Selection of Physiological Input Modalities for Emotion Recognition

Thomas Christy, Ludmila I. Kuncheva, and Kerry W. Williams
School of Computer Science, Bangor University, LL57 1UT, UK.
E-mail: t.christy@bangor.ac.uk

Technical Report # CS-TR-002-2012, Bangor University, UK

***Abstract*—Emotion recognition is an extremely challenging problem in the heart if affective computing. In many applications, for example affective gaming, it is paramount to capture emotion with minimum inconvenience to the user, and feed the result into the human-computer interface. This paper describes the results of an experiment with the recently created and released for public use DEAP data set. The set contains the recordings of 7 physiological modalities for 32 users, each user viewing 40 video clips and grading them for emotional content. Two classification tasks were considered: high/low arousal and high/low valence. Feature rankings were calculated separately for each user by the RFE-SVM method. A stability index revealed high discrepancy of these rankings across users. Arguing that everyday-life applications require a small and highly discriminatory subset of modalities, all combinations of the 7 modalities were examined. No suitable subset of modalities could be identified although the EEG modality was present in all subsets on the Pareto frontier for the two classification tasks.**

***Keywords*-Emotion classification, multiple modalities, feature selection**

I. Introduction

Emotions are an important part of the human psyche and play a vital role in our everyday life [9]. Affective Computing has seen a dramatic rise over the past decade [7], [29], [30] permeating various disciplines such as computer science, electronic engineering and psychology. Emotion can be detected using a myriad of behavioural and physiological modalities, some of which require sensors and devices to be attached to the user. The intuition is that the more modalities are used simultaneously, the more accurate the recognition will be. However, for applications such as affective gaming to become a reality, a low cost, convenient, unobtrusive and comfortable system is required. This brings up the question of *wearability* of the input devices [5].[1] It is possible that certain combinations of modalities may be sufficient for the purposes of the application, rendering the remaining modalities and their input devices redundant.

In this study we seek to find a highly discriminative subset of seven physiological modalities for emotion recognition. The recently created data set DEAP [21] was chosen for the analyses. The rest of the paper is organised as follows. A review of the main modalities for emotion detection is given in Section II. The methods involved in the modality selection analyses are detailed in Section III. The DEAP data set [21] is described in Section IV. Section V contains the experiment, and Section VI offers a discussion of the findings, our recommendations and conclusions.

II. Modalities for emotion recognition

Modalities used for emotion recognition can be grouped into two categories: physiological and behavioural as summarised in Figure 1. While a comprehensive explanation of each modality goes beyond this paper, we give a basic summary of the physiological modalities recorded in the DEAP data set.

Physiological modalities are measured from sensors attached to the body. Fairclough [13] warns about the difficulties in recognising emotion from physiological measurements. The complex relationship between experienced states and their expression via the central and peripheral nervous systems is a fundamental problem for psychophysiology which could be misunderstood by those outside the discipline. It should not be assumed that psychophysiological measurement provides a literal, isomorphic representation of a given thought, intention or emotion. With this caveat in mind, seven main physiological modalities are reviewed below.

• **Electroencephalography (EEG)** fig1 (#1)
EEG is one of the most important technologies in modern neuroscience [14]. The electrical potentials related to emotion can be projected widely in an intricate pattern across the scalp, and can therefore overlap with potentials evoked by other activities. EEG has been applied for classification of emotions in various contexts [6], [20], [32], [34] and is progressively becoming a portable lightweight technology. It is often assumed that the projections of positive and negative emotions in the left and right frontal lobes of the brain make these two emotions distinguishable by EEG. Practice has shown that the granularity of the information collected from these regions through EEG may be insufficient for detecting more complex emotions [6]. Different success rates

[1] In the rest of the paper, we will often use affective gaming as a potential application area to illustrate our vision and ideas.

