Facial Expression-Based Emotion Classification using Electrocardiogram and Respiration Signals

Dilranjan S. Wickramasuriya, *Student Member, IEEE*, Mikayla K. Tessmer, and Rose T. Faghih, *Member, IEEE*

***Abstract*—Automated emotion recognition from physiological signals is an ongoing research area. Many studies rely on self-reported emotion scores from subjects to generate classification labels. This can introduce labeling inconsistencies due to inter-subject variability. Facial expressions provide a more consistent means of generating labels. We generate labels by selecting locations at which subjects either displayed a visibly averse/negative reaction or laughed in video recordings. We next use a supervised learning approach for classifying these emotional responses based on electrocardiogram (EKG) and respiration signal features in an experiment where different movie/video clips were utilized to elicit feelings of joy, disgust, amusement, etc. As features, we extract wavelet coefficient patches from EKG RR-interval time series and respiration waveform parameters. We use principal component analysis for dimensionality reduction and support vector machines for classification. We achieved an overall classification accuracy of 78.3%.**

***Index Terms*—emotion recognition, continuous wavelet transform, RR-intervals, respiration**

I. Introduction

Automated emotion recognition is expected to play a crucial role in future human-computer interaction systems. Applications will include improved multimedia content recommendations [1], wearable monitoring for maintaining emotional well-being and smart living spaces. Variations in human emotion can be accounted for along two orthogonal axes—valence and arousal [2]. Valence denotes the pleasure-displeasure axis of emotion while arousal denotes its corresponding activation or excitement. A third axis known as dominance axis relates to the degree of control felt. Many emotion recognition methods utilize self-reported valence and arousal scores from subjects for generating classification labels.

Publicly available datasets for emotion recognition from physiological signals include the Database for Emotion Analysis using Physiological Signals (DEAP) [1], the Multi-Modal Database for Affect Recognition and Implicit Tagging (MAHNOB-HCI) [3] and the MEG-based Multimodal Database for Decoding Affective Physiological Responses (DECAF) [4]. The DEAP dataset uses a series of music videos to elicit different emotions in subjects and the MAHNOB-HCI dataset uses movie/video clips for doing so. The DECAF database uses both types of stimuli. The authors of the DEAP dataset noted considerable differences in self-reported arousal, valence and dominance scores provided by the subjects [1] (possibly due to differences in scale interpretation, music tastes and mood) and hence trained subject-specific classifiers for high vs. low emotion (arousal, valence and dominance) categorization. They obtained average valence classification accuracy values of 57.6% and 62.7% using electroencephalography (EEG) and peripheral features respectively. They further described the difficulty of eliciting strong valence responses at *low* arousal; extreme valence values were obtained at *high* arousal more easily. The authors of the MAHNOB-HCI dataset [3] obtained 57% and 45.5% accuracy values using EEG and peripheral signals respectively on three-class pleasant-neutral-unpleasant valence classification. Prior work in the literature has also made use of this dataset for emotion recognition from physiological signals (e.g. [5]–[9]). The accuracy values reported are in the 50–70's% range either on binary high–low or on three-class high–medium–low valence classification. While multimedia features from the video stimuli and eye movements have been used to enhance classification accuracy, we focus specifically on the use of physiological signals here.

Many methods for emotion recognition published in the literature utilize EEG signals. While EEG provides a direct measure of central nervous system activity, scalp EEG recordings are susceptible to motion artifact contamination and present a considerable challenge during wearable monitoring. Moreover, most classification approaches utilize the scores provided by the subjects themselves during trials or expert/population ratings as labels. As noted above, considerable inter-subject variability can exist in scores [1], and expert/population ratings may not necessarily correspond to what each particular subject felt. Hence, we use the subjects' own facial expressions to manually label the data. Manual facial expression analysis for generating emotion annotations was also used in [10], [11].

Both electrocardiogram (EKG) and respiration signals have simple, repetitive elements to them; R-peak detection in EKGs is also quite robust to noise (R-peaks in an EKG accompany ventricular contraction). RR-interval variations help diagnose cardiac arrhythmia and several other disorders [12]. Three different regions can be identified in the RR-intervals spectrum—Very Low Frequency (VLF, 0.003–0.04 Hz), Low Frequency (LF, 0.04–0.15 Hz) and High Frequency (HF, 0.15–0.4 Hz) [12]. The HF band depends on respiration patterns and the VLF band is chiefly determined by physical activity. The LF band is affected by both the sympathetic and parasympathetic nervous systems, and an increase in LF power is generally interpreted as a rise in sympathetic activity [12]. We utilize EKG and respiration features for valence classification.

