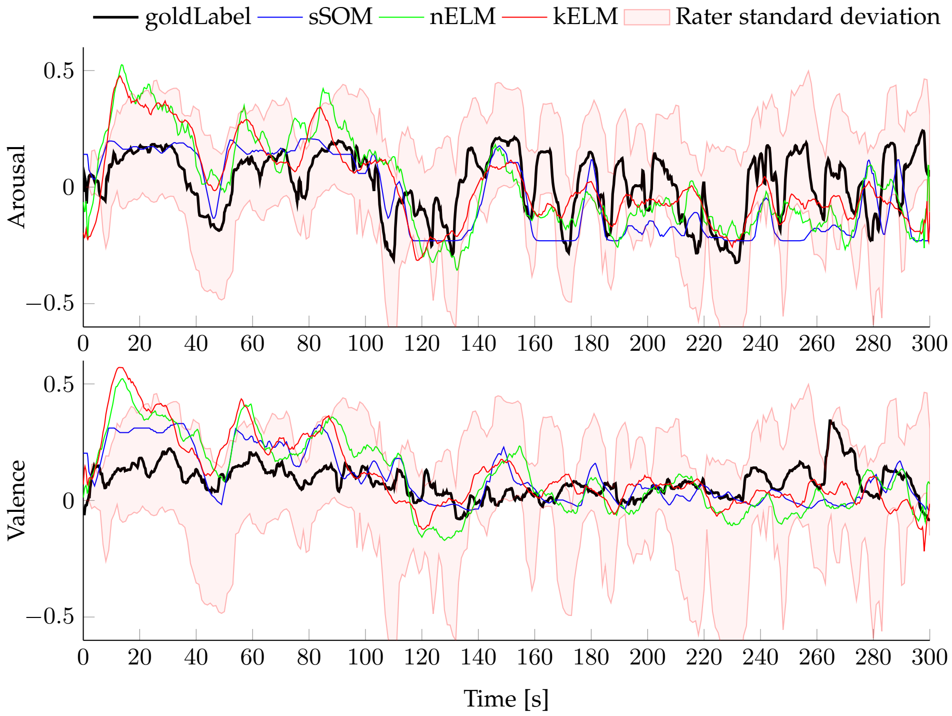

Fig. 4. Comparing the outputs of the models with their targets. Gold-standard for arousal and valence is in bold black line. Model estimations are represented with color lines. The shaded area correspond to the standard deviation of human raters.

TABLE 6
Mean ρ_c of the proposed models and other works on the RECOLA *test* partition.

Classifier	Reference	Arousal	Valence
Vote classifier	Ringeval et al. 2015 [67]	.192	.139
DNN	Cardinal et al. 2015 [56]	.161	.121
Linear SVR	Valstar et al. 2016 [35]	.334	.198
End-to-end	Keren et al. 2017 [64]	.360	.225
sSOM		.404	.273
nELM		**.421**	**.321**
kELM		.367	.293

of the ECG signal. The second group uses only HR and HRV derived features, with an overall better performance, except for [55], which provides a better score for arousal using ECG features. Many of the revised works ([57], [58], [59], [61]) were based on the LSTM model, which was introduced for this database in [52] and it is considered a state-of-the-art model for dimensional affect estimation [7], [57]. In these models, interesting variations have been proposed by defining different loss functions, like the ϵ-insensitive loss in [59] and the concordance correlation in [58], instead of using square-error based functions. These networks use memory inputs to learn from the evolution of the features at different time scales, thus are suitable for estimating time series. In our models, time context information was introduced using the feature derivatives and frame stacking. This short time context provided enough information to achieve competitive results. Moreover, our methods may be more robust when long time recordings are not available. Other works use DNN models [56] and combinations of DNN and LSTM [57]. The SVR with linear kernels is also popular for this task, using L2 [35] and L1-regularized [54] loss functions. One of the outstanding results for arousal is a subject-level fusion (S-fusion) strategy using SVR, achieving a wide difference with other works [55]. However, authors state that the model lack of generalization capabilities, not being able to reach the baseline results of the test partition. This can be explained by the final stage of the model training, which was adjusted using the same developing partition (which was used to measure the performance), thus overfitting the model. On the contrary, works like [63] uses a cross-validation stage to perform hyper-parameter optimization and thus recognition rates for unseen test data are more predictable. The most recent work, which is also better than the others in valence prediction, is an end-to-end approach [64]. In that work, convolutional and recurrent networks are used to learn features and time dynamics directly from HRV signal, thus avoiding hand-made features.

