34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Multimodal Emotion Recognition by Combining Physiological Signals and Facial Expressions: A Preliminary Study

Jukka Kortelainen, *Member, IEEE*, Suvi Tiinanen, Xiaohua Huang, Xiaobai Li, Seppo Laukka, Matti Pietikäinen, and Tapio Seppänen

Abstract— **Lately, multimodal approaches for automatic emotion recognition have gained significant scientific interest. In this paper, emotion recognition by combining physiological signals and facial expressions was studied. Heart rate variability parameters, respiration frequency, and facial expressions were used to classify person's emotions while watching pictures with emotional content. Three classes were used for both valence and arousal. The preliminary results show that, over the proposed channels, detecting arousal seem to be easier compared to valence. While the classification performance of 54.5% was attained with arousal, only 38.0% of the samples were classified correctly in terms of valence. In future, additional modalities as well as feature selection will be utilized to improve the results.**

I. Introduction

THE on-going new mega-trend of ubiquitous computing has resulted in that research on human-centered computing systems has become a major topic of interest both in academia and industry. This means that the future technological solutions for human-computer interaction must have much better capabilities for understanding human behavior than is currently the case. The need has created a new field of research, affective computing, aiming to improve the interaction by including the interpretation of the emotional state of humans to the functionality of machines. By adapting to the emotions, more appropriate response to the user could be given.

One major domain in affective computing is automatic recognition of human emotions. Emotions are fundamental for humans, impacting perception and everyday activities such as communication, learning and decision-making. They are expressed through speech, facial expressions, gestures and other non-verbal clues. Many physiological signals also convey information about the emotional state of humans. Automatic emotion recognition capabilities using various sensor signals can therefore be developed for the computer.

Lately, multimodal approaches for automatic emotion recognition have gained significant scientific interest. These include the combination of facial expressions, audio, and different kinds of physiological signals. While utilizing multiple physiological signals as well as combining facial expressions with audio seem to be rather common (see e.g. [1], [2]), only few studies report the simultaneous application of physiological signals and facial expressions for emotion recognition [3]–[5]. Facial expression is usually the most important non-verbal communication channel. However, the actual expression is affected by the context of the social situation, for which reason it may not reflect the actual feeling of the person. Physiological signals that we are interested in are controlled by autonomous nervous system, which makes them hard to control voluntarily by the test person in different contexts. Thus, combination of these channels may offer potential for more accurate recognition of emotion.

In this paper, multimodal emotion recognition by combining physiological signals and facial expressions is studied. Heart rate variability (HRV) parameters, respiration frequency, and facial expressions are used to classify person's emotions while watching pictures with emotional content. The experimental protocol, parameter extraction from physiological signals and facial expressions, and emotion classification are explained in Section II. In Section III, the results are presented. Section IV concludes the paper by giving a short discussion about the results and future work.

II. Materials and Methods

A. Experimental Protocol and Data

For the study, an experimental protocol with 24 female undergraduate students was carried out. The students participated voluntarily in the experiment for their psychology studies and the consent were signed after they were explained about the experimental protocol.

In the experiment, participants were asked to sit in front of a 17-inch computer screen, while 48 pictures with emotional content from IAPS (International Affective Picture System) were presented. At the same time facial expressions were recorded by an IEEE 1394 firewire camera (Sony DFWVL500, Japan) which was set on the top of the screen. In addition, the participant's heart rate was measured using a Polar S810i heart rate monitoring system.

The 48 pictures used in the experiment were divided into three categories based on their emotional valence rating

This work was supported in part by grant 40297/11 from Tekes.

J. Kortelainen is with the Department of Computer Science and Engineering, BOX 4500, FIN-90014 University of Oulu, Finland (e-mail: jukka.kortelainen@ee.oulu.fi).

S. Tiinanen and T. Seppänen are with the Department of Computer Science and Engineering, University of Oulu, Finland.

X. Huang, X. Li, and M. Pietikäinen are with the Center of Machine Vision Research, Department of Computer Science and Engineering, University of Oulu, Finland.