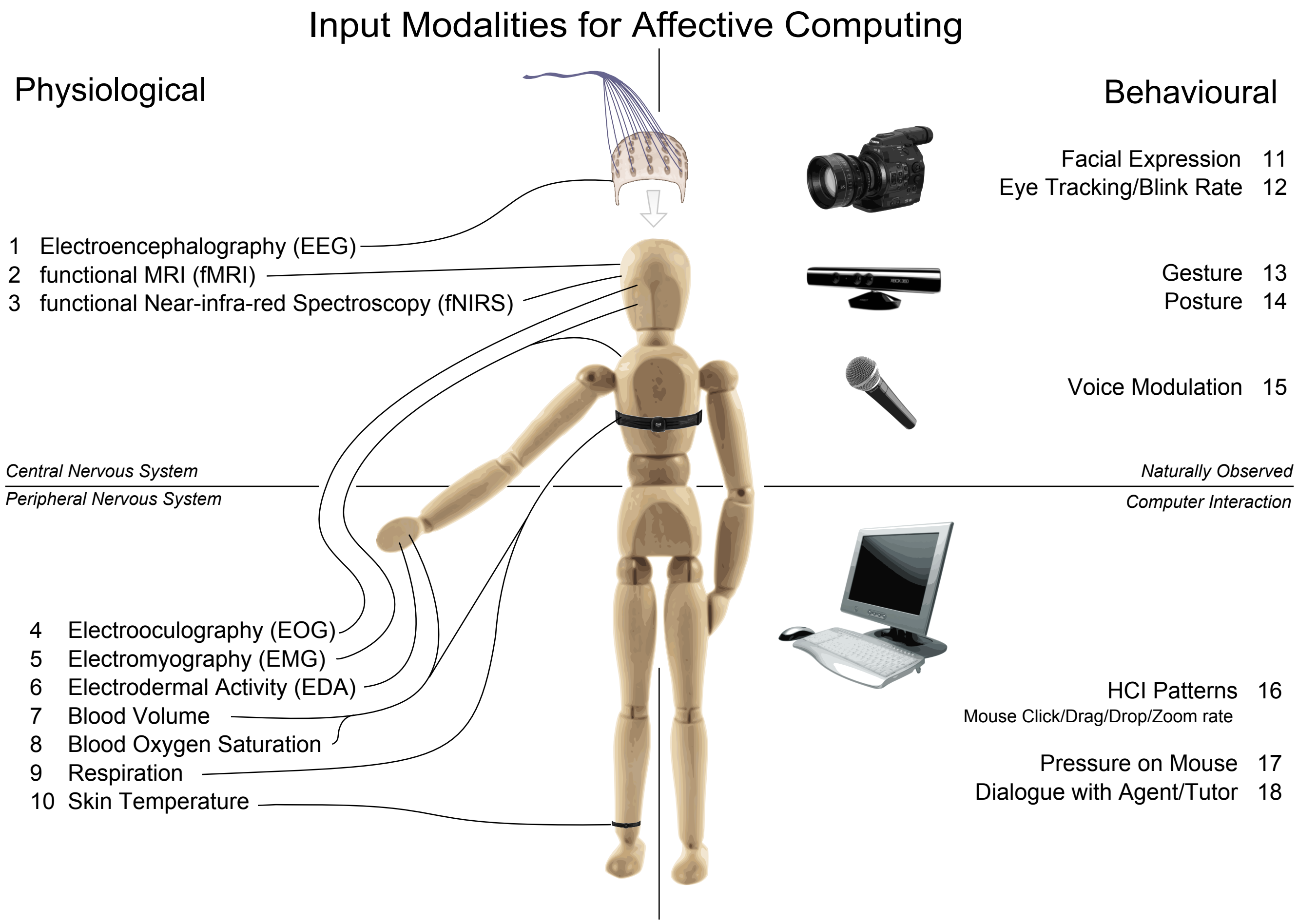

Input Modalities for Affective Computing
Physiological
Behavioural
1 Electroencephalography (EEG)
2 functional MRI (fMRI)
3 functional Near-infra-red Spectroscopy (fNIRS)
Central Nervous System
Peripheral Nervous System
4 Electrooculography (EOG)
5 Electromyography (EMG)
6 Electrodermal Activity (EDA)
7 Blood Volume
8 Blood Oxygen Saturation
9 Respiration
10 Skin Temperature
Facial Expression 11
Eye Tracking/Blink Rate 12
Gesture 13
Posture 14
Voice Modulation 15
Naturally Observed
Computer Interaction
HCI Patterns 16
Mouse Click/Drag/Drop/Zoom rate
Pressure on Mouse 17
Dialogue with Agent/Tutor 18

Figure 1. An illustration of the modalities

of emotion recognition through EEG have been reported in the literature ranging from moderate [8] to excellent accuracy [23], [28]. The reason for the inconclusiveness of the results can be explained with the different experimental set-ups, different ways of eliciting and measuring emotion response, and the type and number of distinct emotions being recognised. For example, there is no consensus about the optimal positioning of the electrodes on the scalp. Figure 2 shows the choices of electrode locations within the 10/20 positioning system, taken from four different studies. There are only two overlapping locations. This brings up the question of what the minimal number of electrodes is and where they should be placed.

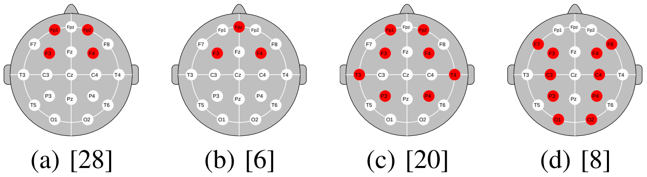

(a) [28] (b) [6] (c) [20] (d) [8]

Figure 2. Positioning of the EEG electrodes (red) for emotion recognition using the standard 10/20 scheme.

• **Electrooculography (EOG)** fig1(#4)
EOG is the measurement of the movement of the eye horizontally and vertically. EOG measurements are made by placing electrode pairs at the top and bottom, and the left and right of the eye. EOG has proven more effective for measuring emotion across different cultures compared to other facial characteristics [36]. Placing electrodes on a customer/game-player might be considered irritating and obtrusive. A high definition camera can be used instead to extract eye movement using image processing.

• **Electromyography (EMG)** fig1(#5)
It is a common observation that people display varying patterns of movement attributed to emotion. These patterns can sometimes be referred to as a nervous twitch, produced involuntarily when a person becomes anxious, nervous, excitable, etc. Professional poker players capitalise

on observing such phenomena to detect if their opponents are bluffing within a game. Electrical activity produced by muscle movement can be used to detect patterns related to emotion. To capture complex movements, EOG sensors should be placed at various body locations, which raises the same question of compromising the user's comfort as with the EEG.

- **Electrodermal Activity (EDA)** fig1(#6)

EDA, also referred to as Galvanic Skin Response (GSR), Skin Conductance Response (SCR) or Psycho Galvanic Reflex (PGR), measures the variance in electrical conductivity through the surface of the skin. EDA readings are effected through the sympathetic nervous system, making it a good indicator of stress and anxiety. EDA suffers from latency, with a delay of approximately one second for a response to be evoked, followed by approximately three seconds for the effect to dissipate. It is among the most basic and low cost physiological modalities available, and is widely used in physiological emotion recognition, including video games [3]. EDA is commonly read between two fingers on either hand, although is not limited to this area of the body [5].

