Detection of affective patterns in physiological signals towards improving automatic emotion recognition

Jonghwa Kim, Elisabeth André

Angaben zur Veröffentlichung / Publication details:

Kim, Jonghwa, and Elisabeth André. 2009. "Detection of affective patterns in physiological signals towards improving automatic emotion recognition." In *Handbook of Pattern Recognition and Computer Vision*, edited by C. H. Chen, 4th ed., 415–34. Singapore: World Scientific. https://doi.org/10.1142/9789814273398_0018.

CHAPTER 3.6

DETECTION OF AFFECTIVE PATTERNS IN PHYSIOLOGICAL SIGNALS TOWARDS IMPROVING AUTOMATIC EMOTION RECOGNITION

Jonghwa Kim and Elisabeth André

Institute of Computer Science,
University of Augsburg,
Eichleitnerstr. 30, 86159 Augsburg, Germany
{*kim, andre*}*@informatik.uni-augsburg.de*

In this chapter, we investigate the potential of physiological signals as reliable channels for emotion recognition. All essential stages of an automatic recognition system are discussed, from the recording of a physiological dataset to a feature-based multiclass classification. In order to collect a physiological dataset from multiple subjects, we developed a musical induction method, without any deliberate lab setting. Four-channel biosensors were used to measure electromyogram, electrocardiogram, skin conductivity, and respiration changes. A wide range of physiological features from various analysis domains is proposed to find the best emotion-relevant features and correlate them with emotional states. The best features extracted are specified in detail and their effectiveness is proven by classification results. Classification of four musical emotions (positive/high arousal, negative/high arousal, negative/low arousal, positive/low arousal) is performed by using an extended linear discriminant analysis (pLDA). Furthermore, by exploiting a dichotomic property of the 2D emotion model, we develop a novel scheme of emotion-specific multilevel dichotomous classification (EMDC) and compare its performance with direct multiclass classification using the pLDA.

1. Introduction

In advanced human-computer interaction (HCI) today, resolving absence of mutual sympathy (rapport) in interaction between human and machine is one of the most important issues. With exponentially evolving technology, it is no exaggeration to say that any interface that disregards human affective states in the interaction - and thus fails to pertinently react to the states - will never be able to inspire confidence. Instead, users will perceive it as cold, untrustworthy, and socially inept. In human communication, expression and understanding of emotions helps achieve mutual sympathy. To approach this in human-computer interaction, we need to equip machines with the means to interpret and understand human emotions without the input of a user's translated intention. Hence, one of the most important prerequisites for realizing such an advanced user interface is a reliable emotion recognition system

which guarantees acceptable recognition accuracy, robustness against any artifacts, and adaptability to practical applications. Developing such a system requires the following stages: to model, analyze, process, train, and classify emotional features measured from the implicit emotion channels of human communication, such as speech, facial expression, gesture, pose, physiological responses, etc. In this chapter we concentrate on identifying emotional cues in various physiological measures.

Recently, numerous studies on engineering approaches to automatic emotion recognition have been published, though research in that field is relatively new compared to the long history of emotion research in psychology and psychophysiology. In particular, many efforts have been deployed to recognize human emotions using audiovisual channels of emotion expression, i.e. facial expressions, speech, and gestures. Little attention, however, has been paid so far to using physiological measures, as opposed to audiovisual emotion channels [1]. This is due to some significant limitations that come with the use of physiological signals for emotion recognition. The main difficulty lies in the fact that it is a very hard task to uniquely map physiological patterns onto specific emotional states. As an emotion is a function of time, context, space, culture, and person, physiological patterns may widely differ from user to user and from situation to situation. Above all, humans use non-discrete labels to describe emotions.

In the next section, we give a brief overview of related research on automatic emotion recognition using physiological signals. Section 3 gives the motivation and rationale for our experimental setting of musical emotion induction and is followed by a detailed explanation of all the biosensors we used. A systematic description of signal analysis methods and classification procedure using extended linear discriminant analysis is given in Section 4. In Section 5, we present the best emotion-relevant ANS features with the recognition results we achieved. In addition, the performance of the novel EMDC scheme is tested and its potential is proven by improved recognition accuracy. In Section 6, we discuss the problems faced during our work including the difficulty in subject-independent recognition. We then conclude with perspectives related to future work.

2. Related Research

A significant amount of work has been conducted by Picard and colleagues at MIT Lab showing that certain affective states may be recognized by using physiological data including heart rate, skin conductivity, temperature, muscle activity and respiration velocity [2]. They used personalized imagery to elicit target emotions from a single subject who had two years' experience in acting, and they achieved an overall recognition accuracy of 81% for eight emotions by using hybrid linear discriminant classification. Nasoz et al. [3] used movie clips based on the study by Gross and Levenson [4] for eliciting target emotions from 29 subjects and achieved an emotion classification accuracy of 83% using the Marquardt Backpropagation al-

gorithm (MBP). In work [5], the IAPS photoset [6] is used to elicit target emotions with positive and negative valence and variable arousal level from a single subject. The arousal and valence dimensions of the emotions were classified separately using a neural network classifier and recognition accuracy rates of 96.6% and 89.9% respectively were achieved.

More recently, an interesting user-independent emotion recognition system was reported by Kim et al. [7]. They developed a set of recording protocols using multimodal stimuli (audio, visual, and cognitive) to evoke targeted emotions (sadness, stress, anger, and surprise) from 175 children aged five to eight. A classification ratio of 78.43% was achieved for three emotions (sadness, stress, and anger) and a ratio of 61.76% for four emotions (sadness, stress, anger, and surprise) by adopting support vector machines as pattern classifier. Most interestingly, analysis steps in the system were fitted to handle relatively short lengths of the input signals (segmented in 50 seconds) compared to previous works that required longer signal lengths of about 2-6 min.

The aforementioned approaches achieved average accuracy rates of over 80% which seem to be acceptable for practical applications. It is true, however, that recognition rates are strongly dependent on the datasets that are used and on the application context. Moreover, the physiological datasets used in most of these works were gathered by using visual elicitation materials in a lab setting. The subjects then "tried and felt" or "acted out" the target emotions while looking at selected photos or watching movie clips that were carefully prearranged to elicit the emotions. In other words, to put it bluntly, the recognition results were achieved for specific users in specific contexts with "forced" emotional states.

Most of the previous works provide evidence of the fact that the accuracy of arousal discrimination is always higher than that of valence differentiation. The reason might be the that the change of the arousal level corresponds directly to the intensity of discharge in ANS activities, such as sweat glands and blood pressure, which is straightforward to measure, while valence differentiation of emotion requires a multifactor analysis of cross-correlated ANS reactions. This finding led us to develop an emotion-specific classification scheme and to calculate a wide range of features in various analysis domains in order to extract valence-relevant features from ECG and RSP signals.