D. S. Wickramasuriya and R. T. Faghih are with the Department of Electrical and Computer Engineering at the University of Houston, Houston, TX 77004 USA. M. K. Tessmer is with the Chemical and Biochemical Engineering Department at Missouri University of Science and Technology, Rolla, MO 65409 USA (e-mail:{dswickramasuriya, rtfaghih}@uh.edu, mktkx7@mst.edu). This work was partly supported by the following NSF grants: 1) 1755780 – CRII: CPS: Wearable-Machine Interface Architectures; 2) 1757949 – REU Site: Neurotechnologies to Help the Body Move, Heal, and Feel Again. Correspondence should be addressed to senior author Rose T. Faghih.

Fig. 1. **Subject facial expressions while watching the movie clips.** The three images on the left depict locations where the subjects laughed (PV class) and the three on the right are taken from locations where the subjects displayed averse reactions (NV class). The video data was accessed as per the End User License Agreement (EULA) signed with the MAHNOB-HCI database adiministrators [3].

II. Methods

A. Data

The MAHNOB-HCI dataset consists of two experiments—an emotion elicitation experiment and a multimedia content tagging experiment. Here, we only used the data from the first experiment where subjects were shown 20 movie clips meant to evoke different emotional responses ranging from fear to joy. A total of 27 subjects took part. Different physiological signals including EEG, EKG, skin conductance and respiration, along with frontal facial video were recorded from the subjects. Video data from some of the subjects are missing due to technical difficulties and not all the subjects provided consent for their recorded data to be published. We excluded these subjects from our analysis and only considered the data from the remaining 19.

B. Labeling

The MAHNOB-HCI dataset contains arousal and valence scores on a scale of 1–9 provided by each subject for each movie clip. However, we decided against using self-reported scores due to inter-subject variability. Moreover, expert/population ratings may not necessarily reflect the emotional response of each individual subject. For instance, one of the movie clips depicting a person placing a worm in his mouth (primarily meant to elicit disgust) was found to be funny by quite a few subjects. We therefore decided to manually label the data using the subjects' own facial expressions. Many clips without significant visible emotional reactions had to be discarded because the subjects largely had neutral expressions during most of the trials.

We viewed all the facial video recordings and manually identified the locations where the subjects either laughed or exhibited a noticeable averse/negative reaction. Occasionally, the subjects would be startled or frightened by something they watched, but then begin to laugh. Such instances were excluded (these would lead to ambiguous labels). Moreover, subjects sometimes smile briefly or even maintain a slight smile over an extended period of time for a particular movie clip. We excluded these locations with brief smiles as well, and considered just the extreme end of the positive emotions, i.e., laughter. These two categories—the laughter and the aversion—were named the positive valence (PV) and negative valence (NV) classes respectively. Fig. 1 shows some video frame examples from both categories. There were 23 instances belonging to the PV class and another 23 in the NV class.

C. Feature Extraction

We extracted EKG and respiration features for valence classification from each of the locations where subjects displayed visible emotional reactions. For each subject, we first detected the R-peaks in the EKG signals using MATLAB's *findpeaks* function and manually corrected erroneous detections. We next re-sampled the RR-interval time series at 7 Hz similar to the default option in PhysioNet's Cardiovascular Signal Toolbox [13]. We next took the continuous wavelet transform (CWT) of this new RR-interval signal utilizing a Morse wavelet (Fig. 2). In our analysis of the wavelet coefficients, we noted the general tendency for large coefficients (i.e., very bright regions in the 2D time-frequency plane) to occur within the 0.05–0.2 Hz range shortly after laughter. As this frequency range is strictly larger than the LF band, we call it the extended LF band (ELF).

The magnitudes of the wavelet coefficients vary from subject to subject, and do not necessarily fall within the same range. Normalization, therefore, is necessary. To do so, we first summed the energies of all the wavelet coefficients in the ELF band at each time instant and then obtained their median value. This is the median energy within the ELF band throughout the entire experiment. We normalized all the wavelet coefficients in the ELF band by dividing by this median value. Next, for each selected point in time where a subject displayed a visible emotional response (i.e., either laughter or an averse reaction) we extracted a 6×100-sized wavelet coefficient patch in the ELF range having the largest absolute sum of coefficients within a minute of that response occurring. If a subject laughed at more than one point when viewing a particular clip, we considered the time period from the first time laughter occurred to the period up to a minute after the last time the subject laughed.

