Features relevance analysis for emotion classification with physiological sensors

Christelle Godin, Fabrice Prost-Boucle, Aurélie Campagne, Sylvie Charbonnier, Stéphane Bonnet, Audrey Vidal

FEATURES RELEVANCE ANALYSIS FOR EMOTION CLASSIFICATION WITH PHYSIOLOGICAL SENSORS

C. Godin[1], F. Prost-Boucle[1], A. Campagne[2], S. Charbonnier[3], S. Bonnet[1], A.Vidal[1]
[1]*Univ. Grenoble Alpes, F-38000 Grenoble, France*
CEA, LETI, MINATEC Campus, F-38054 Grenoble, France
[2]*LPNC, Université Pierre Mendes France, BP 47, F-38040 Grenoble Cedex 9, France*
[3] *Gipsa-Lab Univ. Grenoble Alpes & CNRS, F-38402 Grenoble, France*
christelle.godin@cea.fr, aurelie.campagne@upmf-grenoble.fr, sylvie.charbonnier@gipsa-lab.grenoble-inp.fr, stephane.bonnet@cea.fr, audrey.vidal@cea.fr

Keywords: Emotion recognition, physiological signals, classification, feature selection, heart rate variability, galvanic skin response, eye blinking, arousal, valence, DEAP, MAHNOB-HCI.

Abstract: Emotion classification using physiological sensors can be used for a wide range of applications in the areas of wellness, medicine, entertainment, sport, learning, advertising, human computer interfaces among other. The associated technologies need to be improved in order to be really efficient in real life applications. The sensors should be less obtrusive as possible, and the algorithms that estimate emotion most accurate as possible. The knowledge of the most relevant features for classifying emotions is crucial for both objectives by allowing to select a reduced set of sensors and by optimizing classification algorithms performances. In this paper we analyze the relevance of several features extracted from peripheral physiological sensors by using two databases freely available to the research community. The use of two separate databases allow to analyze if some features are relevant independently from the way the emotions are elicited and from the material used for experiments. We find features extracted from galvanic skin response (GSR) to be relevant for both databases. Eye closing rate and variance of zygomatic electromyography (EMG), only available in one database, are relevant for respectively arousal and valence. The hearth rate variability (HRV) is relevant but only for one of the databases, using an electrocardiogram (ECG) whereas for the other database photopletismography (PPG) was used. Only with a few set of well selected feature and sensors we reach classification performances similar to literature classifiers using more features.

1 INTRODUCTION

Emotion estimation is a topic of interest for intelligent interaction. If a machine is able to recognize the emotional state of its user, it will be possible, for example, to adapt the way the machine interacts with the user so as to enhance the user experience (André et al., 2004). There are several ways to recognize emotional states (Mauss and Robinson, 2009). Emotion is perceptible in facial expression, in the sound of voice and also in motions (Zeng et al., 2009), which are measurable using a camera or a microphone. It is also well known that emotions imply specific brain activities and changes in activity of heart, muscles and sweet glands (Ekman et al., 1983). Hence, sensors measuring physiological activity can be used in order to estimate emotions (Picard et al., 2001). In this study we decided to focus on physiological signals in order to address ambulatory applications. From a theoretical perspective, emotional ambulatory monitoring has the potential to qualify and quantify real-life emotions, discover new emotional phenomena, model real life stimuli (Wilhelm and Grossman, 2010). From a medical perspective, emotion sensing systems could be used for example for personalized psychotherapy (Gaggioli et al., 2014), mental health monitoring (Ertin et al., 2011). Designers could use such systems in order to assess

the customer reactions when discovering a new product. They also could be used from a self-improvement perspective to increase emotional self-awareness of the user during his daily activities.

In literature, several studies to estimate emotions are reported. These studies differ in many ways. Emotions can be induced by various activities : driving (Healey and Picard, 2005), looking at a movie (Fleureau et al., 2012), playing a video game (Yannakakis et al., 2014). The way to characterize emotional states is also different. In some studies, emotions are considered to be discrete states (Healey, 2000) according to discrete emotion theory (Ekman, 2005). In other studies, they are classified using a space in 2 or 3 dimensions such as valence, arousal, dominance (Chanel et al., 2007) according to dimensional theory (Posner et al., 2005). The same physiological signal can be acquired using different kind of sensors (commercial or homemade ones). Different signal processing techniques can also be used in order to extract features from physiological signals and classify emotions. All these differences make it difficult to compare the results of those studies. Comparison between databases is all the more difficult given that most of them are not available to the research community.