3. Setting of Experiment

3.1. *Musical emotion induction*

To collect a database of physiological signals in which the targeted emotions corresponding to the four quadrants in the 2D emotion model (i.e. EQ1, EQ2, EQ3, and EQ4 in Fig. 1) can be *naturally* reflected without any deliberate expression, we decided to use the musical induction method, i.e. to record physiological signals while the subjects were listening to different pieces of music.

A well established mechanism of emotion induction consists in triggering emotions by resorting to imagination or individual memories. Emotional reaction can be triggered by a specific cue and be evoked by an experimental instruction to imagine certain events. On the other hand, it can spontaneously be resurged in memory. Music is a pervasive element accompanying many highly significant events in human social life and particular pieces of music are often connected to significant personal memories. Following this, music can be a powerful cue in awakening emotional experiences and bringing back memories. Since listening to music is often done by an individual in isolation, the possible artifacts of social masking and social interaction can be minimized in the experiment. Furthermore, like odors, music can be treated at lower levels of the brain that are particularly resistant to modifications by later input, contrary to cortically based episodic memory. This is even the case when the listening occurs at the same time as other activities within a social setting since musical emotion cannot co-occur with social interaction in general.

The subjects were three males (one of the co-authors and two student researchers recruited from the authors' lab) aged 25-38 and who all enjoy listening to music in their everyday life. The subjects were not paid, but allowed to perform the experiments during their regular working hours. They individually handpicked four songs that were intended to spontaneously evoke emotional memories and certain moods corresponding to the four target emotions. Figure 1[a] shows the musical emotion model referred to for the selection of their songs. Generally, emotional responses to music varies greatly from individual to individual depending on their unique past experiences. Moreover, cross-cultural comparisons in literature suggest that emotional responses can be quite differentially emphasized by different musical cultures and training. This is why we advised the subjects to choose themselves the songs they believed would help them recall their individual special memories with respect to the target emotions.

For the experiment, we prepared a quiet listening room in our institute in order to ensure that the subjects could experience the emotions evoked by the music undisturbed. For the recording, the subject had to position the sensors following the instructions posted in the room, to put on the headphones, and select a song from his song list saved in the computer. When clicking on the selected song, the recording and music systems were automatically set up by preset values for each song, such as volume, treble, and bass. Most importantly, before the start of the experiment, the subjects were shown how to prepare the skin by using an antiseptic spray and a skin preparation gel for reducing electrode-impedance, and how to correctly position the sensors. Recording schedules were decided by the subjects themselves and the recordings took place whenever they felt like listening to music. They were also free to choose the songs they wanted to listen to. Thus, in contrast

[a]Metaphoric cues for song selection: song1 (positively exciting, energizing, joyful, exuberant), song2 (noisy, loud, irritating, discord), song3 (melancholic, sad memory), song4 (blissful, pleasurable, slumberous, tender)

to methods used in other studies, the subjects were not forced to participate in a lab setting scenario and to use prespecified stimulation material. We believe that this voluntary participation of the subjects during our experiment might help obtain a high-quality dataset with natural emotions.

During the three months, a total of 360 samples (90 samples for each emotion) from three subjects were collected. The signal length of each sample was between 3-5 minutes depending on the duration of the songs.

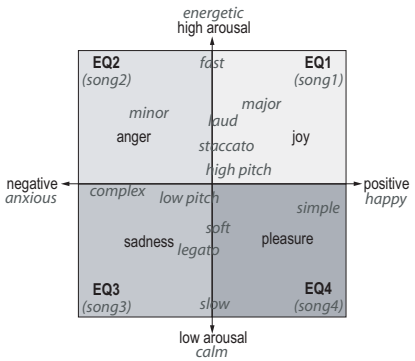

Fig. 1.: Reference emotional cues in music based on the 2D emotion model. EQ1 = positive/high arousal, EQ2 = negative/high arousal, EQ3 = negative/low arousal, EQ4 = positive/low arousal

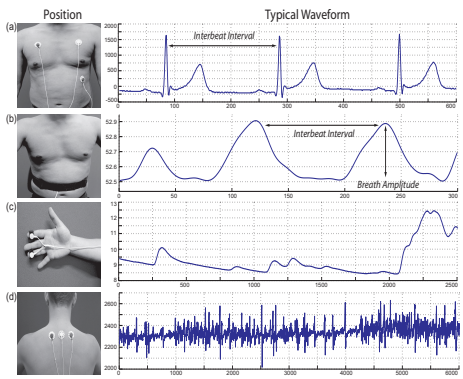

Fig. 2.: Position and typical waveforms of the biosensors: (a) ECG, (b) RSP, (c) SC, (d) EMG.

3.2. *Biosensors*

The physiological signals were acquired using the Procomp[b] Infiniti™ with four biosensors, electromyogram (EMG), skin conductivity (SC), electrocardiogram (ECG), and respiration (RSP). The sampling rates were 32 Hz for EMG, SC, and RSP, and 256 Hz for ECG. The positions and typical waveforms of the biosensors we used are illustrated in Fig. 2.

4. Methodology

The overall structure of our recognition system is illustrated in Figure 3.

After the preprocessing stage for signal segmentation and denoising we calculated 110 features from the 4-channel biosignals and selected the most significant features by using the sequential backward search method. For classification, various machine learning methods (supervised classification in our case) can be used [8].

[b]This is an 8 channel multi-modal Biofeedback system with 14 bit resolution and a fiber optic cable connection to the computer. www.MindMedia.nl

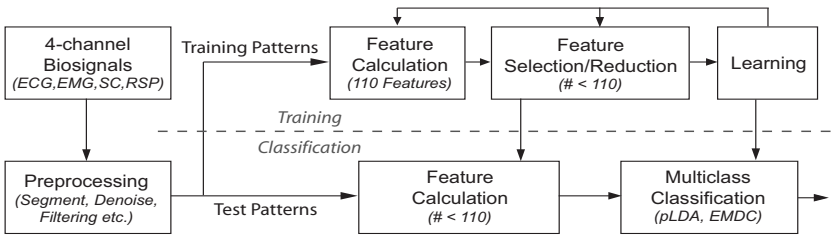

Fig. 3.: Block diagram of supervised statistical classification system for emotion recognition

After having tested some classifiers, such as k-nearest neighbor (k-NN), multilayer perceptron (MLP), and linear discriminant analysis (LDA), we chose the LDA which outperformed with higher recognition accuracy in our case. It should, however, be noted that there is no single best classification algorithm and the choice of the best classification method strongly depends on the characteristics of the dataset to be classified.