Secondly, we extracted two features from the subjects' respiration signals. Laughter tends to be associated with larger inhalations almost immediately afterwards and we extracted the amplitude of the largest inhalation, and the area underneath its waveform in the 15 s following an emotional response. Fig. 3 shows a respiration signal in the 15 s period following such an emotional response. The area underneath the largest inhalation has been shaded and the height of the triangular-shaped region corresponds to the inhalation amplitude.

D. Feature Reduction

Each wavelet time-frequency patch extracted from the RR-interval time series has 600 elements to it. Therefore, we used

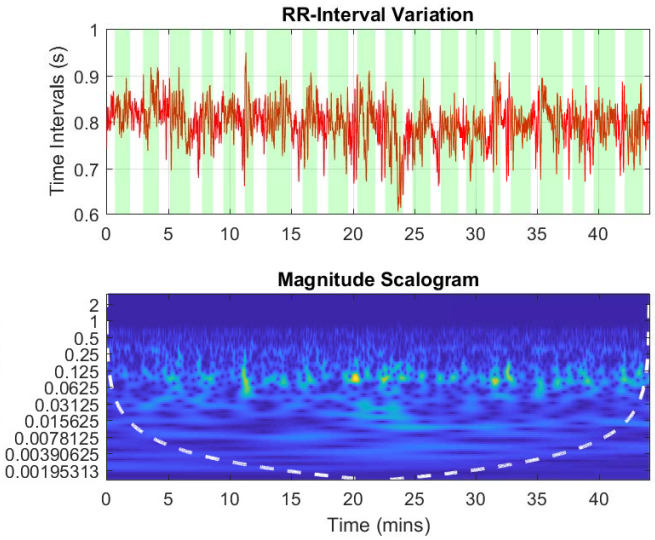

Fig. 2. **An RR-interval times series and its CWT for a particular subject.** The pale green strips in the upper sub-panel denote the movie clip presentation times.

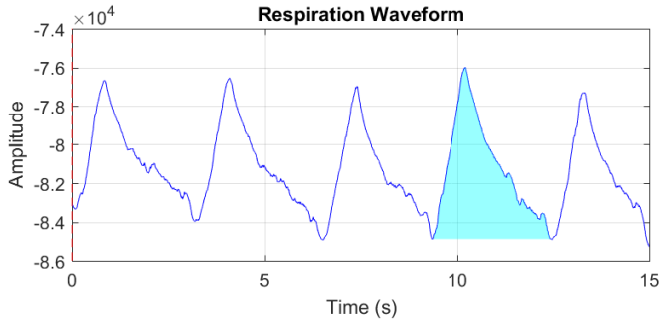

Fig. 3. **Part of a respiration waveform for a particular subject.** The shaded area and its height (corresponding to the largest inhalation/breath) are the two respiration features extracted.

principal component analysis (PCA) to reduce the dimensionality of the patches to five. With the inclusion of the two respiration measurements, a total of seven features were finally chosen for classification. The heart rate and respiration feature distributions are shown in Figs. 4 and 5.

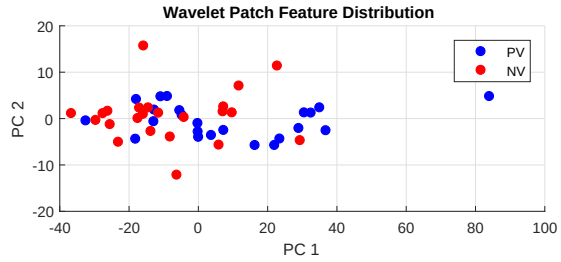

Fig. 4. **Wavelet patch feature distribution.** The 6×100-sized wavelet patches are unrolled into vectors and PCA is applied to reduce their dimensionality. Here, only the first two PCs are shown.

E. Classification

We used a linear support vector machine (SVM) in MATLAB's Classification Learner interface to classify the heart rate and respiration features. An SVM belongs to the category of large-margin classifiers that attempts to find a decision boundary maximizing the distance between two classes. Linear SVMs, unlike kernel-based SVMs, do not transfer the feature vectors into a high-dimensional space for classification.