Related to the discussion on two-dimensions models (Section 5.2), an output-associative RVM (oaRVM) was proposed in [62]. This model is trained using a feed-back of both arousal and valence estimations, as proposed in [79] for audiovisual features. They show that their two-dimensions approach achieved an improvement compared to other single-dimension models. However, authors from [58] could not find important differences in recognition performance between single and two-dimensions approaches for the reported modalities. In our work, the two-dimensions approach is of importance in two cases. First, using both target variables in sSOM makes it possible to visualize relationships with the multidimensional feature space. Secondly, we show that there is a small improvement in the case of a new session to be estimated in real-time, without normalization in a session basis. Moreover, there is practically no increase in the computational cost of training the proposed models for simultaneous arousal and valence estimation in contrast with the heavier computational cost of oaRVM.

The available results on the test partition are shown in Table 6. It can be seen that our methods improved the state-of-the-art on this partition. Among them, nELM approach achieved the best estimations. This could be explained by their random hidden layer generation, providing optimal solutions with good generalization. The sSOM method also achieved competitive results. The patterns between features and targets were effectively modeled with the sSOM structure and the unsupervised training. Summarizing the results of Table 6, estimation of both targets using only HRV was effectively improved with the proposed methods.

6 Conclusions and future work

In this work two new methods for affect recognition have been presented, using features extracted only from the HRV. A novel supervised self-organization model (sSOM) was proposed to improve the recognition accuracy but also to provide a graphical representation of relations between features and targets. Contrary to a black-box model, the sSOM represents a graphical superposition between sensed data and affects. Given the sSOM properties, numerical and categorical variables can be represented, making it a very versatile model for HCI applications. Two novel methods based on extreme learning machines (nELM and kELM), were also applied to these tasks. These models were evaluated in classification and dimensional affect estimation, providing competitive performance compared with state-of-the-art works. In classification, the best results were achieved by kELM in both arousal and valence. In the dimensional estimation task, proposed models outperformed state-of-the-art results in the RECOLA test partition. sSOM obtained very good performance according to the quantitative measures and also provided an alternative way to represent multidimensional data. nELM achieved the best performance with a very low computational cost.

We already shown general properties of the methods and results for classification and regression tasks. Moreover, proposed methods can be used for several applications. sSOM can directly combine features and labels of different nature (categorical or numerical) making it a very versatile model. It could be trained for example to model target affects as engagement or boringness in conjunction with personal traits categories. The graphical representation provided by this model could be exploited in real-time for the communication of affects and personal states, where one can manage to see a relation between the input space and the labels. In addition, ELM proved to be as versatile as SVM with a simplistic framework, as the ELM algorithm for both regression and classification is very similar. It can also manage additional dimensions in the output just by adding an output unit. Moreover, these models have shown that it is possible to face affect recognition using only HRV. The possibility of using a physiological signal like this is promising for out-of-the-lab applications. With better performance on HRV signals and the advances of wearable technology, real-world HCI applications could be seen with such a simple equipment as a wrist-band or a distant web-cam.

In future works we will investigate ways to combine multi-rater information. The point-to-point agreement between raters may be an important clue to determine automatically the confidence around an affect estimation and improve training. Another topic for further research is to improve the dimensional estimation of valence, for which new methods for capturing the temporal dynamics should be explored.

Acknowledgments

This work was supported by the Agencia Nacional de Promoción Científica y Tecnológica (PICT 2015-0977, PICT 2014-1442), and Universidad Nacional del Litoral (PECAP 2014, CAID 2011-519/525), Argentina. RAC is supported by ARC Future Fellowship FT140100824.

References

[1] R. A. Calvo and D. Peters, *Positive computing: Technology for Wellbeing and Human Potential*. The MIT Press, 2014.

[2] R. A. Calvo and S. D'Mello, "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications," *Affective Computing, IEEE Transactions on*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[3] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3, pp. 169–200, 1992.

[4] S. D'Mello and R. A. Calvo, "Beyond the basic emotions: what should affective computing compute?" in *CHI 2013 – Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Paris, France, 2013, pp. 2287–2294.