- **Blood volume (Plethysmography)** fig1(#7-8)

The variation in heart rate is a good indicator of stress and anxiety. As a sign of its pervasive popularity, the use of biometric heart rate reading became the pivotal component of a television game-show, called The Chair [33]. In this show, contestants answered general knowledge questions and were expected to maintain a calm heart rate to win money. The sensing devices are typically in the form of a finger clip, which uses infra-red technology to measure simultaneously heart rate and blood oxygenation.

- **Respiration** fig1(#9)

Emotion can influence breathing rates [4], [19]. The measuring device could be a respiration belt or sensors embedded into clothing. For example, a force feedback vest with embedded breathing rate sensors already features in the avid pro-gamers' arsenal [2]. However, mainstream applications could be hindered by utilising garments to acquire data.

- **Temperature** fig1(#10)

Body temperature is affected by emotion, specifically joy, anger and sadness [24], [25], and has been used for emotion recognition in video games [5], [35].

Widespread use of emotion recognition in human-computer interface is preconditioned on selecting a highly informative subset of modalities which can be seamlessly integrated into the user's environment. In computer gaming, for example, the presence of scientific laboratory equipment would diminish the chances of a pleasurable experience even for the most enthusiastic gamer. The issue of sensor comfort and ease of use is an important topic and should become a pivotal factor in affective video game peripheral design [1], [5].

III. Feature selection

A. *Modalities as features*

Emotion recognition can be cast as a typical pattern recognition task. Emotions are rarely identified from a single snapshot of the sensors or the brain state. Each modality usually generates consecutive measurements spanning the expected duration of the emotion. One possible way forward is to concatenate all data into one single feature vector and apply the classical pattern recognition approaches and methods [12], [18]. In theory, intricate cross-modality relationships can be identified from the concatenated vector through feature selection. As a speculative example, suppose that the EEG at time t, together with the galvanic skin response half a second earlier and the facial expression a second later may together form a very indicative feature combination. However, there are downsides of the concatenation approach. There are computational challenges given that the data from any brain-measuring modality alone contains hundreds or thousands of features. Therefore identifying useful cross-modal relationships is hampered by the curse of dimensionality. The second approach is to use each modality individually and combine the label outputs.

Looking for a small and highly discriminative subset of modalities, we ran feature selection using the publicly available DEAP database[3] [21].

B. *Feature selection methods*

Feature selection is one of the most widely discussed topics in pattern recognition. A myriad of insightful and comprehensive surveys have been devoted to it [10], [15], [26]. The two major questions that a feature selection method must address are: (1) Are the features evaluated individually, and if not, how do we traverse the class of all candidate subsets? (2) What criterion do we apply to evaluate the merit of a given subset of features?

Much of the difficulties in emotion recognition are due to the complex relationship between experienced states and their expression via the central and peripheral nervous systems [13]. Such relationships may be smeared and obscured due to imperfect measuring technology. Affective computing modalities are characterised by large variability across users, sensitivity to the environment and low predictive accuracy. This makes the feature selection task difficult, and requiring large multi-user and multi-run data sets. To account for this problem, basic feature selection and classification methods should be attempted. The simplest solution is to check each feature separately (univariate methods), resulting in a ranking of the features. This approach is adopted by

[2] http://en.wikipedia.org/wiki/3rd_Space_Vest

[3] http://www.eecs.qmul.ac.uk/mmv/datasets/deap/

Koelstra et al. [21] for selection of features in emotion classification. For two classes (low valence/high valence or low arousal/high arousal), the quality of feature x is measured by

$$J(x) = \frac{|\mu_1 - \mu_2|}{\sigma_1^2 + \sigma_2^2} \tag{1}$$

where μ_1 and μ_2 are the estimates of the class means and σ_1 and σ_2 are the respective standard deviations. An empirical threshold of 0.3 is proposed and features whose criterion J is greater than this threshold form the selected set. While univariate feature selection methods have a proven record in various application areas where data has high dimensionality [2], [31], multivariate methods have been advocated for such data [27]. As true multivariate feature selection with traversing and evaluating possible subsets will still be computationally infeasible, a "pseudo-multivariate" approach can be applied. The result is again an individual ranking of the features but some interdependencies between them are taken into account. The most successful pseudo-multivariate feature ranking method to date is the Recursive Feature Elimination (RFE) based on the Support Vector Machine classifier (SVM) [16]. The SVM classifier alone is a suitable feature ranker. Using a linear kernel, the SVM classifier builds an optimal linear boundary between two classes of interest in the original feature space. The boundary is calculated so as to be as far as possible from the closest points from the opposite classes, called the support vectors. The coefficients of the boundary measure the importance of the features

$$J_{SVM}(x_i) = |w_i|, \;\; \mathbf{w} = [w_1, \ldots, w_K]^T, \tag{2}$$

where X is the feature vector, feature x_i is the ith component of X, and the boundary is given by $\mathbf{w}^T X + b$. The recursive elimination is done by starting with the whole feature set X, training an SVM classifier and dropping a given number of features with the lowest J_{SVM}. A new SVM is trained on the remaining features, and a new elimination is done. For simplicity, suppose that one feature is dropped in each elimination round. The result from the RFE-SVM is a ranking of the original features. RFE-SVM has been successfully applied for classification of fMRI data [11]. Both criteria are used here for ranking data coming from affective modalities.