Yet, a few research teams have made their database available, with the goal to provide the scientific community with a common basis. Let us cite "Eight emotion" (Healey and Picard, 2002), "Driver"(Healey and Picard, 2005), "DEAP" (Koelstra et al., 2012) and "MAHNOB-HCI" (Soleymani et al., 2012). In all the studies using these databases, feature selection and classification algorithms were used, but the relevance of each physiological sensor and each extracted feature to emotion assessment was not analyzed in depth. Yet, selecting the right sensors and the right features is important if one wants to design a device that minimizes the number of sensors and the power consumption. Moreover it is well known that the choice of features contributes to the performance of the classifiers (Janecek et al., 2008). Finally, underlining which features are important for emotional assessment could contribute to a better understanding of physiological emotional processes.

The goal of this paper is to evaluate the relevance of particular features to emotion classification. Furthermore, we will aim at identifying features that appear to be relevant whatever the database used (meaning those features could be used to classify emotions whatever the way emotions are induced). To achieve this goal, we explore and compare peripheral signals from the DEAP and MANHOB-HCI databases. Indeed, those two databases were recorded in a mainly similar way, the physiological signals recorded and the emotional representation being about the same.

The outline of the paper is as follows. We will describe the content of each database in section 2, the pre-processing in section 3, the feature selection methods in section 4. The results presented in section 5, will be discussed in section 6.

2 DATABASES

DEAP and MANHOB-HCI (hereafter called MAHNOB) databases contain behavioral and physiological data measured in participants watching small videos (of around 1 min). A comparative analysis of their content is presented in Table 1. Their respective experimental protocol can be found in (Koelstra et al., 2012) and (Soleymani et al., 2012). Several differences observed between the two databases, on the video type, the number of participants, the number of videos per participant, and the kind of behavioral and physiological signals recorded.

Table 1: content of the DEAP and MAHNOB-HCI databases (+=available, -= unavailable)

	DEAP	MAHNOB-HCI
Video type	video clip	movie
Duration	1min	Approx. 1min
P=Number of participant	32	27
V=Number of video per participant	40	20
Emotion assessment labels	Arousal, Valence, Dominance, Liking, Familiarity	Arousal, valence, dominance, predictability
EEG	+	+
GSR	+	+
ECG	-	+
PPG	+	-
Respiration	+	+
Temperature	+	+
EMG	+	-
EOG	+	-
Other	-	Camera, eye gaze,

		sounds

For both databases, emotions were assessed on basis of their arousal level, their valence and their dominance. The two first indicators are the most widely used to characterize emotions in a dimensional scale (Posner et al., 2005). Arousal reflects the emotional activation/intensity (from calm to excited/from low to high intensity), valence reflects pleasure associated with the emotion (from pleasant to unpleasant) and dominance the coping potential at the emotional situation (from low to high control). Each label is ranked from 0 to 10 by each subject after viewing the video.

3 PRE-PROCESSING

3.1 Normalization of the emotional assessments (labels)

In our study, in order to minimize inter-individual variability, we normalized labels for each participant. For participant number μ, for the video number γ we considered the label given by:

$$l^{\mu,\gamma} = (l_0^{\mu,\gamma} - \overline{l_0}^{\mu})/\sigma_{l_0}{}^{\mu} \quad (1)$$

where the label can be $l = (v, a, d)$ for valence, arousal and dominance, $l_0^{\mu,\gamma}$ is the label given in the database, $\overline{l_0}^{\mu}$ and $\sigma_{l_0}{}^{\mu}$ are respectively the mean and standard deviation of the label over the V videos for participant number μ. In this paper, we consider only valence and arousal, the labels the most commonly used in literature.

However, we noticed that in both databases dominance was highly correlated to valence (correlation score is 0.82 for DEAP and 0.9 for MAHNOB when we consider the labels and the features averaged by video over the participants). This could be due to a high link between emotion and motivation generated by both sets of videos.

3.2 Emotion classes

In most of the papers considering emotional state estimation, the labels levels are divided into classes of intensity. For each label (valence, arousal), two classes are considered in DEAP (Koelstra et al., 2012) whereas three classes are used in MAHNOB (Soleymani et al., 2012). In order to have similar results for both databases, we consider two classes. The labels being normalized, we consider the video belongs to the low label class (H0) when $l^{\mu,\gamma} < 0$ and to the high label class (H1) when $l^{\mu,\gamma} \geq 0$.

3.3 Physiological signal

In this study, considering we want to target ambulatory applications, we have chosen to focus on peripheral physiological sensors that are wearable and non-obtrusive. From this perspective, EEG and other modalities (camera, eye gaze and sounds) are excluded. Then, we consider Galvanic Skin Responses (GSR), Electro-CardioGram (ECG) (for MAHNOB), Photo-Plethysmogram (PPG) (for DEAP), respiration amplitude, temperature, Electro-MyoGram (EMG), and Electro-OcculoGram (EOG).

