Emotion Recognition Based on DEAP Database Physiological Signals

Tamara Stajić
University of Belgrade, School of Electrical Engineering
Belgrade, Serbia
tasa.stajic@gmail.com

Jelena Jovanović
University of Belgrade, School of Electrical Engineering
Belgrade, Serbia
jelenajovanovic0119@gmail.com

Nebojša Jovanović
University of Belgrade, School of Electrical Engineering
Belgrade, Serbia
nebojsa.php@gmail.com

Milica M. Janković
University of Belgrade, School of Electrical Engineering
Belgrade, Serbia
piperski@etf.rs

2021 29th Telecommunications Forum (TELFOR) | 978-1-6654-2585-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/TELFOR52709.2021.9653286

Abstract—**Recognizing and accurately classifying human emotion is a complex and challenging task. Recently, great attention is paid to the emotion recognition methods using three different approaches: based on non-physiological signals (like speech and facial expression), based on physiological signals or based on hybrid approaches. Non-physiological signals are easily controlled by the individual, so these approaches have downsides in real world applications. In this paper, an approach based on physiological signals which cannot be willingly influenced (electroencephalogram, heartrate, respiration, galvanic skin response, electromyography, body temperature) is presented. Publicly available DEAP database was used for the binary classification (high vs. low) considering four frequently used emotional parameters (arousal, valence, liking and dominance). We have extracted 1490 features from the dataset, reduced to less than 15% (200 most significant features) and applied three different classification approaches – Support Vector Machine, Boosting algorithms and Artificial Neural Networks.**

Keywords—emotion recognition, machine learning, physiological signals, DEAP database

I. Introduction

Emotion is a complex behavioral phenomenon which includes different levels of neural activations and chemical reactions in the human brain [1]. Emotion is a combination of human thought, feeling and behavior, and can be defined as a physiological reaction to different external stimuli [2]. For decades, emotions and emotion recognition have attracted a lot of attention which resulted in a variety of approaches that could be grouped into two distinct categories [2]. First group consists of methods based on non-physiological data such as speech [3] and facial expressions [4]. The advantage of this approach is the fact that the data is easily collected, without the need for any specialized and costly equipment. However, non-physiological signals can be willingly controlled which means that the individual can mask their emotion, and cause uncertainty in the classification that cannot be detected and removed. Second group relies on physiological data such as electroencephalography (EEG) [2], electromyography (EMG) [5], electrocardiography (ECG) [6], galvanic skin response (GSR) [7] etc. This approach allows better correlation with actual emotional state, but at the same time makes it harder to set up the experiment, requires special equipment and subject preparation. Noise inherently present in these signals can also present an obstacle for reliable emotion recognition.

Hybrid approaches imply multimodal methods for emotion recognition that combine non-physiological and physiological approaches. Huang et al. [8] proposed a combination of facial expressions and EEG signals for emotion recognition. Same approach was used by Tan et al. [9]. A Python package for the same task called `MindLink-Eumpy` was introduced by Li et al. [10]. In theory, this allows taking the best of both methods which should result in higher accuracy.

The most common dimensional space used for describing emotions is the arousal/valence space, where emotions are described in terms of the intensity - going from ‘inactive’ to ‘active’ in the arousal dimension, and from ‘unpleasant’ to ‘pleasant’ in the valence dimension [11]. Aside from valence and arousal, other parameters commonly used in literature to present emotions are dominance (ranging from ‘helpless’ to ‘in control’) and liking. These four parameters, alongside familiarity, are also used in the most used open database of physiological signals for emotion classification - DEAP [12].

The reported accuracies in the paper that introduced the DEAP database [12] are 65.1%, 62.7% and 67.7%, and the F1 score 61.8%, 60.8% and 63.4% for arousal, valence and liking, respectively. Torres-Valencia et al. [11] reported the highest accuracy value of 75% for binary arousal classification (high vs. low) when combining EEG, GSR, and ECG signals, and 58.7% accuracy for binary valence classification in case of using only EEG signals (F1 scores was not reported). Yang et al. [13] opted for an approach using a multi-column CNN-based model using EEG signals and reported accuracies of 90% and 90.6% for valence and arousal respectively.

An important aspect of emotion recognition is subjectivity. Emotion itself is a subjective occurrence which makes it difficult to generalize. There are two general approaches regarding this issue – inter-subject and cross-subject. Even though inter-subject classification gives higher accuracy in general, its applicability in real world cases is limited because it requires model recalibration or retraining for each new user, which can be very costly and time consuming. In the study presented in this paper we chose the cross-subject approach because of the higher real-world applicability. The main goal of our

This research was supported by the Ministry for Education, Science and Technology Development of Serbia, Belgrade, Serbia.