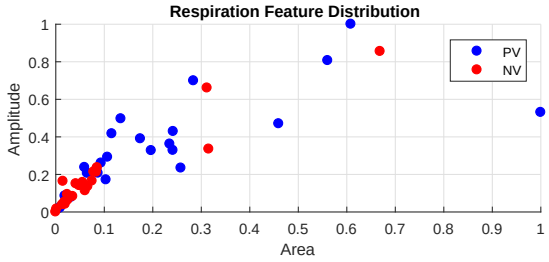

Fig. 5. **Respiration feature distribution.** The features are the amplitude and area of the largest breath shortly after an emotional response.

III. Results

We evaluated a linear SVM classifier using 10-fold cross-validation and the results are shown in Table I.

TABLE I
VALENCE CLASSIFICATION ACCURACY (PV IS THE POSITIVE CLASS)

Sensitivity (%)	Specificity (%)	Accuracy (%)
73.91	82.61	78.3

Several methods based on EEG and peripheral feature extraction published in the literature report classification accuracy values in the 50-70's% range on binary or three-class emotional valence recognition using movie clips as the stimuli [5]–[9]. Here, we obtained an accuracy close to 80% considering that we restrict ourselves to a smaller version of the problem, i.e., classification of the extreme valence reactions when subjects either laughed or visibly displayed averse reactions.

IV. Discussion and Conclusions

Valence classification based on physiological signals often utilizes subject-provided scores or expert/population ratings. Inter-subject variability in the scores can however be problematic as it introduces a lack of consistency in the labels. While the use of expert/population ratings solves this problem, these ratings may not necessarily correspond to what each particular subject felt when exposed to a trial stimulus. Here, we manually identified locations where subjects displayed emotional responses based on facial video recordings to generate labels. Annotating emotional valence based on a manual analysis of facial expressions was also performed in [10], [11]. Our objective here was to obtain cleaner labels via manual facial expression analysis. Generating labels in this manner could however, be susceptible to errors if subjects were to mask their expressions deliberately or not show any expressions.

We extracted two different types of features derived from EKG and respiration signals and achieved reasonably high

classification accuracy using linear SVMs. Many existing arousal and valence classification algorithms use multi-modal features including EEG and skin conductance. The accuracy of our method could be improved further by the inclusion of additional features. Skin conductance [14]–[16] and electromyogram data [17] are known to contain information regarding a person's emotions.

Future work would include investigating physiological signal changes elicited using musical video stimuli instead of movie clips. Eliciting emotions using music videos as opposed to movie clips is more challenging. Here, emotion classification accuracy can be lower as well [4]. Fig. 6 shows a preliminary result from a subject's RR-interval time series in the DEAP dataset (where emotions were elicited using music videos). Two very bright spots are visible in the time-frequency plane at approximately 0.125 Hz during clip 2 and clip 7 (clip 7 concludes shortly before the 10 min. mark). The subject laughed at both these time instances. Therefore, the wavelet-based decomposition of RR-intervals appears to encode valence information even for musical stimuli.

Moreover, machine learning is the dominant method in the literature for emotion recognition. Developing a Bayesian filter [18] for tracking valence within a state-space framework, similar to the skin conductance-based arousal tracking algorithm proposed in [19], would be another potential future direction.

This work presents preliminary results for valence classification using heart rate and respiration features. Further investigation with more subjects would be necessary to confirm its viability for clinical emotion recognition applications.

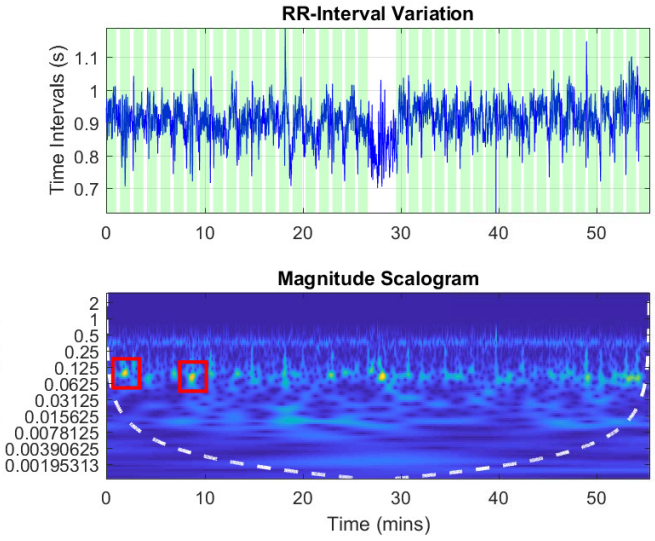

Fig. 6. **An RR-interval time series and its CWT for subject 10 in the DEAP dataset**. Two high energy bright spots can be seen on the spectrogram during the times corresponding to music video #2 and #7. The bright spots appear at approximately 0.125 Hz and coincide with the subject laughing.