[5] J. Russell, "Core Affect and the Psychological Construction of Emotion," *Psychological Review*, vol. 110, no. 1, pp. 145–172, 2003.

[6] R. A. Calvo and S. Mac Kim, "Emotions in text: Dimensional and categorical models," *Computational Intelligence*, vol. 29, no. 3, pp. 527–543, 2013.

[7] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *IMAVIS*, vol. 31, no. 2, pp. 120–136, 2013.

[8] S. D. Kreibig, "Autonomic Nervous System Activity in Emotion: A Review," *Biological Psychology*, vol. 84, no. 3, pp. 394–421, 2010.

[9] S. H. Fairclough, "Fundamentals of physiological computing," *Interacting with Computers*, vol. 21, no. 1-2, pp. 133–145, 2009.

[10] R. Picard, E. Vyzas, and J. Healey, "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[11] R. L. Mandryk, M. S. Atkins, and A. I. N. Press, "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.

[12] M. D. van der Zwaag, J. H. Janssen, and J. H. Westerink, "Directing Physiology and Mood through Music: Validation of an Affective Music Player," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 57–68, 2013.

[13] C. Liu, K. Conn, N. Sarkar, and W. Stone, "Physiology-based Affect Recognition For Computer-Assisted Intervention of Children with Autism Spectrum Disorder," *International Journal of Human-Computer Studies*, vol. 66, pp. 662–677, 2008.

[14] G. Valenza, L. Citi, A. Lanatá, E. P. Scilingo, R. Barbieri, G. Valenza, L. Citi, and A. Lanata, "Revealing Real-Time Emotional Responses: a Personalized Assessment based on Heartbeat Dynamics," *Scientific reports*, vol. 4, no. 4998, p. 4998, 2014.

[15] O. AlZoubi, R. A. Calvo, and R. H. Stevens, "Classification of EEG for affect recognition: an adaptive approach," in *Advances in Artificial Intelligence*, 2009, pp. 52–61.

[16] R. Khosrowabadi, H. C. Quek, A. Wahab, and K. K. Ang, "EEG-based Emotion Recognition Using Self-Organizing Map for Boundary Detection," *International Conference on Pattern Recognition*, pp. 4242–4245, Aug. 2010.

[17] R. Khosrowabadi, C. Quek, K. K. Ang, and A. Wahab, “ERNN: A Biologically Inspired Feedforward Neural Network to Discriminate Emotion From EEG Signal,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 3, pp. 609–620, 2014.

[18] C.-K. Wu, P.-C. Chung, and C.-J. Wang, “Representative Segment-Based Emotion Analysis and Classification with Automatic Respiration Signal Segmentation,” *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 482–495, 2012.

[19] S. Ioannou, V. Gallese, and A. Merla, “Thermal infrared imaging in psychophysiology: Potentialities and Limits,” *Psychophysiology*, Jun. 2014.

[20] M.-Z. Poh, N. C. Swenson, and R. W. Picard, “A wearable sensor for unobtrusive, long-term assessment of electrodermal activity.” *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 5, pp. 1243–52, May 2010.

[21] R. W. Picard, S. Fedor, and Y. Ayzenberg, “Multiple Arousal Theory and Daily-Life Electrodermal Activity Asymmetry,” *Emotion Review*, no. 1974, 2014.

[22] J. Hernandez, I. Riobo, A. Rozga, G. Adowd, and R. Picard, “Using electrodermal activity to recognize ease of engagement in children during social interactions,” in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2014.

[23] M. Soleymani, M. Pantic, and T. Pun, “Multimodal Emotion Recognition in Response to Videos,” *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211–223, 2012.

[24] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, “ECG Pattern Analysis for Emotion Detection,” *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 102–115, 2012.

[25] S. Jerritta, M. Murugappan, K. Wan, and S. Yaacob, “Classification of emotional States from electrocardiogram signals: a non-linear approach based on Hurst,” *Biomedical Engineering Online*, vol. 12, no. 1, p. 44, May 2013.

[26] W. Handouzi, C. Maaoui, A. Pruski, and A. Moussaoui, “Objective model assessment for short-term anxiety recognition from blood volume pulse signal,” *Biomedical Signal Processing and Control*, vol. 14, pp. 217–227, Nov. 2014.