C. Stability of feature ranking

Ideally, the features selected for recognising emotion will be the same for all persons, opening up the possibility of creating an emotion recognition device for the mass game user. In reality, the person-to-person differences are considerable, suggesting that an individually tailored selection approach may be better. Individual feature selection is faced with at least two difficulties. First, a number of input modalities must be kept because the subset selected for a particular may differ from the subset for the next one. This defeats the object here because we are looking to propose a single reliable, wearable and inexpensive set of modalities. Second, the selection will be done using a small training set - the labelled set for one person -, which may lead to spurious results. Therefore we consider a stability index [22] for measuring the agreement between the individual feature rankings.

The stability index is based on the concept of consistency between feature subsets. Suppose that there are two feature rankings, R_A and R_B of the T features in Y. Let A be the set of the top k features according to R_A and B be the set of the top k features according to R_B. The Consistency Index between A and B is

$$I_C(A,B) = \frac{r - \frac{k^2}{T}}{k - \frac{k^2}{T}} = \frac{rT - k^2}{k(T-k)}, \tag{3}$$

where r is the number of common features in sets A and B. The maximum value of the index, $I_C(A,B) = 1$, is achieved when A and B contain the same features (note that A and B are sets, and there is no order of the features within). The minimum value of the index is bound from below by -1. The limit value is attained for $k = \frac{T}{2}$ and $r = 0$. It should be mentioned that $I_C(A,B)$ is not defined for $k = 0$ and $k = T$. These are the trivial cases where either no feature is selected or all features are selected. They are not interesting from the point of view of comparing feature subsets, so for completeness we can assume $I_C(A,B) = 0$ for both cases. Finally, $I_C(A,B)$ will assume values close to zero for independently drawn A and B because r is expected to be around $\frac{k^2}{T}$.

A stability index for L sets, $S_1, S_2, \ldots, S_L$, all of cardinality k, coming from different rankings, is the average pairwise consistency

$$\mathcal{I}_S(\mathcal{A}(k)) = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} I_C(S_i(k), S_j(k)). \tag{4}$$

The consistency index is used here to determine to what extent the feature subsets differ across potential game users.

IV. Data and feature extraction

A. DEAP dataset

The dataset contains physiological signals recorded for 32 different users. Each user was shown 40 music video clips and for each video seven physiological modalities were recorded: EEG, EMG, EOG, Skin Temperature, Respiration pattern, Blood volume pressure and GSR. The 40 video clips were carefully pre-selected so that their intended arousal and valence values span as large as possible an area of the arousal-valence space. Each participant was asked to grade each clip after the viewing giving discrete values from 1 to 9 for arousal, valence, dominance and liking. The participant's labels for arousal and valence were taken as the true labels of the predominant emotion experiences throughout viewing the clip.

The EEG channels were recorded using 32 sensors (channels) placed on the subject's head. Eight further channels were used for the remaining physiological modalities. The EMG and EOG were recorded using four sensors positioned on the face to register the electrical activity linked with respective muscle movement. The GSR was captured using sensors places on the subjects ring and middle fingers, and body temperature was recorded using the a sensor on the subjects little finger. Breathing patterns and heart rate were captured using a respiration belt.

Our analyses were carried out on the preprocessed version of the data provided in MATLAB format. The data for each of the 32 users are stored in a 3d array of size 40(videos)×40(modality channels)×8046(values). The values span 63 seconds, sampled at 128 Hz: 60 seconds trial and 3 seconds pre-trial for establishing a baseline that was subsequently removed. The preprocessing included also removing artefacts from EOG and applying a bandpass frequency filter from 4.0–45.0 Hz to the EEG data.

B. Feature extraction

Following Koelstra et al. [21], we extracted 261 features from the seven modalities as detailed in Table I. We were unable to reproduce exactly the set from their study but added medians and interquartile ranges as estimates of central tendency which are less sensitive to the presence of outliers in the data.[4] The calculated features revealed anomalies in the data, which could be attributed to a temporary malfunctioning of some sensors or change of examination conditions. This was not unexpected in a real life experiment even in a strictly controlled environment and carefully selected stimuli. The problems with extracting reliable features from the data amplify the concerns about using the physiological modalities for affective gaming, especially when real-time emotion recognition is required.

C. Visualisation

To examine the sources of variability in the data we calculated the principal components of the standardised data and plotted the 1280 data points in the space of the first two principal components. The labels of the points in Figure 3 (a) correspond to the 32 users. Figures 3 (b) and (c) show the labels for valence and arousal. The plots show that the arousal and valence classification into low/high is likely to be quite challenging as there are no clear patterns of the classes. On the other hand, the data for each user it tightly grouped indicating that users could be easily recognised.