4.1. *Preprocessing*

Different types of artifacts were observed in all the four channel signals, such as transient noise due to movement of the subjects during the recording, mostly at the beginning and at the end of each recording. Thus, uniformly for all subjects and channels, we segmented the signals into final samples of a 160 seconds each, obtained by taking the middle part of each signal. It is important to note that the EMG signal generally requires additional pre-processing such as deep smoothing or signal separation, depending on the position of the sensor, because the nature of the signal is such that all the muscle fibers within the recording area of the sensor contract at different rates. In our case, the EMG signal contains artifacts generated by heart beat and respiration, since we positioned the sensor at the upper trapezius muscle. Using an adaptive bandpass filter we removed the artifacts (Fig. 4). For other signals we used pertinent lowpass filters to remove noises without loss of information.

4.2. *Measured features*

From the four channel signals we calculated a total of 110 features from various analysis domains including conventional statistics in time series, frequency domain, geometric analysis, multiscale sample entropy, subband spectra, etc. For the signals with non-periodic characteristics, such as EMG and SC, we focused on capturing the amplitude variance and localizing the occurrences (number of transient changes) in the signals. In the following sections, we describe the feature calculation methods in detail.

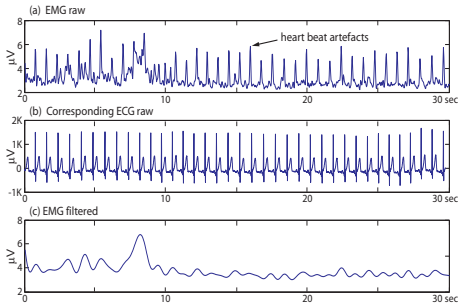

Fig. 4.: Example of EMG signal with heart beat artifacts and denoised signal

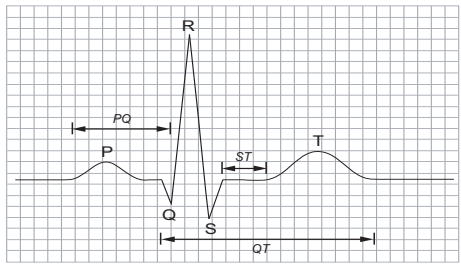

Fig. 5.: QRS waveform in ECG signal. Usual lengths: P-wave (0.08-0.10s), QRS (0.06-0.10s), PR-interval (0.12-0.20s), QT_c-interval ($QT/\sqrt{RR} \leq 0.44$s) [9]

4.2.1. *Electrocardiogram (ECG)*

ECG measures depolarized electrical changes of muscular contraction associated with cardiovascular activity. In general, the ECG is measured at the body surface along the axis of the heart and results from the activation first of the two small heart chambers, the atria, and then of the two larger heart chambers, the ventricles. The contraction of the ventricles produces the specific waveform known as the QRS complex (see Fig. 5).

To obtain the subband spectrum of the ECG signal we used the typical 1024 points fast Fourier transform (FFT) and partitioned the coefficients within the frequency range 0-10 Hz into eight non-overlapping subbands with equal bandwidth. First, as features, power mean values of each subband and fundamental frequency (F0) are calculated by finding maximum magnitude in the spectrum within the range 0-3 Hz. To capture peaks and their locations in subbands, subband spectral entropy (SSE) is computed for each subband. Entropy plays an important role in information theory as a measure of disorganization or uncertainty in a random variable. In pattern recognition it is generally used to measure the degree of a classifier's confidence. To compute the SSE, it is necessary to convert each spectrum into a probability mass function (PMF) like form. Eq. 1 is used for the normalization of the spectrum.

$$x_i = \frac{X_i}{\sum_{i=1}^{N} X_i}, \qquad \text{for } i = 1 \ldots N \tag{1}$$

where X_i is the energy of i^th frequency component of the spectrum and $\tilde{\mathbf{x}} = \{x_1 \ldots x_N\}$ is to be considered as the PMF of the spectrum. In each subband the SSE is computed from $\tilde{\mathbf{x}}$ by

$$H_{\text{sub}} = -\sum_{i=1}^{N} x_i \cdot \log_2 x_i \tag{2}$$

By packing the eight subbands into two bands, i.e., subbands 1-3 as the low frequency (LF) band and subbands 4-8 as the high frequency (HF) band, the ratios of the LF/HF bands are calculated from the power mean values and the SSEs.

In biomedical engineering, the analysis of the local morphology of the QRS waveform and its time varying properties has been a standard method for assessing cardiac health. Importantly, heart rate variability (HRV) is one of the most often used measures for ECG analysis. To obtain the HRV from the continuous ECG signal, each QRS complex is detected and the RR intervals (all intervals between adjacent R waves) or the normal-to-normal (NN) intervals (all intervals between adjacent QRS complexes resulting from sinus node depolarization) are determined. We used the QRS detection algorithm of Pan and Tompkins [10] in order to obtain the HRV time series. Figure 6 shows examples of R wave detection and interpolated HRV time series, referring to the increases and decreases over time in the NN intervals.

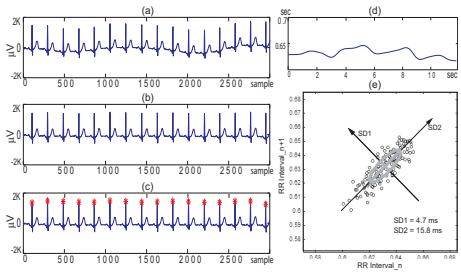

Fig. 6.: Example of ECG Analysis: (a) raw ECG signal with respiration artifacts, (b) detrended signal, (c) detected RR interbeats, (d) interpolated HRV time series, (e) Poincaré plot of the HRV time series.

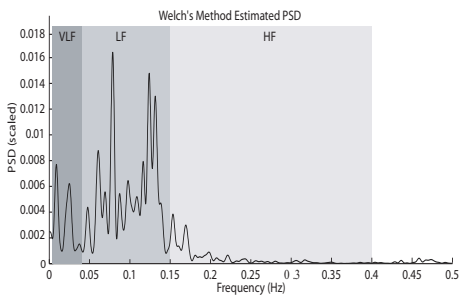

Fig. 7.: Example of heart rate spectrum in three subbands using the 1024-points fast Fourier transform.

In the time-domain of the HRV, we calculated statistical features including mean value, standard deviation of all NN intervals (SDNN), standard deviation of the first difference of the HRV, the number of pairs of successive NN intervals differing by more than 50 ms (NN50), the proportion derived by dividing NN50 by the total number of NN intervals. By calculating the standard deviations in different distances of RR interbeats, we also added Poincaré geometry in the feature set to capture the nature of interbeat (RR) interval fluctuations. Poincaré plot geometry is a graph of each RR interval plotted against the next interval and provides quantitative information of the heart activity by calculating the standard deviations of the distances of $R-R(i)$ to lines $y=x$ and $y=-x+2*R-R_m$, where $R-R_m$ is the mean of all $R-R(i)$. Figure 6 (e) shows an example plot of the Poincaré geometry. The standard deviations SD_1 and SD_2 refer to the fast beat-to-beat variability and longer-term variability of $R-R(i)$ respectively.