DEAP signals were acquired at 512Hz sampling frequency but were down sampled to 128Hz. In MANHOB, the acquisition sampling frequency was 256Hz. In concordance with DEAP, we down sampled the signals to 128Hz.

3.4 Extracted Features

In (Koelstra et al., 2012) and (Soleymani et al., 2012), the authors propose a large list of potential features to characterize emotion. In a first approach, we used all the features proposed in both papers, for a total of a hundred parameters. We found that just a few parameters were relevant for modeling emotional state. In addition, a lot of proposed parameters were highly correlated (i.e. they represent redundant information), these features reflecting similar physiological mechanisms by definition. In this paper, we reduced the set of parameters to the list presented in table 2, in order to simplify the analysis. Each feature is identified by a feature number (given into brackets) which will be used afterwards. This leads to 15 parameters for MAHNOB and 20 parameters for DEAP (where EMG and EOG are recorded, contrary to MAHNOB-HCI). As for labels, in order to minimize the inter-individual variability, the measured features were normalized over the videos for each participant as follows:

$$f_i^{\mu,\gamma} = (f_{i,0}^{\mu,\gamma} - \overline{f_{i,0}}^{\mu})/\sigma_{f_{i,0}}{}^{\mu} \quad (2)$$

with $i = \{1, \dots, N\}$, N the number of features to be analyzed (N=15 for MAHNOB et N=20 for DEAP), $f_{i,0}^{\mu,\gamma}$ are the features extracted from the database

before normalization, $\overline{f_{i,0}}^{\mu}$ and $\sigma_{f_{i,0}}{}^{\mu}$ their mean and standard deviation over the V videos for the participant number μ.

Table 2: features extracted from physiological measures.

Modality	Extracted features
GSR	(1) Average of the derivative, (2) % of neg. samples in the derivative, (3) number of local minima
ECG/PPG	(4) Average of heart rate, (5) average of inter-beat intervals, (6) standard deviation of heart rate, (7) Root Mean Square of Successive Differences (RMSSD), (8) Standard Deviation of Successive Differences (SDSD), (9) Heart Rate Variability (HRV) power in the bands VLF [0.01-0.04] Hz, (10) LF [0.04-0.15] Hz and (11) HF [0.15-0.5] Hz,
Respiration	(12) Standard deviation, (13) range (greatest breath), (14) average peak to peak time
Temperature	(15) Average skin temperature
EMG	(16) Zigomatic variance (17) trapezius variance
EOG	(18) Horizontal Variance (19) Vertical variance (20) Eye blinking rate

4 FEATURE SELECTION

A lot of feature selection methods are used in the literature (Guyon and Elisseeff, 2003), (Saeys et al., 2007) in order to analyse a feature relevance and to select a subset of features. They are generally divided into filters and wrappers. Filter methods rank the features independently of a classifier by giving a score for each feature, estimating a class separability criteria. Wrapper methods are classifier dependent. They select the subset of features that is the most relevant for a given classifier. In this section, we describe the feature selection methods chosen and the way they are implemented in our analysis.

4.1 Correlation

Given labels are continuous values, correlation between features and labels appears to be a natural method to use.

We use the correlation coefficient given by:

$$R_{i,l} = \frac{\sum_{\gamma=1}^{V}\sum_{\mu=1}^{P}(f_i^{\mu,\gamma} - \bar{f}_i)(l^{\mu,\gamma} - \bar{l}_l)}{\sum_{\gamma=1}^{V}\sum_{\mu=1}^{P}(f_i^{\mu,\gamma} - \bar{f}_i)^2(l^{\mu,\gamma} - \bar{l}_l)^2} \quad (3)$$

This coefficient reflects the linear dependence between the feature f_i and the label l. If those variables are perfectly linearly dependant ($f_i = \alpha\, l$), then $R_{i,l} = \pm 1$, the sign corresponding to the sign of α. If there is no linear link between label and feature, $R_{i,l} = 0$. Feature relevance can then be ranked by sorting the computed absolute values of the coefficient in ascending order.

4.2 Fisher score

In classification tasks, the Fisher score is a traditional method for feature selection. The objective of this score is to evaluate the ratio between inter-class variability and intra-class variability. It is given by:

$$SC_{i,l} = \frac{\sum_{c=1}^{C} n_c\,(\mu_{i,c} - \mu_c)^2}{\sum_{c=1}^{C} n_j\,{\sigma_{i,c}}^2} \quad (4)$$

where C is the number of classes, n_c the number of samples of class c , $\mu_{i,}$ is the mean value of feature f_i over the dataset, $\mu_{i,c}$ and $\sigma_{i,c}$ are the mean and standard deviation of f_i on class c.