S. Laukka is with the LearnLab, Department of Educational Sciences and Teacher Education, University of Oulu, Finland.

TABLE I
ESTIMATED HRV PARAMETERS

Parameter (unit)	Description
HR (beat/min)	Mean heart rate per minute
RR (ms)	Mean heart rate interval
SDNN (ms)	Standard deviation of RR intervals
RMSSD (ms)	Square root of the mean squared difference of successive RRs
NN50	Number of pairs of successive RRs that differ by more than 50 ms
pNN50 (%)	Number of NN50 divided by total number of RRs
LF (ms^2)	Low frequency (LF, 0.04-0.15 Hz) power
HF (ms^2)	High frequency (HF, 0.15-0.4 Hz) power
LF/HF (%)	LF/HF ratio i.e. sympathovagal balance
LFn (%)	Normalized LF power, calculated by dividing LF with sum of LF and HF
HFn (%)	Normalized HF power, calculated by dividing HF with sum of LF and HF

from the IAPS: pleasant, neutral, and unpleasant. Each of the categories contained 16 pictures. In the experiment, the pictures were presented one by one in a randomized order. For each picture, the participant's task included three sections: watching the picture on the screen for 20 seconds, reporting her feeling about the picture orally according to the Self-Assessment Manikin (SAM, [6]) scale using valence (SAMV; 1 for positive, 2 for neutral, and 3 for negative) and arousal (SAMA; 9 point scale; 1 for very calm and 9 for very aroused), and orally describing the content of the content of the picture for and observer.

B. Physiological Signals

Standard HRV parameters [7] were calculated from beat-to-beat RR-intervals provided by the monitoring system. The parameters were determined from the sequences corresponding to the presentation of the pictures. Estimated time and frequency domain parameters with their descriptions are listed in Table I. For frequency domain analysis the RR-interval signal was first interpolated at 4Hz and linear trends were removed. Power spectrum densities (PSD) were calculated by using autoregressive (AR) modeling (Burg's method) with model order 18, which is generally used when analyzing RR-intervals [7]. A representative RR-interval signal after interpolation and the corresponding PSD during the presentation of one picture is given in Fig. 1.

Respiration frequency during the presentation of the pictures was estimated from the RR-intervals using spectral decomposition [8]. In this approach, the roots of the AR coefficients are determined, after which their phase angles as frequencies that lie on the high frequency range are detected. The pole which produces the highest peak of spectrum determines the respiration frequency.

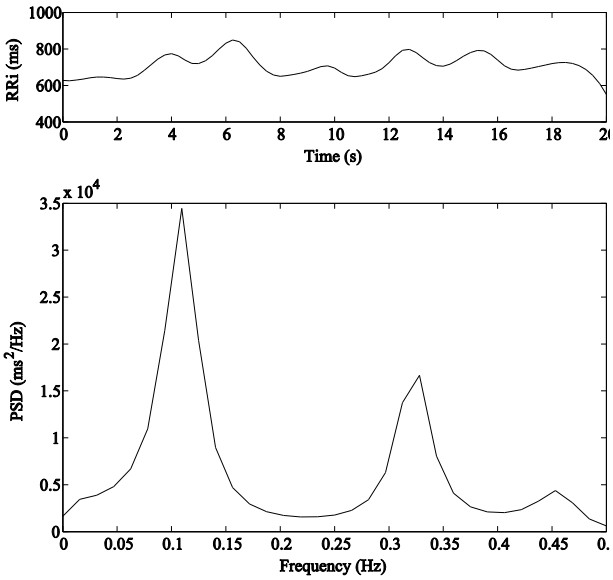

Fig. 1. A representative RR-interval signal after interpolation and the corresponding PSD during the presentation of one picture.

C. Facial Expressions

In facial expression research, the classification of new data is usually approached by cross-database validation. In other words, before applied to the new data, the classifier is trained using a database with annotated facial expression data. In the proposed study, the extended Cohn-Kanade database (CK+) [9] was used. The CK+ database contains facial images labeled by seven basic emotions (happy, anger, disgust, fear, sadness, surprise, and contempt) and neutral emotion, from which 600 facial images were manually chosen and further classified as positive, neutral, or negative.