978-1-6654-2585-8/21/$31.00 ©2021 IEEE

study is a broad analysis of all available physiological data from the DEAP database, as well as evaluation of different machine learning algorithms for the purpose of emotion recognition.

II. Methodology

A. *The DEAP database*

DEAP database consists of 40 physiological signals from 32 subjects recorded while watching 40 different music videos. After each video, the subjects gave subjective ratings based on which emotions are labelled. The dataset of physiological signals includes: 32-ch electroencephalogram (EEG), 2-ch electrooculogram (EOG), plethysmogram, respiration pattern, 2-ch electromyography (on zygomaticus major muscle, zEMG and trapezius muscle, tEMG), galvanic skin response (GSR) and body temperature. The sampling rate for all data was set to 512 Hz. DEAP database also includes recordings of facial expressions, but that data was not considered in our research.

In this paper, we used the pre-processed data available at https://www.eecs.qmul.ac.uk/mmv/datasets/deap/. The parsed pre-processed and down sampled (by factor 4) data has the dimension 40x40x8064 which correspond to (number of videos) x (number of physiological channels) x (number of samples in one recording).

B. *Signal processing and feature extraction*

Data analysis was done in the Python programming language (Python Software Foundation, Delaware, USA). Aside from standard libraries used for scientific analysis like `NumPy` [14] and `SciPy` [15], we used the `pyphysio` library [16] for signal processing and feature extraction. The complete project code and further information is available at the following GitHub repository: https://github.com/nebojsa55/EmotionRecognition.

A review of all extracted features (extracted from 8064 samples for each subject and for each video) is given in Table 1.

The focus of EEG analysis was on statistical features of different frequency bands such as alpha (8-12 Hz), beta (13-30 Hz), gamma (30+Hz), and theta (4-8 Hz). Other EEG features included power spectral density (PSD) in different bands and Hjorth features (activity, mobility, and complexity) [17]. The resulting set consisted of 44 features for each of the 32 EEG channels.

Respiratory signal features were extracted in the same way as features for heart rate variability - using the `hrvanalysis` [18] Python library. Before the analysis, the signal was filtered using a low-pass Butterworth filter (order 2, f_{low}=32 Hz).

Galvanic skin response (GSR) signal has two basic components – DC component which represents general activity of the sweat glands, and skin conductance response (SCR) component that is a good indicator of arousal level due to external sensory and cognitive stimuli [19]. A low-frequency drift was extracted from the GSR signal by applying a Moving Average (MA) filter, which was then subtracted from the GSR signal. This way the SCR component was singled out and additionally filtered by low pass (LF) fir filter (f_{low}=0.2 Hz) to obtain LF SCR signal and by very low pass (VLF) fir filter (f_{low}=0.08 Hz) to obtain VLF SCR.

TABLE I. Extracted features, PSD – power spectral density, STD– standard deviation, VLF – very low frequency, LF – low frequency, HF – high frequency, RMS – root mean square, SCR – skin conductance response