The best feature is the one with the highest inter-class variability and the lowest intra-class variability. Hence, the best features are ones which have the highest scores SC.

We used this Fisher score in our study for the classification task by considering two classes for each label (C=2: low label, high label).

We also considered the feature score obtained by considering each video as a separate class (C=V). Indeed, we observed for each video a very large variability in the labels between participants. This result suggests that it was not easy for the participants to rate the videos in the (arousal, valence, dominance) space. The Fisher analysis with V classes was done in order to analyse the feature relevance independently of the way the labels have been assigned.

4.3 Other feature selection methods

We also test other feature selection methods extracted from (Zhao et al., 2010). Those methods are Chi-square Score (Liu and Setiono, 1995), Gini index (Gini, 1912), Information Gain (Cover and Thomas, 2012), CFS (Hall and Smith, 1999) and FCBF (Yu and Liu, 2003). In order to analyse the results, for each method we rank the features, the number 1 being the first selected feature and the number 20 the last one.

4.4 Bayes classification

Finally, we considered the binary classification task (low label, high label, C=2).

We used a Naïve Bayes classifier. This classifier assumes that the features are independent and can be mapped for each class by a normal distribution. $N(\mu_{i,c}, \sigma_{i,c})$. The first step of the classification consists in learning the parameters of the classifier. For the naïve Bayes classifier, the parameters to learn are the mean and variance of the features over each class $(\mu_{i,c}, \sigma_{i,c})$. Then in the evaluation phase, for each example, the probability of its membership to each class is computed. The class with the maximum probability is allocated to the example. The features are selected using a selective forward search using Bayes classifier. Beginning with an empty set of features, we add, at each step, the new feature that (combined with the previously selected features) results in the highest classification accuracy.

In order to evaluate the performance of the classifiers, we use the accuracy and its 95% upper and lower bounds.

To evaluate the generalization power of their classifier, (Koelstra et al., 2012) trained a classifier for each participant and performed a leave-one video out cross-validation. (Soleymani et al., 2012) trained a participant independent classifier by considering all the examples of their database and performed a leave one participant out cross-validation. In our opinion, the second approach is better suited when one wants to select features whose relevance does not depend of the databases used or of the participants. Our assessment criterion is the mean percentage of accuracy over all the participants

5 RESULTS

5.1 Features vias label relevance

In this section, we analyse how features are related to labels (valence, arousal).

In order to identify which feature should be relevant for the classification task, we use Fisher score introduced in 2.2. The results are illustrated in figure 1. Each bar graph represents the value of the score for each of the numbered features in both databases (DEAP in blue and MAHNOB in red). The highest bars represent the most relevant features. The absolute value of the correlation gives a new indicator about the feature's relevance. The correlation's sign indicate if an increase of the label is associated to an increase or a decrease of the feature. Correlation results are presented in figure 2.

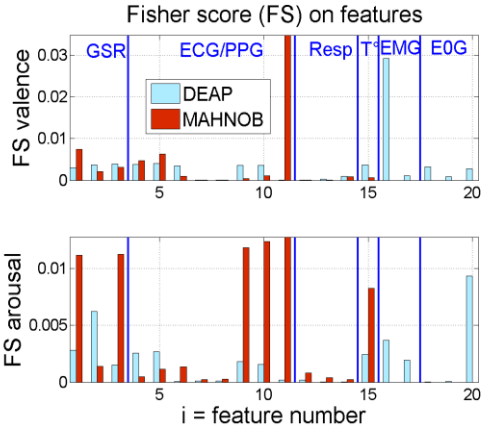

Figure 1: Fisher score of each feature for the 2 class problem for valence (high figure) and arousal (low figure).

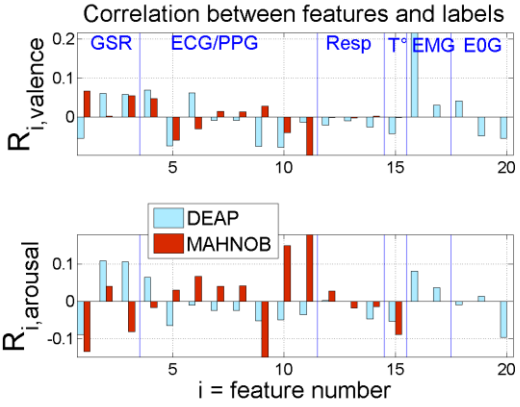

Figure 2: correlation values between features and valence (high figure) and arousal (low figure).