Table I
THE 261 FEATURES EXTRACTED FROM THE 7 MODALITIES

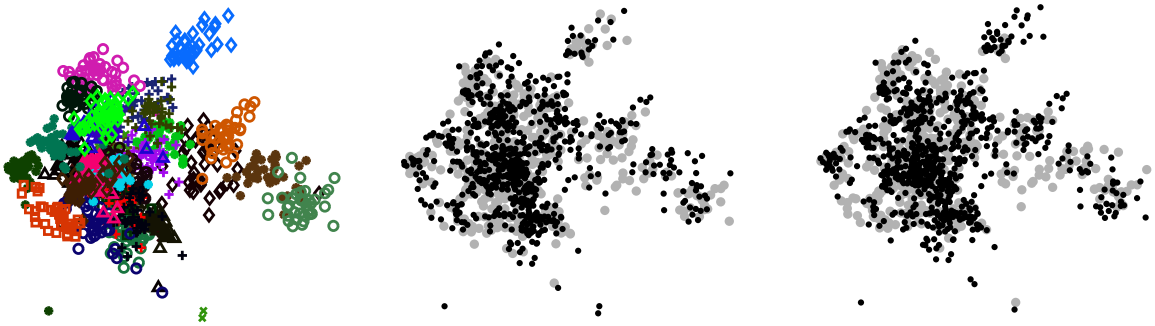
Modality	Features
EEG 32 channels	4–8 Hz theta 8–10 Hz alpha 1 (slow alpha) 10–12 Hz alpha 2 (relaxed and alert alpha) 12–20 Hz beta 1 20–40 Hz beta 2 (high beta) 40–50 Hz gamma for each electrode (6×32 = 192 features)
EMG, EOG 4 channels	Mean of the signal Standard deviation Number of local maxima Median Interquartile range (5×4 = 20 features)
GSR 1 channel	Mean of the signal Mean derivative Mean across the negative derivative values Proportion of negative derivative values Number of local minima Mean rising time 10 spectral power values in [0–2.4] Hz bands Median Interquartile range (17×1 = 17 features)
Respiration pattern 1 channel	Mean of the signal Standard deviation Mean derivative Band energy ratio (difference between the logarithm of energy between the lower (0.05-0.25Hz) and the higher (0.25-5Hz) bands) Breathing rhythm (spectral centroid) Breathing rate 10 spectral power values in [0–2.4] Hz bands Median peak to peak time Median of the signal Interquartile range (19×1 = 19 features)
Heart rate 1 channel	Mean Heart Rate (HR) Mean HR variability (measured as the std of the peak-to-peak intervals) Energy ratio between frequency bands [0.04–0.15] Hz and [0.15–0.5] Hz Spectral power in the bands [0.1–0.2] Hz, [0.2–0.3] Hz, [0.3–0.4] Hz (6×1 = 6 features)
Skin temperature 1 channel	Mean of the signal Mean derivative Spectral power in bands [0–0.1] Hz and [0.1–0.2] Hz Median Interquartile range (6×1 = 6 features)

V. EXPERIMENT

Reliable, real-time, unobtrusive emotion recognition is important for affective human-computer interface. The purpose of the experiment is to investigate whether reasonable classification accuracy can be achieved with a subset of physiological modalities.

We seek to answer the following questions:

[4] The code for the feature extraction was written in MATLAB and is available by request from the corresponding author.



(a) The 32 users (b) Valence (low-grey/high-black) (c) Arousal (low-grey/high-black)

Figure 3. Three labellings of the data set plotted in the space of the first two principal components.

1) Which are the most discriminatory features? How stable are the feature rankings across users?
2) What is the classification accuracy and the F1 measure with the top ranked features?
3) Can we achieve reasonable classification accuracy with a subset of modalities?

A. *User-based feature selection*

To answer question 1, feature selection was carried out with RFE-SVM. We used the Weka implementation with the default settings [17]. In the first experiment, the classes were Low Valence and High Valence. The threshold on the valence scale was set at 5 as in [21]. In the second experiment, the class labels were Low Arousal and High Arousal, also cut off at threshold 5 on the arousal scale. Thirty two separate feature selection runs were done, one for each user in the database, generating 32 rankings of the 261 features. Thus each feature received 32 rank values, and the total feature rank was calculated as the average thereof. If the rankings were identical, there would be a feature ranked 1 in all runs, a feature ranked 2 in all runs, etc. Therefore we are interested to discover a small coherent group of features with low ranks. Unfortunately, no such group could be identified in either experiment. The average feature ranks for the valence and the arousal experiments are plotted in Figure 4. The mean rank is shown with a dashed line.

The stability of the rankings was evaluated for all subsets of cardinality $k = 1, \ldots, T$ using (4). The curves for the valence and the arousal experiments are shown in Figure 5. The stability index peaks around 50-70 features but the value of the index is less than 0.05 which suggests that the rankings are close to independent. Judging by this result, and by the lack of agreed low ranked features, we refrain from proposing a feature subset to be used across different users. The top features for each modality for the two classification tasks are shown in Table II. The following observations can be made.

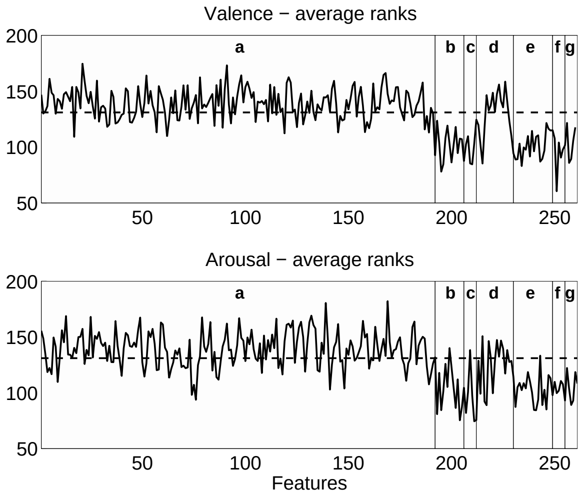


Figure 4. Average feature ranks across the 32 users. Modalities: a-EEG, b-EOG, c-EMG, d-GSR, e-Respiratory, f-Heart rate,g-Skin temperature

- The EEG modality is ranked last in both experiments. One reason for this is that the EEG pattern of emotion is overly user-specific, and does not generalise as easily as the other modalities across different users.
- There is no dominant modality and there is no redundant modality. Top features from all modalities, apart from EEG, are ranked high in the two overall lists, reinforcing the view that the emotion presentation is complex and multifaceted.