[27] C. D. Katsis, N. S. Katertsidis, and D. I. Fotiadis, “An integrated system based on physiological signals for the assessment of affective states in patients with anxiety disorders,” *Biomedical Signal Processing and Control*, vol. 6, no. 3, pp. 261–268, Jul. 2011.

[28] D. Novak, M. Mihelj, and M. Munih, “A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing,” *Interacting with Computers*, vol. 24, no. 3, pp. 154–172, May 2012.

[29] W. Wen, G. Liu, N. Cheng, J. Wei, P. Shangguan, and W. Huang, “Emotion recognition based on multi-variant correlation of physiological signals,” *Affective Computing, IEEE Transactions on*, vol. 5, no. 2, pp. 126–140, 2014.

[30] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, “DEAP: A database for emotion analysis using physiological signals,” *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[31] Z. Zhang, Z. Pi, and B. Liu, “TROIKA: A General Framework for Heart Rate Monitoring Using Wrist-Type Photoplethysmographic Signals During Intensive Physical Exercise,” *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 522–531, 2015.

[32] J. Hernandez, D. J. McDuff, and R. W. Picard, “Biophone: Physiology monitoring from peripheral smartphone motions,” in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. IEEE, 2015, pp. 7180–7183.

[33] M.-Z. Poh, D. J. McDuff, and R. W. Picard, “Non-contact, automated cardiac pulse measurements using video imaging and blind source separation,” *Optics express*, vol. 18, no. 10, pp. 10 762–74, May 2010.

[34] M. A. Quiros-Ramirez, S. Polikovsky, Y. Kameda, and T. Onisawa, “Towards developing robust multimodal databases for emotion analysis,” 2012, pp. 589–594.

[35] M. Valstar, J. Gratch, F. Ringeval, M. T. Torres, S. Scherer, and R. Cowie, “AVEC 2016 – Depression, Mood, and Emotion Recognition Workshop and Challenge,” 2016.

[36] G. B. Huang, “An Insight into Extreme Learning Machines: Random Neurons, Random Features and Kernels,” *Cognitive Computation*, vol. 6, no. 3, pp. 376–390, 2014.

[37] G. Huang, G. B. Huang, S. Song, and K. You, “Trends in extreme learning machines: A review,” *Neural Networks*, vol. 61, pp. 32–48, 2015.

[38] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, “Extreme learning machine for regression and multiclass classification.” *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, vol. 42, no. 2, pp. 513–29, 2012.

[39] H. Monkaresi and R. A. Calvo, “Feasibility of a low-cost platform for physiological recording in affective computing applications,” in *The 10th International Conference on Body Area Networks (BodyNets 2015).*, vol. 1, no. 1, 2015, p. 2012.

[40] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, “Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions,” *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pp. 1–8, Apr. 2013.

[41] G. Stegmayer, M. Pividori, and D. H. Milone, “A very simple and fast way to access and validate algorithms in reproducible research,” *Briefings in Bioinformatics*, vol. 17, no. 1, pp. 180–183, 2016.

[42] G. Valenza, A. Lanatà, and E. P. Scilingo, “The Role of Nonlinear Dynamics in Affective Valence and Arousal Recognition,” *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 237–249, 2012.

[43] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. Scilingo, “Recognizing Emotions Induced by Affective Sounds through Heart Rate Variability,” *IEEE Transactions on Affective Computing*, vol. 3045, no. c, pp. 1–9, 2015.

[44] H. Monkaresi, R. A. Calvo, and H. Yan, “A Machine Learning Approach to Improve Contactless Heart Rate Monitoring Using a Webcam,” *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1153–60, 2014.

[45] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D. Mello, “Automated Detection of Engagement using Video-Based Estimation of Facial Expressions and Heart Rate,” *IEEE Transactions on Affective Computing*, vol. 3045, no. c, pp. 1–14, 2016.

[46] M. Nicolaou, H. Gunes, and M. Pantic, “A multi-layer hybrid framework for dimensional emotion classification,” *Proceedings of the 19th ACM international conference on Multimedia - MM '11*, p. 933, 2011.

[47] B. Schuller and F. Weninger, “Ten recent trends in computational paralinguistics,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7403 LNCS, pp. 35–49, 2012.