For valence, the most relevant features for DEAP and MAHNOB are respectively the numbers 16 (zygomatic EMG) and 11 (HRV in HF). For zigomatic, the sign of the correlation coefficient is positive, probably because a positive emotion generates smiles. In contrast, an inverse pattern is observed for the high frequency HRV, which appears higher for negative valence.

For arousal, feature 20 (eye blinking rate) is relevant for DEAP and corresponds to a decrease of the eye blinking rate with arousal. The Fisher and correlation scores tend to show that features 9, 10, 11, are relevant for MAHNOB also. The very low frequency HRV decreases whereas low and high frequencies increase with arousal. For both databases, features 1 and 15 (Average of the GSR derivative, skin temperature) appear to be relevant for arousal estimation. The correlation sign indicate

that skin resistance and skin temperature decreases with arousal. However it should be noted that even the highest correlations and Fisher scores stay low. The tables below summarize the results obtained with the other feature selection methods from (Zhao et al., 2010).

Table 3: rank of each feature (line) with respect to each feature selection method (arrow)

Valence DEAP						
N°	**FS**	**X2**	**Gini**	**Info Gain**	**CFS**	**FCBF**
1	**11**	**3**	**3**	**3**		
2	6	4	4	4		
3	3	5	5	5		
4	4	6	6	6		
5	2	7	7	7		
6	9	8	8	8		
7	19	9	9	9		
8	18	10	10	10		
9	8	11	11	11		
10	7	12	12	12		
11	20	13	13	13		
12	17	14	14	14		
13	16	15	15	15		
14	14	16	16	16		
15	5	17	17	17		
16	**1**	**1**	**1**	**1**	**1**	**1**
17	13	18	18	18		
18	**10**	**2**	**2**	**2**	**2**	**2**
19	15	19	19	19		
20	12	20	20	20		

Valence MAHNOB						
N°	**FS**	**X2**	**Gini**	**Info Gain**	**CFS**	**FCBF**
1	**2**	**2**	**2**	**2**		
2	6	3	3	3		
3	5	4	4	4		
4	4	5	5	5		
5	3	6	6	6		
6	8	7	7	7		
7	12	8	8	8		
8	13	9	9	9		
9	11	10	10	10		
10	**7**	**1**	**1**	**1**	**1**	**1**
11	1	11	11	11		
12	15	12	12	12		
13	14	13	13	13		
14	9	14	14	14		
15	10	15	15	15		

Arousal DEAP						
N°	**FS**	**X2**	**Gini**	**Info Gain**	**CFS**	**FCBF**
1	4	2	2	2		
2	2	3	3	3		
3	11	4	4	4		
4	6	5	5	5		
5	5	6	6	6		
6	16	7	7	7		
7	14	8	8	8		
8	15	9	9	9		
9	9	10	10	10		
10	10	11	11	11		
11	12	12	12	12		
12	13	13	13	13		
13	19	14	14	14		
14	18	15	15	15		
15	7	16	16	16		
16	3	17	17	17		
17	8	18	18	18		
18	20	19	19	19		
19	17	20	20	20		
20	**1**	**1**	**1**	**1**	**1**	**1**

Arousal MAHNOB						
N°	**FS**	**X2**	**Gini**	**Info Gain**	**CFS**	**FCBF**
1	**5**	**2**	**2**	**2**	**1**	**2**
2	7	5	5	5		
3	3	6	6	6		
4	12	7	7	7		
5	8	8	8	8		
6	9	9	9	9		
7	14	10	10	10		
8	13	11	11	11		
9	4	4	4	4	2	3
10	**2**	**1**	**1**	**1**	**3**	**1**
11	**1**	**3**	**3**	**3**		
12	11	12	12	12		
13	10	13	13	13		
14	15	14	14	14		
15	6	15	15	15		

Feature 1 (GSR average of derivative) stays relevant for all the feature selection methods and for both valence and arousal. The most relevant features are not common between the databases. For DEAP they are feature 16 (EMG zygomatic) and 18 (EOG horizontal) for valence and 20 (eye blinking rate) for arousal not available for MAHNOB. For MANHOB they are feature 10 (HRV LF) and 11 (HRV HF) computed from ECG not available for DEAP.

Those differences between MAHNOB and DEAP may be justified by the fact that MAHNOB does not contain EMG and EOG and by the fact that heart rate is measured via PPG in MAHNOB whereas it is

measured by ECG in DEAP. The ECG signal allows a better localisation of R peaks than PPG signal.

5.2 Fisher score of individual features on videos

In the previous section, small values for correlation and Fisher scores were obtained suggesting a weak link between features and labels. The difficulty for the participant to choose a level of arousal and valence could justify this result. Indeed, for each video, we noted a large variability in rating between participants for each label: the average standard deviation of video's valence and arousal, after normalization, are respectively 0.68 and 0.89 for DEAP and 0.51 and 0.76 for MAHNOB.