B. *Classification accuracy*

Following the protocol in [21], to answer question 2 we ran separate classification experiments with each user. The linear SVM classifier was applied in a leave-one-out protocol. For this experiment we used MATLAB and the SVM classifier from the Bioinformatics Toolbox. Table III gives the classification accuracies and the F1 measure for

Table II
THE TOP RANKED FEATURES FOR ALL MODALITIES

Valence			Arousal		
Feature name	Rank	Position	Feature name	Rank	Position
Trapezius EMG - median	74.4	(1)	Heart rare - band energy ratio	60.4	(1)
Horizontal EOG - mean	80.8	(4)	Horizontal EOG - median	78.0	(2)
Respiration - spectral power	84.3	(6)	Respiration - spectral centroid	83.0	(3)
GRS - number of local minima	88.8	(13)	Trapezius EMG - median	84.5	(4)
Skin temperature - spectral power	89.0	(15)	GSR - proportion of negative derivative	85.3	(6)
Heart rare - spectral power	92.9	(18)	Skin temperature - spectral power	89.3	(14)
EEG Channel PO3 - beta1	93.7	(19)	EEG Channel F3 - gamma	109.1	(40)

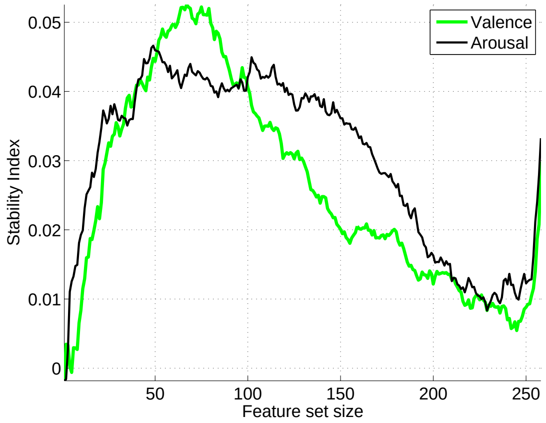


Figure 5. Stability index of the feature rankings of the 32 users.

the following experiments:[5]

• **All features**. This was done as a baseline experiment.

• **Selection 1**. We applied the selection criterion, threshold and procedure from the Koelstra et al.'s study (equation (1)), individually for each user. For each user, features were selected on the training fold of the data (39 video clips), and SVM was trained, and the testing was done on the left-out clip.

• **Selection 2**. Within the same leave-one-out protocol, an SVM classifier was trained and the weights were used to rank the features on the training fold of 39 clips. The top T features were used to train a second SVM classifier which was subsequently tested on the testing fold (the left-out clip). T was taken to be the number of features derived by Selection 1.

• **Selection 3**. The same as Selection 2 but the number of features was set to $T = 10$.

• **Selection 4**. We attempted a fourth selection idea whereby the class label of the clip is returned only if the labels of Selections 2 and 4 agree. The returned labels were 68% for Valence and 71% for Arousal. The results shown below are only for the data points with returned labels and are marked with '*' in Table III.

Table III also contains the standard deviations and the p-values of a t-test using the Statistics Toolbox of MATLAB. For the F1 measure, we compared the mean across the users to 0.5, corresponding to the chance value for this measure. For the the classification accuracies, we compared the means across the users with the accuracy of the "largest prior" classifier using a paired t-test. This is the classifier that always predicts the most prevalent class from the training data.

The results indicate that the F1 measure is provably better than chance for the Valence experiment for all strategies. For the Arousal experiment, the results were not so clear-cut. Strategies 3, 4 and 5 did not achieve statistically significant clearance above chance. Reducing the number of features leads to deterioration of the classifier. The classification accuracy for the Valence experiment exceeds the largest prior but only strategies 3 and 4 were found to be significantly better. On the other hand, the classification accuracy in the Arousal experiment is lower than that of the largest prior classifier. Only selection 4 achieves higher accuracy but the paired t-test did not find the difference significant. The overall conclusion from this part is that Valence can be classified into Low/High categories with a small subset of features while Arousal classification is not as straightforward. Taking the conclusion of the previous subsection, the subsets of features selected through the SVM are quite user-specific.

C. All combinations of modalities

We run an experiment with all $2^7 - 1 = 127$ possible combinations of modalities. All features from the respective modalities were used. Feature selection was not attempted because some of the modalities have too few features anyway. The 32 data sets, one per user, were used again with the leave-one-out protocol for each data set. An SVM

[5] To avoid confusion, here we explain how the F1 measure was calculated. The two classes in each experiment were taken in turn to be "the class of interest". F1 was calculated as

$$\text{F1} = 2\times\text{recall}\times\text{precision} / (\text{recall} + \text{precision})$$

for each class of interest. The two values were averaged to give the final F1 value. If all objects were labelled by the classifier in only one of the classes, F1 was taken to be undefined, and was assigned Not-A-Number constant in MATLAB. By averaging the two F1 measures instead of weighting them by the class prevalences we eliminate the influence of prior probabilities.