Entropy-based features from the HRV time series were also considered. Based on the so-called *approximate entropy* and *sample entropy*, a multiscale sample entropy (MSE) was introduced [11] and successfully applied to physiological data, especially for analysis of short and noisy biosignal. Given a time series $\{X_i\} = \{x_1, x_2, ..., x_N\}$ of length N, the number $(n_i^{(m)})$ of similar m-dimensional vectors $y^{(m)}(j)$ for each sequence vectors $y^{(m)}(i) = \{x_i, x_{i+1}, ..., x_{i+m-1}\}$ is determined by measuring their respective distances. The relative frequency to find the vector $y^{(m)}(j)$ within a tolerance level δ is defined by

$$C_i^{(m)}(\delta) = \frac{n_i^{(m)}}{N - m + 1} \tag{3}$$

The approximate entropy, $h_A(\delta, m)$, and the sample entropy, $h_S(\delta, m)$ are defined as

$$h_A(\delta, m) = \lim_{N\to\infty} [H_N^{[m]}(\delta) - H_N^{(m+1)}(\delta)], \tag{4}$$

$$h_S(\delta, m) = \lim_{N\to\infty} -\ln \frac{C^{(m+1)}(\delta)}{C^{(m)}(\delta)}, \tag{5}$$

where

$$H_N^{(m)}(\delta) = \frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} \ln C_i^{(m)}(\delta), \tag{6}$$

Because it has the advantage of being less dependent on the time series length N, we applied the sample entropy h_S to coarse-grained versions $(y_j^{(\tau)})$ of the original HRV time series $\{X_i\}$,

$$y_j(\tau) = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i, \quad 1 \le j \le N/\tau, \quad \tau = 1, 2, 3, ... \tag{7}$$

The time series $\{X_i\}$ is first divided into N/τ segments by non-overlapped windowing with length of scale factor τ and then the mean value of each segment is calculated. Note that for scale one $y_j(1) = x_j$. From the scaled time series $y_j(\tau)$ we obtain the m-dimensional sequence vectors $y^{(m)}(i, \tau)$. Finally, we calculate the sample entropy h_S for each sequence vector $y_j(\tau)$. In our analysis we used $m = 2$ and fixed $\delta = 0.2\sigma$ for all scales, where σ is the standard deviation of the original time series x_i. Note that using the fixed tolerance level δ as a percentage of the standard deviation corresponds to initial normalizing of the time series and it thus ensures that h_S does not depend on the variance of the original time series, but only on their sequential ordering.

In the frequency-domain of the HRV time series, three frequency bands are of general interest: the very-low frequency (VLF) band (0.003-0.04 Hz), the low frequency (LF) band (0.04-0.15 Hz), and the high frequency (HF) band (0.15-0.4 Hz). From these subband spectra, we computed the dominant frequency and power

of each band by integrating the power spectral densities (PSD) obtained by using Welch's algorithm, as well as the ratio of power within the low-frequency band to that within the high-frequency band (LF/HF). Since parasympathetic activity dominates at high frequency, the LF/HF ratio is generally thought to distinguish sympathetic effects from parasympathetic effects. Figure 7 shows the heart rate spectrum from one of the subjects.

4.2.2. *Respiration (RSP)*

RSP signal (breathing rate and intensity) is commonly acquired by measuring physical change of the thoracic expansion with a rubber band around the chest or belly and contains less artifact in general than the other sensors using electrodes, e.g., ECG, EMG, SC etc. Including the typical statistics of the raw RSP signal, we calculated similar types of features, such as the ECG features, the power mean values of three subbands (obtained by dividing the Fourier coefficients within the range 0-0.8 Hz into non-overlapped three subbands with equal bandwidth), and the set of subband spectral entropies (SSE). In order to investigate inherent correlation between respiration rate and heart rate, we considered a novel feature content for the RSP signal. Since an RSP signal exhibits a quasi periodic waveform with sinusoidal properties, it does not seem unreasonable to conduct an HRV-like analysis for the RSP signal, i.e. to estimate breathing rate variability (BRV). After detrending using the mean value of the entire signal and lowpass filtering, we calculated the BRV by detecting the peaks in the signal using the maxima ranks within each zero-crossing (Fig. 8).

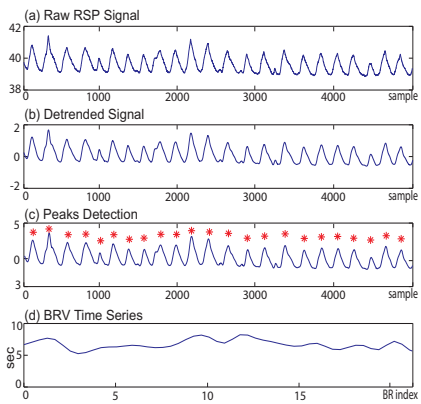

Fig. 8.: BRV analysis for RSP signal: (a) raw RSP signal with Fs = 32Hz, (b) lowpassed and detrended signal of (a), (c) peaks detection, (d) BRV time series.

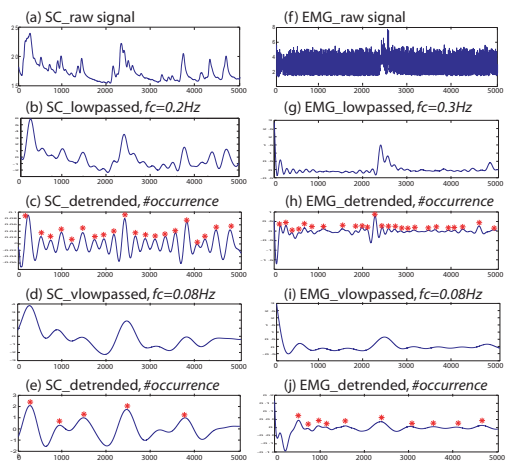

Fig. 9.: Analysis Examples of SC and EMG signals

From the BRV time series, we calculated the mean value, SD, SD of the first difference, MSE, Poincaré analysis, etc. In the spectrum of the BRV, peak frequency, power of the two subbands, the low-frequency band (0-0.03Hz) and the high-frequency band (0.03-0.15 Hz), and the ratio of the power within the two bands (LF/HF) were calculated.

4.2.3. *Skin Conductivity (SC)*

The SC signal includes two types of electrodermal activity, the DC level component and the skin conductance response (SCR). The DC level in the SC signal indicates a general activity of the perspiratory glands influenced by body temperature or external temperature. The SCR is the distinctive short waveform in the SC signal and is considered to be useful for emotion recognition as it is linearly correlated with the intensity of arousal responding to internal/external stimuli. The mean value, standard deviation, and mean of first and second derivations were extracted as features from the normalized SC signal and the low-passed SC signal using a cutoff frequency of 0.2 Hz. To obtain a detrended SCR waveform without DC-level components, we removed the continuous, piecewise linear trend in the two low-passed signals, i.e., the very low-passed (VLP) and the low-passed (LP) signal with a cutoff frequency of 0.08 Hz and 0.2 Hz, respectively (see Fig. 9 (a)-(e)).