[48] A. Mencattini, E. Martinelli, F. Ringeval, B. Schuller, and C. D. Natale, “Continuous Estimation of Emotions in Speech by Dynamic Cooperative Speaker Models,” *IEEE Transactions on Affective Computing*, vol. 3045, no. c, pp. 1–14, 2016.

[49] W. Schuller, “Acquisition of Affect,” in *Emotions and Personality in Personalized Services*. Springer International Publishing, 2016, ch. II, pp. 57–80.

[50] A. Haag, S. Goronzy, P. Schaich, and J. Williams, “Emotion Recognition Using Bio-sensors: First Steps Towards an Automatic System,” in *Affective Dialogue Systems*, 2004, pp. 36–48.

[51] J. N. Bailenson, E. D. Pontikakis, I. B. Mauss, J. J. Gross, M. E. Jabon, C. a.C. Hutcherson, C. Nass, and O. John, “Real-time classification of evoked emotions using facial feature tracking and physiological responses,” *International Journal of Human-Computer Studies*, vol. 66, no. 5, pp. 303–317, 2008.

[52] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-p. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, “Prediction of Asynchronous Dimensional Emotion Ratings from Audiovisual and Physiological Data,” *Pattern Recognition Letters*, 2014.

[53] F. Povolný, P. Matêjka, M. Hradîš, A. Popková, L. Otrusina, and P. Smrž, “Multimodal Emotion Recognition for Avec 2016 Challenge,” in *AVEC '16 Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 75–81.

[54] B. Sun, S. Cao, L. Li, J. He, and L. Yu, “Exploring Multimodal Visual Features for Continuous Affect Recognition,” in *AVEC '16 Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 83–88.

[55] R. Weber, V. Barrielle, C. Soladié, and R. Séguier, “High-Level Geometry-based Features of Video Modality for Emotion Prediction,” in *AVEC '16: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 51–58.

[56] P. Cardinal, N. Dehak, A. Lameiras, J. Alam, and P. Boucher, “ETS System for AV+EC 2015 Challenge,” in *AVEC '15 Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. New York, New York, USA: ACM, 2015, pp. 17–23.

[57] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, “Multimodal Affective Dimension Prediction Using Deep Bidirectional Long

Short-Term Memory Recurrent Neural Networks," in *AVEC '15 Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 73–80.

[58] S. Chen and Q. Jin, "Multi-modal Dimensional Emotion Recognition using Recurrent Neural Networks," in *AVEC '15 Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. New York, New York, USA: ACM, 2015, pp. 49–56.

[59] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long Short Term Memory Recurrent Neural Network based Multimodal Dimensional Emotion Recognition," in *AVEC '15 Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. New York, New York, USA: ACM, 2015, pp. 65–72.

[60] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller, "Facing Realism in Spontaneous Emotion Recognition from Speech: Feature Enhancement by Autoencoder with LSTM Neural Networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH), to be Publ.*, San Francsico, CA, 2016.

[61] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang, "Avec '16: Proceedings of the 6th international workshop on audio/visual emotion challenge," in *AVEC '16 Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. New York, NY, USA: ACM, 2016, pp. 97–104.

[62] A. Manandhar, K. D. Morton, P. A. Torrione, and L. M. Collins, "Multivariate Output-Associative RVM for Multi-Dimensional Affect Predictions," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 10, no. 3, pp. 365–372, 2016.

[63] M. Kachele, P. Thiam, F. Schwenker, and M. Schels, "Ensemble Methods for Continuous Affect Recognition: Multi-modality, Temporality, and Challenges," in *AVEC '15 Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. New York, New York, USA: ACM, 2015, pp. 9–16.

[64] G. Keren, T. Kirschstein, E. Marchi, F. Ringeval, and B. Schuller, "End-to-end learning for dimensional emotion recognition from physiological signals," in *Proceedings 18th IEEE International Conference on Multimedia and Expo (ICME)*. New York, New York, USA: ACM, 2017, pp. 985–990.

[65] M. Robnik-Sikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Machine Learning Journal*, vol. 53, pp. 23–69, 2003.

[66] K. L. Gwet, *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters.*, 4th ed. Advanced Analytics, LLC, 2014.

[67] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "The AV + EC 2015 Multimodal Affect Recognition Challenge: Bridging Across Audio, Video, and Physiological Data," in *AVEC '15 Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. New York, New York, USA: ACM, 2015, pp. 3–8.