One approach could be to consider that every participant reacts to a given video with the same emotion. In that case, each video could be considered as an emotional class. In order to analyse the relevance of each feature in this perspective we have computed the Fisher score of each feature by considering each video as an emotional class (videos classification task). The results are presented in figure 3. For a better visualization of the results, it should be noted that scores are plotted using a y logarithmic axis due to their large difference in range.

For DEAP, the Fisher scores remain all very low, which suggests that emotional videos used in this database do not induce strong emotional reactions or that the reaction differs from one participant to another. We can also assume that HRV in frequency bands is not a relevant measure when PPG is used, due to a lack of precision in the identification of peaks R. Eye closing rate (feature number 20) remains the most relevant feature.

In contrast, for MAHNOB, the scores are actually higher than those obtained with 2 classes and this more particularly for features 9, 10 and 11 (HRV in frequency bands). This result suggests that individual physiological reactions are more related to the content of the videos than to the ratings of emotions using arousal and valence labels.

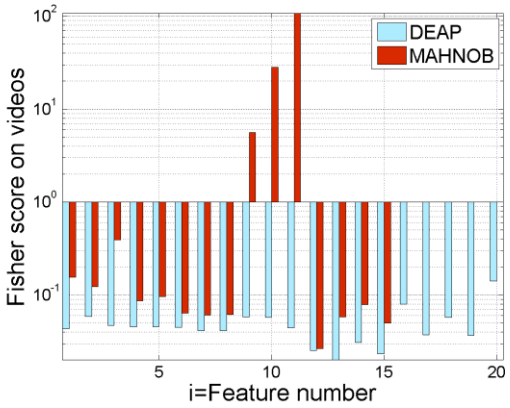

Figure 3: Fisher scores, each video being considered as one class.

5.3 Bayes Classification

The results obtained by a selective forward search using Naïve Bayes classifier are presented in figure 4 for each label and for each database. The *x* axis shows the number of features used by the classifier. The *y* axis presents the mean accuracy reached by the classifier and its 95% confidence bounds. As can be seen, the classification rates are low, only slightly higher than random classifiers, but they are comparable to those of (Koelstra et al., 2012), which were 62% for valence and 57% for arousal for DEAP, using 106 features. The same results are reached here with one carefully selected feature. Those of MAHNOB in (Soleymani et al., 2012) were 46% for arousal and 45% for valence for the 3 class problem (low, medium, high label). They cannot be directly compared with our results, which are obtained for 2 classes only. Moreover in (Soleymani et al., 2012) they decided to establish the classes using emotional keywords instead of valence and arousal ratings. As our main purpose was to compare the feature relevance we chose to consider the same approach for both databases.

As can be seen, the combination of several features does not strongly improve the results. In Table 3, the selected features are presented in their order of selection. The same relevant features as those observed in the previous analyses are selected. Feature 10 and 11 appear first and second in MAHNOB for arousal, and valence. Feature 6 (standard deviation of heart rate) also seems to be interesting. For GSR we find again feature 16 (zygomatic) for valence and feature 2 (related to GSR) for arousal.

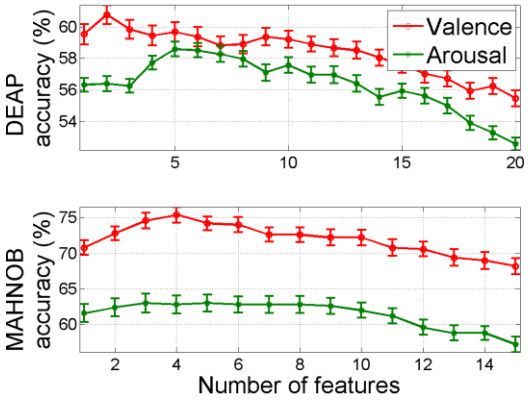

Figure 4: classification accuracy with 95% lower and upper bound for each database and each label (2 classes) with respect to the number of features (forward selection).

Table 4: Features in selected order by forward search

DEAP valence	16 2
MAHNOB valence	10 11 1 14
DEAP arousal	2 15 17 20 16
MAHNOB arousal	10 11 6

Figure 5 shows that better results are obtained for MAHNOB than for DEAP. This can be explained by the fact that MAHNOB's films induced stronger emotions than DEAP's video clips. Another reason can be that PPG was not good enough to estimate correctly HRV, which is the feature that obtains the best results for MAHNOB.

6 DISCUSSION

The objective of this work was to determine which features extracted from physiological peripheral sensors are relevant for emotion assessment. Two databases freely available to the research community were used. Several feature selection analysis were done.