The baseline of the SC signal was calculated and subtracted to consider only relative amplitudes. By finding two consecutive zero-crossings and the maximum value between them, we calculated the number of SCR occurrences within 100 seconds from each LP and VLP signal, the mean of the amplitudes of all occurrences, and the ratio of the SCR occurrences within the low-passed signals (VLP/LP).

4.2.4. *Electromyography (EMG)*

For the EMG signal, we calculated types of features similar to those of the SC signal. The mean value of the entire signal, the mean of the first and second derivations, and the standard deviation were extracted as features from the normalized and low-passed signals. The occurrence number of myo-responses and the ratio of that within VLP and LP signals were also added to the feature set and were determined in the same way as the SCR occurrence but using cutoff frequencies with 0.08 Hz (VLP) and 0.3 Hz (LP) (see Fig. 9 (f)-(j)).

In the end, we obtained a total of 110 features from the 4-channel biosignals; 53 (ECG) + 37 (RSP) + 10 (SC) + 10 (EMG).

4.3. *Classification*

4.3.1. *Feature selection*

A large number of algorithms for feature subset selection have been proposed in the literature [12]. Although sequential backward selection (SBS) is computationally

more demanding than sequential forward selection (SFS), we decided to use SBS in our recognition system because it outperformed SFS and other methods in the feature space. Nevertheless we note that the performance of all the selection methods proposed is strongly dependent on the given dataset.

We did not consider integrating a dimensionality reduction method in our recognition scheme, such as principle component analysis (PCA) and Fisher projection, which are commonly used in combination with a classifier. Dimensionality reduction amounts to projecting high-dimensional data to a lower dimensional space with a minimal loss of information. This means that new features are created by the transformation of original feature values, rather than by selecting a feature subset from a given feature set. Such feature reduction methods were not suitable for the purpose of our work since we seeked to determine the best emotion-relevant features which preserve their origins of analysis domain and value. We use Fisher projection exclusively to preview the distribution of the features.

4.3.2. *Classifying using extended linear discriminant analysis*

In discriminant analysis, for a given dataset, three scatter matrices, within-class (S_w), between-class (S_b), and mixture scatter matrices (S_m) are defined as follows;

$$S_b = \sum_{i=1}^{c} N_i(\mu_i - \overline{\mathbf{x}})(\mu_i - \overline{\mathbf{x}})^T = \mathbf{\Phi}_b \mathbf{\Phi}_b^T, \tag{8}$$

$$S_w = \sum_{i=1}^{c} \sum_{j \in C_i} (\mathbf{x}_j - \mu_i)(\mathbf{x}_j - \mu_i)^T = \mathbf{\Phi}_w \mathbf{\Phi}_w^T, \tag{9}$$

$$S_m = S_b + S_w = \sum_{i=1}^{N} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T = \mathbf{\Phi}_m \mathbf{\Phi}_m^T, \tag{10}$$

where N is the number of all samples, N_i is the number of samples in class $C_i (i = 1, 2, ..., c)$, μ_i is the mean of the samples in class C_i, and $\overline{\mathbf{x}}$ is the mean of all samples, i.e.,

$$\mu_i = \frac{1}{N_i} \sum_{i \in C_i} \mathbf{x}_i, \tag{11}$$

$$\overline{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{c} \mathbf{x}_i = \frac{1}{N} \sum_{i=1}^{c} N_i \mu_i. \tag{12}$$

Note that the mixture scatter matrix S_m is the covariance matrix of all samples regardless of their class assignments and all the scatter matrices are designed to be invariant under coordinate shifts. The idea in LDA is to find an optimal transformation W which satisfies

$$\mathcal{J}(W) = \underset{W}{\operatorname{argmax}} \frac{|W^T S_b W|}{|W^T S_w W|}, \tag{13}$$

such that the separation between classes is maximized while the variance within a class is minimized (Fisher's criterion). Finding the optimal W is equivalent to finding the generalized eigenvectors satisfying $S_b W = \lambda S_w W$, for $\lambda = 0$. Transformation W can be obtained by applying the eigenvalue decomposition to the matrix $S_w^{-1} S_b$ if S_w is nonsingular, or to the matrix $S_b^{-1} S_w$ if S_b is nonsingular, and taking the rows of the transformation matrix to be the eigenvectors corresponding to the $n-1$ largest eigenvalues. Applying the singular value decomposition (SVD) on the scatter matrices of the training set is a stable way to compute the eigenvalue decomposition [13]. Since there are at most $c-1$ nonzero generalized eigenvectors of the scatter matrix, the upper bound of the number of retained dimensions in classical LDA is $c-1$ and the dimensionality can be further reduced, for example, by incorporating in W only those eigenvectors corresponding to the largest singular values determined in the scatter SVD. Given the transformation W, classification can be performed in the transformed space based on some distance measures d, such as Euclidean distance. The new instance, $\mathbf{v}$, is classified to

$$\underset{k}{\operatorname{argmin}}\, d\ (\mathbf{v}W, \overline{\mathbf{x}}_k W) \tag{14}$$

where $\overline{\mathbf{x}}_k$ is the centroid of k-th class and $k = 1, 2, ..., c$.

Note that a limitation of conventional LDA is that its objective function requires that one of the scatter matrices be nonsingular. It means that for a given c-class, p-dimensional classification problem, at least $c + p$ samples are required to guarantee that the within-class scatter matrix S_w does not become singular. To deal with the singularity problem, several extended LDA methods are proposed such as PCA+LDA, pseudoinverse LDA, regularized LDA, and LDA using generalized singular value decomposition (GSVD). In our work we used the pseudoinverse LDA (pLDA), a natural extension of classical LDA, applying the eigenvalue decomposition to the matrix $S_b^+ S_w$, $S_w^+ S_b$, or $S_m^+ S_b$. Pseudoinverse matrix is a generalization of the inverse matrix and exists for any $m \times n$ matrix. The computationally simplest way to get the pseudoinverse is using SVD; if $A = U\Sigma V^T$ is the singular value decomposition of A, then the pseudoinverse $A^+ = V\Sigma^+ U^T$. For a diagonal matrix such as Σ, we get the pseudoinverse by taking the reciprocal of each nonzero element on the diagonal.

5. Results

5.1. *Classification using SBS + pLDA*

The confusion matrix in Table 1 presents the correct classification ratio ($\mathcal{CCR}$) of subject-dependent (Subject A, B, and C) and subject-independent (All) classification where the features of all the subjects are simply merged and normalized. We used the leave-one-out cross-validation method where a single observation taken from the samples is used as the test data while the remaining observations are used

for training the classifier. This is repeated such that each observation in the samples is used once as the test data.