An overview of the methodology used for facial expression classification is illustrated in Fig.2. Firstly, the eyes are automatically detected by applying the algorithm presented in [10]. If the detection fails, the localization is performed manually. Secondly, then face roll rotation is estimated and corrected. The facial images are then cropped from the full images by fixing the eyes position. All facial images are normalized to 128×128 resolution. Local binary pattern (LBP) [11] operator is used as the facial feature. LBP is a gray-scale invariant texture primitive statistic, which has shown excellent performance in the classification of various kinds of textures. The operator is defined as:

$$\mathrm{LBP}_{S,R} = \sum_{s=0}^{S-1} f(g_s - g_c)2^s \qquad (1)$$

where

$$f(g_s - g_c) = \begin{cases} 1, g_s - g_c \geq 0 \\ 0, g_s - g_c < 0 \end{cases} \qquad (2)$$

and g_c is the gray value of the center pixel and g_s is the gray value of S equally spaced pixels on a circle of radius R at this center pixel. The facial images are divided into 6×6 non-overlapping blocks before application of the LBP operator with $R = 3$ and $S = 8$ for each block. The resulting LBP

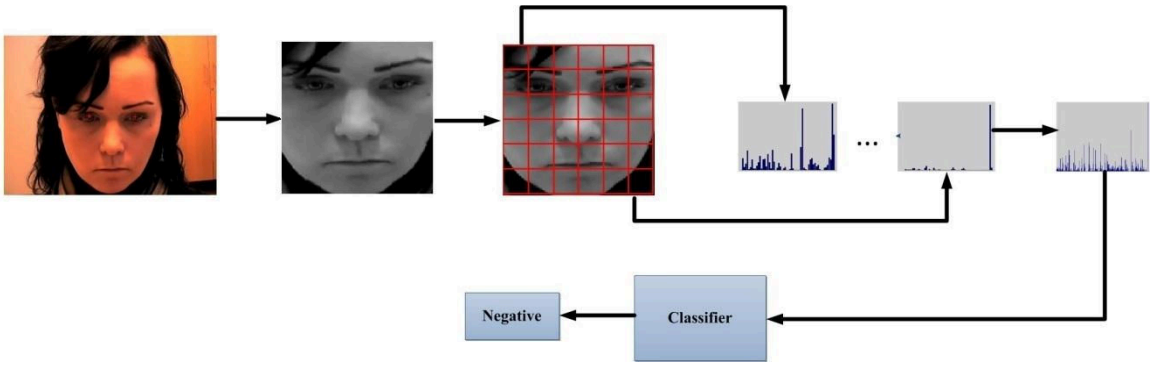

Fig. 2. The framework of facial expression recognition.

histograms are concatenated into a facial description. In order to keep the majority patterns, uniform LBP is used. The approach produces feature data with the dimensionality of 2124 (6×6×59 bins from each block).

The preprocessing and feature extraction were done for the training database images and testing images extracted from videos recorded during the experimental protocol. The testing images included the frames from 1152 video recordings with duration of 20 s each (15 frames/s, frame resolution 320p×240p). The testing images were classified as positive, neutral, or negative using a support vector machine with a linear kernel [12] after training with the database images. In order to be able to later combine facial expression data with the physiological signal data, the percentage of positive, neutral, and negative frames during the presentation of each picture was calculated. This approach results in three dimensional data (values varying between 0 and 100) representing the facial expression classification results.

D. Emotion Classification

The HRV parameters, respiration frequency, and percentages of positive, neutral, and negative frames during the presentation of pictures were combined to form 15-dimensional feature data. The data comprised of 1152 samples (24 participants × 48 pictures). The features were normalized to have zero mean and unit standard deviation. A *k* nearest neighbors (KNN) classification was performed to determine the valence and arousal for each sample. The determination was performed by using the SAMV and SAMA values of the participant's other samples. Before the classification, the SAMA values were changed to range from 1 to 3, i.e. the values 1-3, 4-6, and 7-9 were changed to 1, 2, and 3, correspondingly.