[68] J. Pan and W. J. Tompkins, "A Real-Time QRS Detection Algorithm," pp. 230–236, 1985.

[69] A. Accardo, G. Addio, D. Maestri, D. Vitale, G. Furgi, and F. Rengo, "Fractal dimension and power-law behavior reproducibility and correlation in chronic heart failure patients," in *Signal Processing Conference*, 2002.

[70] U. Rajendra Acharya, K. Paul Joseph, N. Kannathal, C. M. Lim, and J. S. Suri, "Heart rate variability: a review," *Medical & biological engineering & computing*, vol. 44, no. 12, pp. 1031–51, Dec. 2006.

[71] L. I.-k. Lin, "A Concordance Correlation Coefficient to Evaluate Reproducibility Author(s): Lawrence I-Kuei Lin Source:," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.

[72] M. M. Bradley and P. J. Lang, "The international Affective Picture System (IAPS) in the Study of Emotions and Attention," in *Handbook of emotion elicitation and assessment*, 2007, ch. 2, pp. 29–46.

[73] T. Kohonen, "Essentials of the self-organizing map," *Neural Networks*, vol. 37, pp. 52–65, 2013.

[74] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006.

[75] A. Ben-Israel and T. N. E. Greville, *Generalized inverses: theory and applications*, 2nd ed. Springer, 2001.

[76] J. Gruber, D. S. Mennin, A. Fields, A. Purcell, and G. Murray, "Heart rate variability as a potential indicator of positive valence system disturbance: A proof of concept investigation," *International Journal of Psychophysiology*, vol. 98, no. 2, pp. 240–248, 2015.

[77] M. Bradley and P. J. Lang, "The International affective digitized sounds (IADS): stimuli, instruction manual and affective ratings." Center for the Study of Emotion and Attention, Tech. Rep., 1999.

[78] H. Gunes and M. Pantic, "Automatic, Dimensional and Continuous Emotion Recognition," *International Journal of Synthetic Emotions*, vol. 1, no. 1, pp. 68–99, Jan. 2010.

[79] M. A. Nicolaou, H. Gunes, and M. Pantic, "Output-associative RVM regression for dimensional and continuous emotion prediction," *Image and Vision Computing*, vol. 30, no. 3, pp. 186–196, Mar. 2012.

Leandro A. Bugnon received the Bioengineering degree (Hons.) from Universidad Nacional de Entre Ríos (UNER), Argentina, in 2013 and is currently pursuing the Ph.D. degree on Engineering oriented to Computational Intelligence, Signals and Systems from Universidad Nacional del Litoral (UNL), Argentina. He is with the Research Institute for Signals, Systems and Computational Intelligence - sinc(i) (FICH-UNL/CONICET) since 2013. From 2013 he has a doctoral scholarship at the National Scientific and Technical Research Council (CONICET). His research interests include automatic learning, pattern recognition, signal and image processing, with applications to affective computing and biomedical signals.

Rafael Calvo is Professor at the University of Sydney, and ARC Future Fellow. He has worked in Carnegie Mellon University, Universidad Nacional de Rosario (Argentina) and on sabbaticals at the University of Cambridge and the University of Memphis. Rafael also has worked as an Internet consultant for projects in the US, Australia, Brazil, and Argentina. He is the author of two books and many publications in the fields of learning technologies, affective computing and computational intelligence. Rafael is Editor of the Oxford Handbook of Affective Computing and co-author of "Positive Computing" (MIT Press) with Dorian Peters.

Diego Milone received the Bioengineering degree (Hons.) from National University of Entre Rios (UNER), Argentina, in 1998, and the Ph.D. degree in Microelectronics and Computer Architectures from Granada University, Spain, in 2003. He was with the Department of Bioengineering and the Department of Mathematics and Informatics at UNER from 1995 to 2002. Since 2003 he is Full Professor in the Department of Informatics at National University of Litoral (UNL). From 2009 to 2011 was Director of the Department of Informatics and from 2010 to 2014 was Assistant Dean for Science and Technology. Since 2006 he is a Research Scientist at the National Scientific and Technical Research Council (CONICET). Since 2015 he is Director of the Research Institute for Signals, Systems and Computational Intelligence (CONICET-UNL). His research interests include statistical learning, pattern recognition, signal processing, neural and evolutionary computing, with applications to speech recognition, affective computing, biomedical signals and bioinformatics.