Table III
CLASSIFICATION ACCURACY AND F1 MEASURE FOR SELECTED FEATURES (● DENOTES STATISTICAL SIGNIFICANCE AT LEVEL 0.05)

Method	Mean	Std	p-value	
		F1 - Valence		
All features	62.49	7.29	7e-011	●
Selection 1	62.20	9.65	5e-008	●
Selection 2	58.50	9.99	4e-005	●
Selection 3	59.99	9.66	2e-006	●
Selection 4	63.25	13.52	5e-006	●
		F1 - Arousal		
All features	63.99	11.17	6e-008	●
Selection 1	60.17	12.10	4e-005	●
Selection 2	53.17	10.92	0.1101	
Selection 3	52.45	10.35	0.2055	
Selection 4	52.81	12.83	0.2485	
		Accuracy - Valence		
Largest prior classifier	58.60	7.30	–	
All features	62.42	7.31	0.0552	
Selection 1	62.19	9.63	0.1146	
Selection 2	61.56	8.35	0.1484	
Selection 3	63.98	8.54	0.0040	●
Selection 4*	68.71	10.20	0.0000	●
		Accuracy - Arousal		
Largest prior classifier	64.40	10.40	–	
All features	64.14	11.35	0.8870	
Selection 1	60.70	12.43	0.0342	
Selection 2	61.56	10.53	0.1737	
Selection 3	64.14	9.97	0.8801	
Selection 4*	66.36	12.91	0.3045	

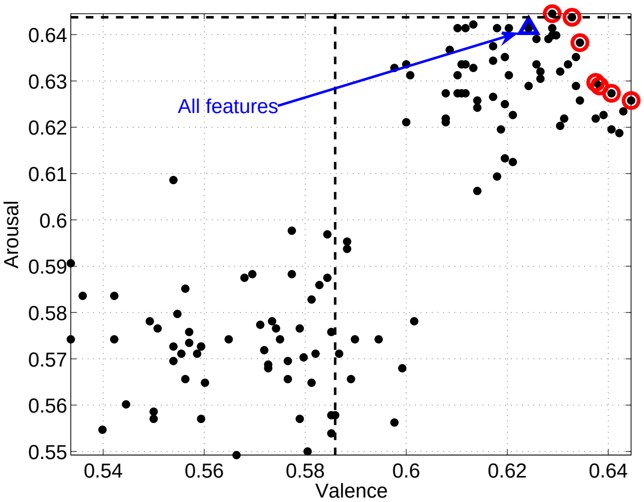


Figure 6. Scatterplot of the valence and arousal classification accuracies for all 127 combinations of the 7 modalities

Table IV
THE TOP RANKED FEATURES FOR ALL MODALITIES

Modalities					Valence	Arousal
EEG	EOG	RESP			64.45	62.58
EEG	EOG	EMG	RESP		63.44	63.83
EEG	EOG	GSR	RESP		63.83	62.89
EEG	EOG	GSR	Skin t^o		64.06	62.73
EEG	EOG	EMG	RESP	HR	62.89	64.45
EEG	EOG	EMG	RESP	Skin t^o	63.28	64.38
EEG	EOG	GSR	RESP	Skin t^o	63.75	62.97

classifier was built and evaluated on each data set for the Valence and Arousal experiments. Figure 6 shows the 127 combinations as points in the 2-d space of the valence and arousal accuracies. The point corresponding to using all modalities is marked with a blue triangle. In addition, we identified and marked with red circles the Pareto-optimal combinations of modalities, also detailed in Table IV. These are non-dominated combinations, which means that for any such combination, there is no other combination that has higher accuracy in both Valence and Arousal experiments.

The class prevalences are marked with dashed lines on the figure. It can be seen that many combinations exceed the prior probability for the Valence experiment and barely any combinations score better than the prior probability for the Arousal experiment, reinforcing the findings in the previous subsection.

Figure 6 reveals an interesting 2-cluster structure of the data. To examine this further, we plotted in Figure 7 the same layout in grey colour and overlaid the points for the individual modalities in black. The black dots indicate that the respective modality is a part of the combination. The Figure shows that the better cluster (higher accuracies on both Valence and Arousal) is entirely due to the EEG modality. This suggests that although individual features from this modality did not appear high in the overall ranking (Table II), important discriminatory information is contained in the relationship between the EEG channels. Black points in the lower cluster on the right side of the dashed line show that Low/High valence can be recognised with accuracy above chance by modalities that do not include EEG. This however requires further data collection in tailor-made settings.

VI. DISCUSSION AND CONCLUSIONS

The results in our experiment with the DEAP data base are in unison with Koelrstra's results. Note that we did not use the features extracted from the video clips because here we are only interested in modalities measured from the human player, assuming that the emotional charge of the stimulus is not known. We attempted a further feature selection and modality selection but did not succeed in identifying a small with high discrimination potential. All combinations of modalities with high ranks included the EEG features, which would require that the user wears a cap or at least a headset with a sufficient number of electrodes on it. Classification accuracy is likely to increase with the development of more reliable and robust data collection devices as well as with

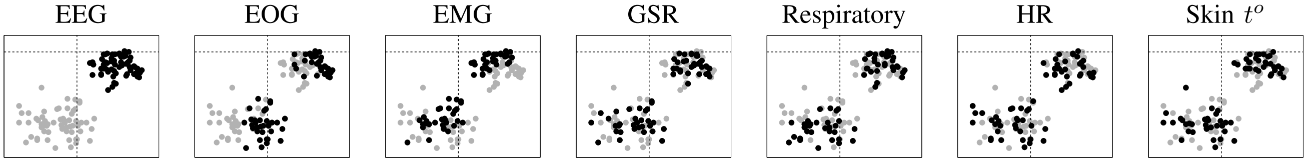


Figure 7. Modalities

the development of adroit techniques for feature extraction from these data.