The first result is that the classification task using valence and arousal labels was not easy, for both databases. Indeed correlation coefficients, Fisher scores, and classification accuracies were low. One reason can be that valence and arousal labels are not representative enough of the participants' physiological reaction. Participants may have found it difficult to rate their emotional state. This is confirmed by the high scores obtained by HRV power in three frequency bands on DEAP database, when videos are considered as emotional classes. This result presumes that better classification results could be obtained using other labels than labels based on self-evaluation, which are very subjective. In (Soleymani et al., 2012) they decided to use emotional keywords in order to create three classes for valence and arousal. However, classification rates were also relatively low. Another reason can be that the videos watched by the participants are not emotionally stimulating enough and that other reactions are superimposed on emotional reactions.

A second result is that better results were obtained with the MAHNOB database than with the DEAP database, whatever the criteria used: correlation coefficients, classification accuracies and fisher scores. Several reasons can justify this result. Firstly, we assume that movies (in MAHNOB) induced more emotions and less emotional variability between participants than video clips (in DEAP). This is coherent with the standard deviations of labels that were lower in MAHNOB. In addition, higher Fisher scores were observed for features HRV and (to a lesser extent) for features related to skin response, which are known to be sensitive to emotions (Lang et al., 1993). However, it should be noted that HRV estimation in the three frequency bands is more accurate when R peaks are extracted from ECG signal (available in MAHNOB) than from PPG signal (available in DEAP). This may partially explain the differences between databases. Finally, it is possible that movies and video-clips differ by the nature of induced discrete emotions.

It was also interesting to identify the relevant features for each label, their variation, and whether they correspond to well-known physiological reactions. The most important feature seems to be the HRV power in three frequency bands. Those features are commonly used in emotion estimation (Kreibig, 2010). It was shown that power in very low frequency band decreases whereas power in low and high frequencies band increases with arousal. This feature was only relevant for MAHNOB perhaps because it was not correctly estimated using PPG in DEAP.

For DEAP, eye blinking rate for arousal and variance of zygomatic EMG for valence, were both well-known relevant features for respectively vigilance and attention (Wei and Lu, 2012; Campagne et al., 2005), and smiles (Fleureau et al., 2012). Unfortunately eye blinking rate was not available in MANHOB-HCI database. One possibility would be to identify it from EEG frontal signal (Roy et al., 2014). Features extracted from GSR show a significant correlation with arousal for both DEAP and MANHOB databases. This result is not surprising given the close relationship found in the literature between the skin activity level and individual's emotional state (Lang et al., 1993).

Finally, in previous studies (Koelstra et al., 2012) (Soleymani et al., 2012), the optimal size of the feature space for emotion classification had not been evaluated. Using a forward search algorithm associated with a Bayesian classifier, optimal accuracy was achieved only with a few set of features (from 1 to 5) and this accuracy is equivalent to previous study (Koelstra et al., 2012).

7 CONCLUSION

In this work, we aimed at identifying user and database independent features for emotional estimation, using wearable physiological sensors. The features related to GSR were found to be the only ones relevant and available for DEAP and MAHNOB databases (both freely available to the research community). Other features were found to be more relevant for one of the two databases, such as features extracted from ECG in MAHNOB or those extracted from EOG and zygomatic EMG in DEAP. Those results should be confirmed by new experiments, which should use the most complete set of sensors possible, including all the signals recorded in DEAP and MANHOB databases, in order to obtain result comparable with those databases. It would also be interesting to measure both PPG and ECG in order to confirm our hypothesis that PPG is not precise enough for HRV spectral analysis.

AKNOWLEGMENT

Portions of the research in this paper uses the MAHNOB database collected by Professor Pantic and the iBUG group at imperial College London, and in part collected in collaboration with Prof. Pun and his team of University of Geneva, in the scope of MAHNOB project financially supported by the European Research Council under the European Community's 7th Framework Programme (FP7/2007-2013)/ERC Starting Grant agreement N°203143.

REFERENCES

André, E., Rehm, M., Minker, W., Bühler, D., 2004. Endowing spoken language dialogue systems with emotional intelligence, in: proceedings affective dialogue systems 2004. Springer, pp. 178–187.

Chanel, G., Ansari-Asl, K., Pun, T., 2007. Valence-arousal evaluation using physiological signals in an emotion recall paradigm, in: Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on. pp. 2662–2667.

Cover, T.M., Thomas, J.A., 2012. Elements of information theory. John Wiley & Sons.

Ekman, P., 2005. Basic Emotions. Psychol. Rev. - PSYCHOL REV 99, 45 – 60.

Ekman, P., Levenson, R.W., Friesen, W.V., 1983. Autonomic nervous system activity distinguishes among emotions. Science 221, 1208–1210.