V. Acknowledgment

"Portions of the research in this paper uses the MAHNOB Database collected by Professor Pantic and the iBUG group at Imperial College London, and in part collected in collaboration with Prof. Pun and his team of University of Geneva, in the scope of MAHNOB project financially supported by the European Research Council under the European Community's 7th Framework Programme (FP7/20072013) / ERC Starting Grant agreement No. 203143 [3]."

References

[1] S. Koelstra, C. Muhl, 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 Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, 2012.

[2] J. A. Russell, "Evidence of convergent validity on the dimensions of affect.," *J. Personality and Social Psychology*, vol. 36, no. 10, p. 1152, 1978.

[3] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, 2012.

[4] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-based multimodal database for decoding affective physiological responses," *IEEE Trans. Affect. Comput.*, vol. 6, no. 3, pp. 209–222, 2015.

[5] S. Chen, Z. Gao, and S. Wang, "Emotion recognition from peripheral physiological signals enhanced by EEG," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp. 2827–2831, 2016.

[6] H. Xu and K. N. Plataniotis, "Subject independent affective states classification using EEG signals," in *Proc. IEEE Global Conf. Signal and Information Processing*, pp. 1312–1316, 2015.

[7] M. B. H. Wiem and Z. Lachiri, "Emotion sensing from physiological signals using three defined areas in arousal-valence model," in *Proc. Int. Conf. Control, Automation and Diagnosis*, pp. 219–223, 2017.

[8] H. Ferdinando, T. Seppänen, and E. Alasaarela, "Comparing features from ECG pattern and HRV analysis for emotion recognition system.," in *Proc. IEEE Conf. Computational Intelligence in Bioinformatics and Computational Biology*, pp. 1–6, 2016.

[9] Y. Zhu, S. Wang, and Q. Ji, "Emotion recognition from users' EEG signals with the help of stimulus videos," in *Proc. IEEE Int. Conf. Multimedia and Expo*, pp. 1–6, 2014.

[10] M. Soleymani, S. Asghari-Esfeden, M. Pantic, and Y. Fu, "Continuous emotion detection using EEG signals and facial expressions," in *Proc. IEEE Int. Conf. Multimedia and Expo*, pp. 1–6, 2014.

[11] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 17–28, 2015.

[12] G. Ernst, *Heart rate variability*. Springer, 2014.

[13] A. N. Vest, G. Da Poian, Q. Li, C. Liu, S. Nemati, A. J. Shah, and G. D. Clifford, "An open source benchmarked toolbox for cardiovascular waveform and interval analysis," *Physiological Measurement*, vol. 39, no. 10, p. 105004, 2018.

[14] R. T. Faghih, P. A. Stokes, M.-F. Marin, R. G. Zsido, S. Zorowitz, B. L. Rosenbaum, H. Song, M. R. Milad, D. D. Dougherty, E. N. Eskandar, and R. Barbieri, "Characterization of fear conditioning and fear extinction by analysis of electrodermal activity," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Medicine and Biology Society (EMBC)*, pp. 7814–7818, 2015.

[15] M. R. Amin and R. T. Faghih, "Inferring autonomic nervous system stimulation from hand and foot skin conductance measurements," in *Proc. 52nd Asilomar Conf. Signals, Systems and Computers*, Oct 2018.

[16] M. R. Amin and R. T. Faghih, "Sparse deconvolution of electrodermal activity via continuous-time system identification," *IEEE Trans. Biomed. Eng.*, 2019.

[17] D. S. Wickramasuriya and R. T. Faghih, "Online and offline anger detection via electromyography analysis," in *Proc. IEEE Healthcare Innovations and Point of Care Technologies Conf.*, pp. 52–55, 2017.

[18] X. Deng, R. T. Faghih, R. Barbieri, A. C. Paulk, W. F. Asaad, E. N. Brown, D. D. Dougherty, A. S. Widge, E. N. Eskandar, and U. T. Eden, "Estimating a dynamic state to relate neural spiking activity to behavioral signals during cognitive tasks," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Medicine and Biology Society (EMBC)*, pp. 7808–7813, 2015.

[19] D. S. Wickramasuriya, C. Qi, and R. T. Faghih, "A state-space approach for detecting stress from electrodermal activity," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Medicine and Biology Society (EMBC)*, 2018.