Table 1.: Recognition results in rates (*error 0.00 = $\mathcal{CCR}$ 100%*). # of samples: 120 for each subject and 360 for All

Subject A ($\mathcal{CCR}$ % = **81%**)

	EQ1	EQ2	EQ3	EQ4	total*	error
EQ1	**22**	4	1	3	30	0.27
EQ2	3	**26**	1	0	30	0.13
EQ3	1	2	**23**	4	30	0.23
EQ4	3	0	1	**26**	30	0.13

Subject B ($\mathcal{CCR}$ % = **91%**)

	EQ1	EQ2	EQ3	EQ4	total*	error
EQ1	**27**	3	0	0	30	0.10
EQ2	3	**25**	1	1	30	0.17
EQ3	0	2	**28**	0	30	0.07
EQ4	0	1	0	**29**	30	0.03

Subject C ($\mathcal{CCR}$ % = **89%**)

	EQ1	EQ2	EQ3	EQ4	total*	error
EQ1	**28**	0	2	0	30	0.07
EQ2	0	**30**	0	0	30	0.00
EQ3	0	0	**24**	6	30	0.20
EQ4	0	0	5	**25**	30	0.17

All: Subject-independent ($\mathcal{CCR}$ % = **65%**)

	EQ1	EQ2	EQ3	EQ4	total*	error
EQ1	**62**	9	8	11	90	0.31
EQ2	15	**57**	13	5	90	0.37
EQ3	9	6	**58**	17	90	0.36
EQ4	8	5	21	**56**	90	0.38

*: Actual total # of samples

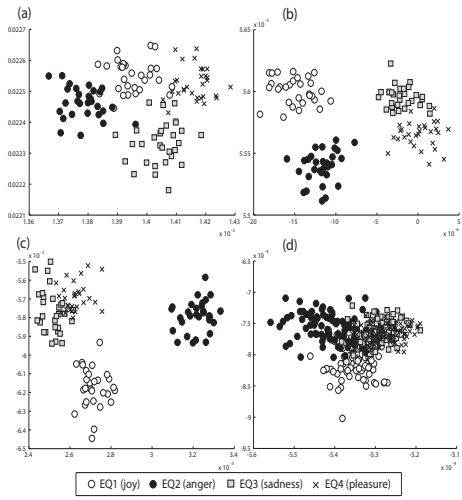

Fig. 10.: Comparison of feature distributions of subject-dependent and subject-independent case. (a) Subject A, (b) Subject B, (c) Subject C, (d) Subject-independent

The table shows that the $\mathcal{CCR}$ varies from subject to subject. For example, the best accuracy was 91% for subject B and the lowest was 81% for subject A. Not only does the overall accuracy differ from one subject to the next, but the $\mathcal{CCR}$ of the single emotions varies as well. For example, EQ2 was perfectly recognized for subject C while it caused the highest error rate for subject B. It was three times mixed up with EQ1 which is characterized by opposite valence. As the confusion matrix shows, the difficulty in valence differentiation can be observed for all subjects. Most classification errors for Subject A and B lie in false classification between EQ1 and EQ2 while an extreme uncertainty can be observed in the differentiation between EQ3 and EQ4 for Subject C. On the other hand, it is very meaningful that relatively robust recognition accuracy is achieved for the classification of emotions that are reciprocal in the diagonal quadrants of the 2D emotion model, i.e., EQ1 vs. EQ3 and EQ2 vs. EQ4. Moreover, the accuracy is much better than that of arousal classification. The $\mathcal{CCR}$ of subject-independent classification was not comparable to that obtained for subject-dependent classification. As shown in Figure 10, merging the features of all subjects does not refine the discriminating information related to the emotions, but rather leads to scattered class boundaries.

We also tried to differentiate the emotions based on the two axes, arousal and valence, in the 2D emotion model. The samples of four emotions were divided into groups of negative valence (EQ2+EQ3) and positive valence (EQ1+EQ4) and into groups of high arousal (EQ1+EQ2) and low arousal (EQ3+EQ4). By using the same methods, we then performed a two-class classification of the divided samples for arousal and valence separately. It turned out that emotion-relevant ANS specificity can be observed more conspicuously in the arousal axis regardless of subject-dependent or independent cases. Classification of arousal achieved an acceptable $\mathcal{CCR}$ of 97-99% for the subject-dependent recognition and 89% for the subject-independent recognition, while the results for valence were 88-94% and 77%, respectively.

5.2. *Finding the best emotion-relevant ANS features*

In most literature dealing with emotion-relevant ANS specificity, a tendency analysis of physiological changes has been used to correlate ANS activity with certain emotional states, e.g. EQ1 with increased heart rate or anxiety with increased skin conductivity. Even for multiclass classification problems, however, such a direction analysis of physiological changes is not sufficient to capture accompanying multimodal ANS reactions that are cross-correlated with each other when using multichannel biosensors. Therefore, we tried to first identify the significant features for each classification problem and thereby to investigate class-relevant feature domain and interrelation between the features for a certain emotion.

In Table 2, the best emotion-relevant features, which we determined by ranking the features selected for all subjects (including Subject All) in each classification problem, are listed in detail by specifying their values and domains. One interesting result is that each classification problem respectively links together with a certain

Table 2.: Best emotion-relevant features extracted from four channel physiological signals. Arousal classes: EQ1+EQ2 vs. EQ3+EQ4, Valence classes: EQ1+EQ4 vs. EQ2+EQ3, Four classes: EQ1/EQ2/EQ3/EQ4.

Classes	Best emotion-relevant Features (Ch_value_domain, C: ECG, R: RSP, S: SC, M: EMG)
Arousal	C_std(diff)_HRVtime, C_**sd2_PoincareHRV**, C_powerLow_HRVspec, R_**meanEnergy_SubSpectra**, R_**sd2_PoincareBRV** R_**mean_MSE**, S_mean_RawLowpassed, S_std_RawLowpassed, M_#occurrenceRatio_RawLowpassed M_mean_RawNormed
Valence	C_**sd2_PoincareHRV**, C_meanEnergy_SubSpectra, C_ratioLH_HRVspec, C_mean_MSE, C_mean(diff)_MSE R_**meanEnergy_SubSpectra**, R_mean(diff)_SubSpectra, R_sd1_PoincareBRV, R_**sd2_PoincareBRV**, R_**mean_MSE** S_mean(diff)_RawNormed, M_mean(diff)_RawNormed
Four Emotions	C_mean_HRVtime, C_std_HRVtime, C_std(diff)_HRVtime, C_mean(diff)_MSE, C_mean_MSE, C_mean_SSE C_**sd2_PoincareHRV**, C_mean_SubSpectra, R_**meanEnergy_SubSpectra**, R_mean_SSE, R_mean_BRVtime R_sd1_PoincareBRV, R_**sd2_PoincareBRV**, R_**mean_MSE**, R_power_BRVspec, S_std_RawLowpassed S_mean(diff)_RawNormed, S_mean(diff(diff))_RawLowpassed, S_mean_RawNormed, S_#occurrence_RawLowpassed M_mean(diff)_RawNormed