III. Results

The dependencies between different features and SAMV and SAMA values were investigated. Generally, the correlation between the features and valence was found to be low. In Fig. 3, the samples are presented according to the percentage of positive frames and SAMV. Slightly more positive frames were observed in the video recordings when positive pictures were shown. The physiological parameters did not seem to correlate well with valence. However, with arousal, the results were found to be better. Fig. 4 shows the dependency between HFn parameter and SAMA. The decrease in the high frequency power indicates the decrease of parasympathetic nervous activity due to emotional strain.

The classification results were in line with the above-described findings. Fig. 5 and Fig. 6 illustrate the performance of valence and arousal as a function of *k*, respectively. With valence, the best results (38.0%) were

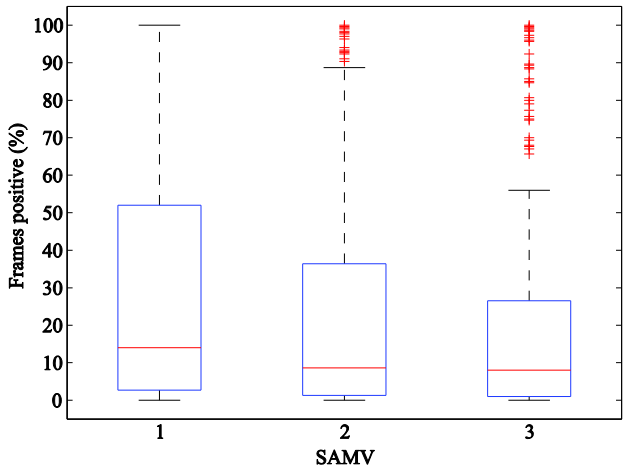

Fig. 3. The samples presented according to the percentage of positive frames during the presentation of pictures and SAMV.

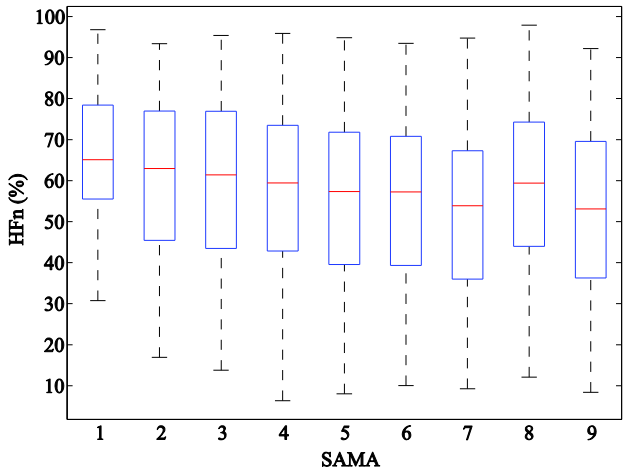

Fig. 4. The samples presented according to the normalized high frequency power (HFn) and SAMA.

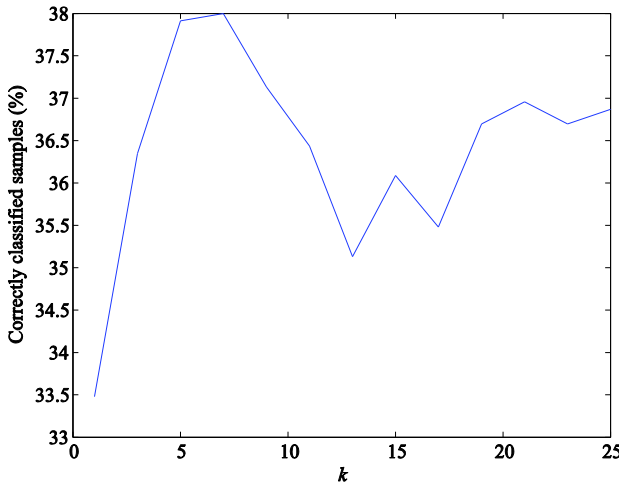

Fig. 5. The results of valence classification as a function of *k*.