Ertin, E., Stohs, N., Kumar, S., Raij, A., al' Absi, M., Shah, S., 2011. AutoSense: Unobtrusively Wearable Sensor Suite for Inferring the Onset, Causality, and Consequences of Stress in the Field, in: Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems, SenSys '11. ACM, New York, NY, USA, pp. 274–287.

Fleureau, J., Guillotel, P., Huynh-Thu, Q., 2012. Physiological-Based Affect Event Detector for Entertainment Video Applications. IEEE Trans. Affect. Comput. 3, 379–385.

Gaggioli, A., Pallavicini, F., Morganti, L., Serino, S., Scaratti, C., Briguglio, M., Crifaci, G., Vetrano, N., Giulintano, A., Bernava, G., Tartarisco, G., Pioggia, G., Raspelli, S., Cipresso, P., Vigna, C., Grassi, A., Baruffi, M., Wiederhold, B., Riva, G., 2014. Experiential Virtual Scenarios With Real-Time Monitoring (Interreality) for the Management of Psychological Stress: A Block Randomized Controlled Trial. J. Med. Internet Res. 16, e167.

Gini, C., 1912. Variabilite e mutabilita (Italian). Mem. Metodol. Stat.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Hall, M.A., Smith, L.A., 1999. Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper., in: FLAIRS Conference. pp. 235–239.

Healey, J.A., 2000. Wearable and automotive systems for affect recognition from physiology (Thesis). Massachusetts Institute of Technology.

Healey, J.A., Picard, R.W., 2005. Detecting stress during real-world driving tasks using physiological sensors. IEEE Trans. Intell. Transp. Syst. 6, 156–166.

Healey, J., Picard, R.W., 2002. Eight-emotion Sentics Data, MIT Affective Computing Group.

Janecek, A., Gansterer, W.N., Demel, M., Ecker, G., 2008. On the Relationship Between Feature

Selection and Classification Accuracy., in: FSDM. Citeseer, pp. 90–105.

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I., 2012. Deap: A database for emotion analysis; using physiological signals. Affect. Comput. IEEE Trans. On 3, 18–31.

Kreibig, S.D., 2010. Autonomic nervous system activity in emotion: A review. Biol. Psychol., The biopsychology of emotion: Current theoretical and empirical perspectives 84, 394–421.

Lang, P.J., Greenwald, M.K., Bradley, M.M., Hamm, A.O., 1993. Looking at pictures: Affective, facial, visceral, and behavioral reactions. Psychophysiology 30, 261–273.

Liu, H., Setiono, R., 1995. Chi2: Feature selection and discretization of numeric attributes, in: 2012 IEEE 24th International Conference on Tools with Artificial Intelligence. IEEE Computer Society, pp. 388–388.

Mauss, I.B., Robinson, M.D., 2009. Measures of emotion: A review. Cogn. Emot. 23, 209–237.

Picard, R.W., Vyzas, E., Healey, J., 2001. Toward machine emotional intelligence: analysis of affective physiological state. IEEE Trans. Pattern Anal. Mach. Intell. 23, 1175–1191.

Posner, J., Russell, J.A., Peterson, B.S., 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev. Psychopathol. 17, 715–734.

Roy, R.N., Charbonnier, S., Bonnet, S., 2014. Eye blink characterization from frontal EEG electrodes using source separation and pattern recognition algorithms. Biomed. Signal Process. Control 14, 256–264.

Saeys, Y., Inza, I., Larrañaga, P., 2007. A review of feature selection techniques in bioinformatics. bioinformatics 23, 2507–2517.

Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M., 2012. A Multimodal Database for Affect Recognition and Implicit Tagging. IEEE Trans. Affect. Comput. 3, 42–55.

Wei, Z.-P., Lu, B.-L., 2012. Online vigilance analysis based on electrooculography, in: The 2012 International Joint Conference on Neural Networks (IJCNN). Presented at the The 2012 International Joint Conference on Neural Networks (IJCNN), pp. 1–7.

Wilhelm, F.H., Grossman, P., 2010. Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment. Biol. Psychol. 84, 552–569.

Yannakakis, G.N., Isbister, K., Paiva, A., Karpouzis, K., 2014. Guest Editorial: Emotion in Games. IEEE Trans. Affect. Comput. 5, 1–2.

Yu, L., Liu, H., 2003. Feature selection for high-dimensional data: A fast correlation-based filter solution, in: ICML. pp. 856–863.

Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S., 2009. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. IEEE Trans. Pattern Anal. Mach. Intell. 31, 39–58.

Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., Liu, H., 2010. Advancing feature selection research. ASU Feature Sel. Repos.