: overall selected features are printed in bold

feature domain. The features obtained from the time/frequency analysis of HRV time series are decisive for the classification of arousal and for the classificaton of the four emotions, while the features from the MSE domain of ECG signals are a predominant factor for correct valence differentiation. More particularly, mutually sympathizing correlate between HRV and BRV (firstly proposed in this paper) has been clearly observed in all the classification problems by the features from their time/frequency analysis and Poincaré domain, *_PoincareHRV* and *_PoincareBRV*. This reveals a manifest cross-correlation between respiration and cardiac activity with respect to emotional state. This is one of the most important findings for future work. In fact, in biomedicine, it is commonly accepted that the respiratory mechanism mediates high frequency components of HRV, but its specific role in affective ANS reactions has so far not been satisfactorily explained. When inhaling, the vagus nerve is impeded and the heart rate begins to increase, whereas this pattern is reversed when exhaling, i.e., the activation of the vagus nerve typically leads to a reduction in heart rate, blood pressure, or both[c]. Apart from its influence on the heart rate, the vagus nerve is also responsible for sweating, several muscle movements in the mouth, and even for speech. It means that most physiological channels we used are innately correlated with each other and respond together as a chain reaction to emotional stimulation. For example, when the parasympathetic nerves overcompensates a strong response from the sympathetic nervous system innervating the sinoatrial node, which occurs in cases of extreme stress or fear, the reduction in heart rate and blood pressure becomes proportionally faster to the intensity of the emotion.

Our feature analysis proves that the correlation between heart rate and respiration is obviously captured by the features from the HRV power spectrum (*_HRVspec*), the fast/long-term HRV/BRV analysis using the Poincaré method, and the multiscale variance analysis of HRV/BRV (*_MSE*). It also demonstrates that the peaks of high frequency range in the HR subband spectrum (*_SubSpectra*) provide information about how the sinoatrial node responds to vagal activity at certain respiration frequencies.

5.3. *Emotion-specific multilevel dichotomous classification*

Most common classifiers are best-suited to handle two-class problems. The pLDA we used is no exception to this and assumes that the covariance matrices of each class are the same or at least close to each other for multiclass ($c > 2$) classification. Consequently, the performance of pLDA in multiclass classification could be suboptimal depending on the difference between the covariance matrices of each class. In our work, we actually used the averaged covariance to directly solve the multiclass

[c]The influence of breathing on the flow of the sympathetic and vagus impulses to the sinoatrial node causes the so called respiratory sinus arrhythmia (RSA). The degree of fluctuation in heart rate is also significantly controlled by regular impulses from the baroreceptors in the aorta and carotid arteries.

problem using a single pLDA classifier. One straightforward way to handle a multiclass problem by using binary classifiers is to decompose the multiple categories into a set of complementary two-class problems.

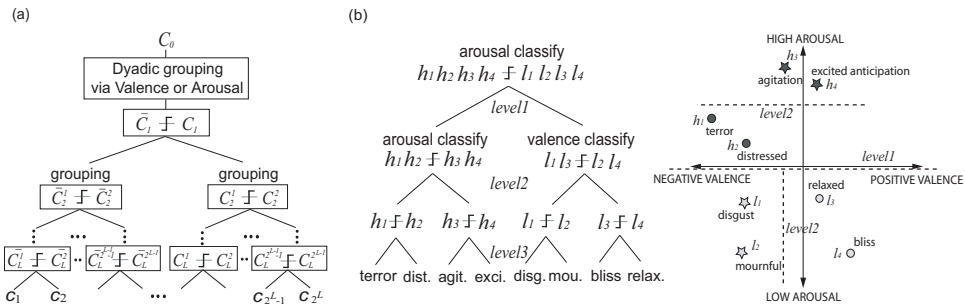

Fig. 11.: Framework of emotion-specific multilevel dichotomous classification (EMDC). (a) Diagram of decomposition process, (b) Decomposition example for an eight-class problem

By taking advantage of supervised classification (where we know in advance which emotion types have to be recognized) we developed an emotion-specific multilevel dichotomous classification (EMDC) scheme. This scheme exploits the property of the dichotomous categorization in the 2D emotion model and the fact that arousal classification yields higher $\mathcal{CCR}$ than valence classification or direct multiclass classification. This proves true in almost all previous works and according to our results as well. Figure 11 illustrates the EMDC scheme and provides an example of the dyadic decomposition for the eight-class problem.

First, the entire training patterns are grouped into two opposing *"superclasses"* (on the basis of valence or arousal), $\bar{C}$ consisting of all patterns in some subset of the class categories and C as all remaining patterns, i.e., $\bar{C} \cap C = \{\}$. This dyadic decomposition using one of the two axes is serially performed until one subset contains only two classes. The grouping axis can be different from each dichotomous level. Then multiple binary classifiers for each level are trained from the corresponding dyadic patterns. Therefore, the EMDC scheme is obviously emotion-specific and effective for a 2D emotion model. Note that the performance of the EMDC scheme is limited by a maximum $\mathcal{CCR}$ of first level classification and makes sense only if the $\mathcal{CCR}$ for one of the two superclasses is higher than that for direct multiclass classification (theoretically this always holds true for our case). Because we used four emotion classes in our experiment, we needed a two-level classification based on arousal and valence grouping for both superclasses in parallel.

Table 3 shows the dichotomous contingency table of recognition results by using the novel EMDC scheme. The best feature sets shown in Table 2 are used for the binary classification at each level. As expected, the $\mathcal{CCR}s$ significantly improve for all class problems. For the classification of four emotions, we obtained an average $\mathcal{CCR}$ of 95% for subject-dependent and 70% for subject-independent classification.

Table 3.: Results using EMDC scheme with the best features

Subject A (CCR % = 94%, 113/120)

	EQ1 & EQ2			EQ3 & EQ4
EQ1 & EQ2		58		2
		EQ1	EQ2	
	EQ1	27	1	
	EQ2	1	29	
EQ3 & EQ4	2			58
				EQ3 EQ4
			EQ3	29 0
			EQ4	1 28

Subject C (CCR % = 94%, 113/120)

	EQ1 & EQ2			EQ3 & EQ4
EQ1 & EQ2		60		0
		EQ1	EQ2	
	EQ1	30	0	
	EQ2	0	30	
EQ3 & EQ4	1			59
				EQ3 EQ4
			EQ3	27 2
			EQ4	4 26

Subject B (CCR % = 98%, 117/120)

	EQ1 & EQ2			EQ3 & EQ4
EQ1 & EQ2		60		0
		EQ1	EQ2	
	EQ1	30	0	
	EQ2	0	30	
EQ3 & EQ4	1			59
				EQ3 EQ4
			EQ3	29 1
			EQ4	1 28

Subject All (CCR % = 70%, 251/360)

	EQ1 & EQ2			EQ3 & EQ4
EQ1 & EQ2		155		25
		EQ1	EQ2	
	EQ1	62	13	
	EQ2	15	65	
EQ3 & EQ4	16			164
				EQ3 EQ4
			EQ3	64 19
			EQ4	21 60

CCR Comparison between pLDA and EMDC

	Subj. A	Subj. B	Subj. C	Subj. All	*Average (ABC)*
pLDA	81%	91%	89%	65%	*87%*
EMDC	**94%**	**98%**	**94%**	**70%**	***95%***

Compared with the results obtained for pLDA, the EMDC scheme achieved an overall $\mathcal{CCR}$ improvement of about 5%-13% in each class problem.