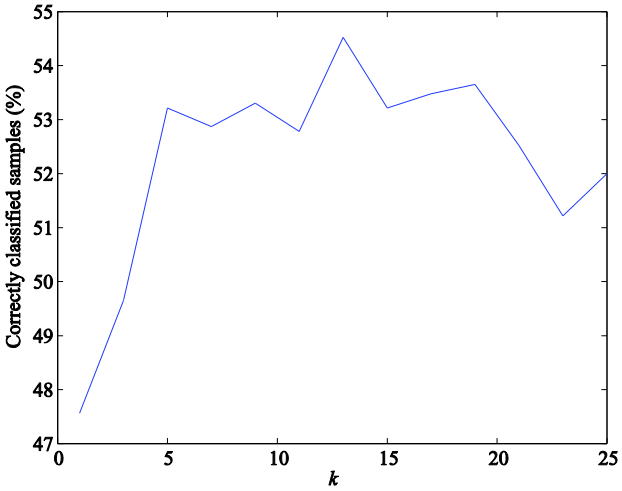

Fig. 6. The results of arousal classification as a function of *k*.

attained when *k* was 7. The performance only slightly exceeds the random guess (33.3%) reflecting the difficulty of the task in the presented experimental protocol. The results were better with arousal: 54.5% of the samples were classified correctly when *k* was 13. The results suggest that, with the proposed features and experimental protocol, detecting arousal seems to be easier compared to valence.

IV. Conclusions and Discussion

In this paper, multimodal emotion recognition by combining physiological signals and facial expressions was studied. HRV parameters, respiration frequency, and facial expressions were used to classify person's emotions while watching pictures with emotional content. The preliminary results with our data show that, over the proposed channels, detecting arousal seems to be easier compared to valence. While the classification performance of 54.5% was attained with arousal, only 38% of the samples were classified correctly in terms of valence.

As mentioned in the introduction, only few studies report the simultaneous application of physiological signals and facial expressions for emotion recognition [3]–[5]. None of these studies utilized respiration frequency or other HRV parameters than heart rate. In our study, the LBP-based facial expression classification was also combined for the first time with physiological signals.

Due to the variation in the material shown to the participants, measurement setup, classification procedure, and features used, results are difficult to compare with literature. One approach to validate the results would be to apply human classification, in which people would be asked to recognize the emotions of the persons in the videos recorded during the experimental protocol. This would provide a good reference for the classification results, especially for valence. In future, also the role of different modalities and the derived features in the classification should be validated. Additional modalities, such as galvanic skin conductance and electrocardiogram will be included in the experimental protocol as well.

References

[1] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Proc. IEEE International Conference in Automatic Face & Gesture Recognition*, Santa Barbara, CA, pp. 827–834, 2011.

[2] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion Recognition: a review", in *Proc. IEEE 7th International Colloquium on Signal Processing and its Applications*, Penang, Malaysia, pp. 410–415, 2011.

[3] Bailenson *et al.*, "Real-time classification of evoked emotions using facial features tracking and physiological responses," *Int. J. Human-Computer Studies*, vol. 66, pp. 303–317, 2008.

[4] C.-Y. Chang, J.-S. Tsai, C.-J. Wang, and P.-C. Chung, "Emotion recognition with consideration of facial expression and physiological signals," in *Proc. Computational Intelligence in Bioinformatics and Computational Biology*, Nashville, TN, pp. 278–283, 2009.

[5] I. Arapakis, I Konstas, and J. Jose, "Using facial expressions and peripheral physiological signals as implicit indicator of topical relevance," in *Proc. 17th ACM international conference on Multimedia*, Beijing, China, pp. 461–470, 2009.

[6] P. Lang, M. Bradley, and B. Cuthbert, "International affective picture system (IAPS): Affective rating of pictures and instruction manual," *Technical Report A-6*. University of Florida, Gainesville, FL, 2005.

[7] "Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology," *Circulation*, vol. 93, pp. 1043–1065, 1996.

[8] L. Zetterberg, "Estimation of parameters for a linear difference equation with application to EEG analysis," *Math. Biosci.*, vol. 5, pp. 227–275, 1969.

[9] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, pp. 94–101, 2010.

[10] Z. Niu, S. Shan, S. Yan, X. Chen, and W. Gao, "2D cascaded AdaBoost for eye Localization," in *Proc. 18th International Conference on Pattern Recognition*, Hong Kong, China, pp. 1216–1219, 2006.

[11] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: application to face recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, pp. 2037–2041, 2006.

[12] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011.