It can be argued that many HCI applications will not require exceptionally high accuracy of emotion recognition. For example, when playing a video game, the user may perceive a game's reaction to their *arousal* level as an adequate response to emotion. It will be of less importance whether the arousal was the consequence of jubilation or frustration. If needed, the context of the game can be used to further gauge the valence of the emotion. Thus modern and ubiquitous HCI may benefit from focusing on a cruder but fast, reliable and robust classification of arousal only.

Finally, we note that the analyses were done with the preprocessed DEAP data, and using the whole signal duration for feature extraction. In reality, to build a responsive affective HCI, emotion must be detected from a short-time interval of data, simulating real-time recognition. It will be interesting to continue this research with analyses of the same modalities in the real-time scenario.

References

[1] C. Abras, D. Maloney-Krichmar, and J. Preece. User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications*, 37(4):445–56, 2004.

[2] W. Altidor, T. M. Khoshgoftaar, and A. Napolitano. A noise-based stability evaluation of threshold-based feature selection techniques. In *IEEE International Conference on Information Reuse and Integration (IRI)*, pages 240–245, 2011.

[3] M. Ambinder. Valve's approach to playtesting: The application of empiricism. Valve Software, 2009.

[4] F.A. Boiten. The effects of emotional behaviour on components of the respiratory cycle. *Biological Psychology*, 49(1-2):29–51, 1998.

[5] A. Bonarini, F. Costa, M. Garbarino, M. Matteucci, M. Romero, and S. Tognetti. Affective videogames: The problem of wearability and comfort. 6764:649–658, 2011.

[6] D.O. Bos. EEG-based Emotion Recognition: the influence of visual and auditory stimuli. *Retrieved from http://hmi. ewi. utwente. nl/verslagen/capitaselecta/CS-Oude/Bos-Danny. pdf*, 2006.

[7] R.A. Calvo and S. DMello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affecitve Computing*, 1(1):18–37, 2010.

[8] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun. Emotion assessment: Arousal evaluation using EEGs and peripheral physiological signals. *Multimedia Content Representation, Classification and Security*, 4105/2006:530–537, 2006.

[9] Charles Darwin. *The Expression of the Emotions in Man and Animals*. John Murray, 1872.

[10] M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis*, 1:131–156, 1997.

[11] F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage*, 43(1):44–58, 2008.

[12] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley & Sons, NY, second edition, 2001.

[13] S.H. Fairclough. Fundamentals of physiological computing. *Interacting with Computers*, 21(1-2):133–145, 2009.

[14] M. Gerven, J. Farquhar, R. Schaefer, R. Vlek, J. Geuze, A. Nijholt, N. Ramsey, P. Haselager, L. Vuurpijl, S. Gielen, and P. Desain. The brain–computer interface cycle. *Journal of Neural Engineering*, 6:041001, 2009.

[15] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[16] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Bioinformatics*, 22(19):2348–2355, 2006.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations*, 11, 2009.

[18] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

[19] Y. Homma and I. Masaoka. Breathing rhythms and emotions. *Experimental physiology*, 93:1011–1021, 2008.

[20] K. Ko, H. Yang, and K. Sim. Emotion recognition using EEG signals with relative power values and Bayesian network. *International Journal of Control, Automation, and Systems*, 7:865–870, 2009.

[21] 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*, pages 1–15, 2011.

[22] L.I. Kuncheva. A stability index for feature selection. In *Proc. IASTED, Artificial Intelligence and Applications*, pages 390–395, Innsbruck, Austria, 2007.

[23] W. Liao, W. Zhang, Z. Zhu, Q. Ji, and W. Gray. *Toward a decision-theoretic framework for affect recognition and user assistance*, volume 64. September 2006.

[24] H.G. Mittelmann and B. Wolff. Effective states and skin temperature: experimental study of subjects with "cold hands" and Raynaud's syndrome. *Psychosomatic Medicine*, 1:271–292, 1939.

[25] H.G. Mittelmann and B. Wolff. Emotions and skin temperature; observations of patients during psychotherapeutic (psychoanalytic) interviews. *Psychosomatic Medicine*, 5(3):211231, 1943.

[26] L.C. Molina, L. Belanche, and A. Nebot. Feature selection algorithms: A survey and experimental evaluation. In *Proc. the IEEE International Conference on Data Mining (ICDM'02)*, Japan, 2002.

[27] K.A. Norman, A.M. Polyn, G.J. Detre, and J.V. Haxby. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10:424–430, 2006.

[28] P. Petrantonakis and L. Hadjileontiadis. Emotion recognition from EEG using higher-order crossings. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):186–197, 2010.

[29] R.W. Picard. Affective computing: From laughter to IEEE. *IEEE Transactions on Affecitve Computing*, 1(1):11–17, 2010.

[30] R.W. Picard. *Affective computing*. The MIT press, 2000.

[31] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507, 2007.

[32] J. Sherwood and R. Derakhshani. On classifiability of wavelet features for EEG-based brain-computer interfaces. *2009 International Joint Conference on Neural Networks*, pages 2895–2902, June 2009.

[33] UK Game Shows.com. The chair. Internet, August 2002.

[34] K. Takahashi. Remarks on emotion recognition from biopotential signals. In *The Second International Conference on Autonomous Robots and Agents*, pages 186–191. Citeseer, 2004.

[35] M. Bonarini, A. Matteucci, M. Tognetti and S. Garbarino. Modeling enjoyment preference from physiological responses in a car racing game. *Computational Intelligence and Games (CIG)*, pages 321–328, 2010.

[36] W. Masuda, T. Yuki and M. Maddux. Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in japan and the united states. *Journal of Experimental Social Psychology*, 43:303–311, 2007.