6. Discussion

We achieved an overall $\mathcal{CCR}$ of 95%, which is more than three times higher than chance probability, for four emotional states from three subjects. This should be sufficient to support the view that emotions, either produced or perceived while listening to music, exist and are accompanied by physiological differences in both the arousal and valence dimensions such that they can eventually be recognized by the machine. At the same time, however, some issues remain in relation to the processing stages of our recognition system.

Recording physiological changes using biosensors is still invasive since the subjects, for example, have to be in physical contact with adhesive electrodes. Furthermore, most biosensors using such electrodes are very susceptible to motion artifacts which we could observe in almost all signals of our dataset. For practical HCI applications, it is therefore necessary to develop non-invasive biosensors, preferably with built-in denoising filters in wirelessly miniaturized form. We expect that today's nano-technology will help design such hardware soon. This would then improve not only the signal quality and the usability of the technology, but also reduce computational costs in the preprocessing stage.

Our analysis results based on the best emotion-relevant features are incontrovertibly useful findings; for example, the consistent tendency of the feature contents to valence and arousal differentiation separately and the proven efficiency of new feature domains that are firstly considered in this paper. We should, however, note

that the effectiveness of the best features might not be universally guaranteed for other datasets or classifiers. First, only three subjects might not be sufficient to generalize the features. Second, the SBS as well as most algorithms for feature selection use a criterion based on a specific classifier and are therefore effective only if the classifier used is known in advance. In addition, such sequential algorithms may lead to suboptimal subsets due to their unidirectional property, i.e. once a feature is added or removed, this action can never be reversed.

By dividing given patterns using the arousal and valence axis in the 2D emotion model, we proposed the EMDC scheme which contributed to a significant improvement of the recognition results. The scheme may, however, still be adjusted in several ways. For instance, since it needs multiple classifiers to be trained for each level, the combination of different classifiers seems to be feasible. By taking advantage of the fact that EMDC enables us to view the classification results of each level in multiresolution aspect (see Table 3), the scheme could be more sophisticatedly designed thanks to the parametric refining of each binary classifier depending on the level.

The reason for the great disparity of $\mathcal{CCR}$ between subject-dependent and independent classification can be explained in many different ways indeed. We mention that one of the main factors in the difficulty of subject-independent classification is the intricate variety of non-emotional individual contexts among the subjects, rather than an individual ANS specificity in emotion. A naive idea for improving the performance of the user-independent system for practical applications would be to first identify the user, prior to starting the recognition process, and then to classify a user's emotion in a user-dependent way. Of course, this is feasible only if the number of users is finite and the users are known to the system, or if the system can cumulatively collect the data of each user in a learning phase.

7. Future Work

One of the most challenging issues in the near future will be to explore multimodal analysis for emotion recognition. We humans use several modalities jointly to interpret emotional states, since emotion affects almost all modes- audiovisual (facial expression, voice, gesture, posture, etc.), physiological (respiration, skin temperature etc.), and contextual (goal, preference, environment, social situation, etc.) states in human communication. In the recent literature, findings concerning emotion recognition by combining multiple modalities have been reported; mostly by fusing features extracted from audiovisual modalities such as facial expression and speech. However, we note that combining multiple modalities by equally weighting them does not always guarantee improved accuracy. The more crucial issue is how to *complementarily* combine the additional modalities. An essential step towards a human-like analysis and finer resolution of recognizable emotion classes would therefore be to find the innate priority among the modalities to be preferred for each

emotional state. Then, an ambitious undertaking might be to decompose an emotion recognition problem into several refining processes using additional modalities, for example: arousal recognition through physiological channels, valence recognition by using audiovisual channels, and then resolving of subtle uncertainties between adjacent emotion classes.

References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, Emotion recognition in human-computer interaction, *IEEE Signal Processing Mag.* **18**, 32–80, (2001).

[2] R. Picard, E. Vyzas, and J. Healy, Toward machine emotional intelligence: Analysis of affective physiological state, *IEEE Trans. Pattern Anal. and Machine Intell.* **23** (10), 1175–1191, (2001).

[3] F. Nasoz, K. Alvarez, C. Lisetti, and N. Finkelstein, Emotion recognition from physiological signals for presence technologies, *International Journal of Cognition, Technology, and Work - Special Issue on Presence.* **6(1)**, (2003).

[4] J. J. Gross and R. W. Levenson, Emotion elicitation using films, *Cognition and Emotion.* **9**, 87–108, (1995).

[5] A. Haag, S. Goronzy, P. Schaich, and J. Williams. Emotion recognition using bio-sensors: First steps towards an automatic system. In *ADA 2004*, pp. 36–48, (2004).

[6] Center for the Study of Emotion and Attention [CSEA-NIMH]. *The International Affective Picture System: Digitized Photographs.* Gainesville, FL: Center for Research in Psychophysiology, University of Florida, (1995).

[7] K. H. Kim, S. W. Bang, and S. R. Kim, Emotion recognition system using short-term monitoring of physiological signals, *Medical & Biological Engineering & Computing.* **42**, 419–427, (2004).

[8] A. Jain, R. Duin, and J. Mao, Statistical pattern recognition: A review, *IEEE Trans. Pattern Anal. and Machine Intell.* **22**(1), 4–37, (2000).

[9] M. B. Mcllory and M. D. Cheithin, *Clinical Cardiology.* (VLANGE Medical Book, 1990), 5th edn edition.

[10] J. Pan and W. Tompkins, A real-time qrs detection algorithm, *IEEE Trans. Biomed. Eng.* **32**(3), 230–323, (1985).

[11] M. Costa, A. L. Goldberger, and C.-K. Peng, Multiscale entropy analysis of biological signals, *Phys. Rev.* **E 71**(021906), (2005).

[12] A. Jain and D. Zongker, Feature selection: Evaluation, application, and small sample performance, *IEEE Trans. Pattern Anal. and Machine Intell.* **19**, 153–163 (Feb., 1997).

[13] D. L. Swets and J. Weng, Using discriminant eigenfeatures for image retrieval, *IEEE Trans. Pattern Anal. and Machine Intell.* **18**(8), 831–836 (Aug, 1996).