Signal	**Features**	**No. of features Total=1490**
EEG [20]	4 features for raw and 4 features for normalized to range (0,1) EEG signals: **Mean, Standard deviation, Mean of first derivative, Mean of second derivative**	32 ch x 8 features
	Hjorth features (Activity, Mobility, Complexity)	32 ch x 3 features
	For alpha, beta and theta band: **PSD** 4 features for raw and 4 features for normalized to range (0,1) EEG signals for alpha, beta and theta band: **Mean, Standard deviation, Mean of first derivative, Mean of second derivative**	32 ch x 3 bands x 9 features
	For alpha, beta and gamma band: **Energy, Recursive energy efficiency**	32 ch x 3 bands x 2 features
HRV [12, 21]	**First derivative, Mean arc-length, RMS, Area-perimeter ratio, Mean and standard deviation, PSD in LF [0.01, 0.08] Hz, in medium [0.08, 0.15] Hz and HF bands [0.15, 0.5] Hz, PSD ratio between [0.04, 0.15] Hz and HF band**	10 features
Respiration [19]	**Maximum amplitude in frequency spectrum** **Mean spectrum in [0.2, 0.5] Hz** **Maximum amplitude in PSD** **Mean PSD in [0.2, 0.5] Hz**	4 features
	Mean, STD, Median and Range of peak-to-peak intervals, STD of first derivative of intervals Mean, maximum, minimum and STD of breathing rate, Number of intervals larger than 50 and 20 ms and their ratio in total number of intervals, Square root of mean of sum of peak-to-peak intervals, Coefficient of interval change and variation, Total PSD, PSD in very low frequency (VLF) range [0.003, 0.04] Hz, low frequency (LF) range [0.04, 0.15] Hz and high frequency (HF) range [0.15, 40] Hz, LF/HF ratio, Normalized power in LF and HF domain	23 features
GSR [12, 19]	4 features for raw SCR and LF SCR: **Mean value, Standard deviation, Mean of first derivative, Mean of second derivative** 4 features for LF and VLF SCR: **Numbers of peaks in LF SCR, Number of peaks in VLF SCR, Number of peaks ratio in LF and VLF SCR, Mean of amplitude of LF and VLF SCR**	12 features
	Zero-crossing rate for LF SCR and VLF SCR	2 features
EMG [12], [19]	**PSD in [4, 40] Hz for zEMG and tEMG signals**	2 features
	4 features for raw and LF tEMG and 4 features for raw and LF zEMG signals: **Mean value, Standard deviation, Mean value of first derivative, Mean value of second derivative** 3 features for LF and VLF tEMG **Number of peaks in LF tEMG, Number of peaks in VLF tEMG, Number of peaks ratio in LF and VLF tEMG** 3 features for LF and VLF zEMG **Number of peaks in LF zEMG, Number of peaks in VLF zEMG, Number of peaks ratio in LF and VLF zEMG**	22 features
Temperature [12]	**Mean, Standard deviation, First derivative, Minimum, Maximum, PSD in [0, 0.1] Hz, PSD in [0.1, 0.2] Hz**	7 features

EMG features were extracted from raw tEMG and zEMG signals, low pass fir filtered tEMG and zEMG signals (f_{low}=0.3 Hz) and very low pass fir filtered tEMG and zEMG signals (f_{low}=0.08 Hz).

Plethysmography measurements represented the change in blood volume, so this signal could be used to estimate beat-to-beat intervals. Heart rate variability (HRV) signal was calculated using the `hrvanalysis` [18] Python library from beat-to-beat intervals extracted from the plethysmography signal.

C. Class labeling

Subjects gave ratings along multiple parameters after every video watched. Of these, we picked four parameters - valence, arousal, dominance and liking. Values for each parameter were in range [1,9]. The classification problem was considered as binary classification for each of the four previously mentioned parameters. Classes were defined as "0" and "1" which represented low and high parameter value, respectively. In this study, the class distinction boundary was set to 4.5. Classes were imbalanced (values for parameters were more than 60% concentrated in the higher range).

D. Feature informativeness

The final extracted set of features was of too high dimensionality for the size of our available data: 1490 features vs 1280 recordings (32 subjects x 40 videos = 1280), so dimension reduction was a necessary step before further analysis. The final set has been limited to 200 features (less than 15% of full feature dataset). For the reduction procedure, we used the Recursive Feature Elimination (RFE) method [22] and ranking based on the mutual information between the concrete feature and the target variable [23]. This analysis was done for each emotional parameter individually. RFE reduction was done by training the classifier on the complete feature set followed by the estimation of the importance of each feature. Features with the lowest importance were removed and the process was repeated until a previously decided number of features was left. For the RFE method implementation, `sklearn` [24] package was used.

E. Classification

After feature extraction and dimensionality reduction, evaluation of different machine learning algorithms was done.

The first evaluated method was Support Vector Machine (SVM) which works by translating the feature space to a higher dimensionality one, where the data becomes linearly separable. The SVM implementation was realized by `sklearn` package.

Boosting algorithms were considered for their historically good performance on tabular datasets. For this problem, `CatBoost` [25] Python package was chosen. Classifier parameters were selected empirically - loss function was LogLoss with a learning rate of 0.001, a maximum tree depth of 5 and subsampling ratio of 0.8.

The third approach that was performed was Artificial Neural Network (ANN): a three-layer network with a `LeakyReLU` activation function. Overfitting was resolved by using batch normalization and dropout regularization. The `PyTorch` [26] library was used for ANN implementation.

III. Results and Discussion

A. The most informative features

Table 2 shows top 3 features for each emotional parameter ranked by mutual information scores.

TABLE II. Highest ranking features, LF-low frequency, VLF-very low frequency

Parameter	Features with highest mutual information scores
Valence	Standard deviation of normalized EEG on electrode F8 Mean of first derivative of normalized EEG on electrode CP1 in the alpha band Energy in the beta band of electrode CP5
Arousal	Number of peaks in the LF zEMG signal PSD in the theta band on electrode FC5 Zero crossing rate of the LF SCR signal
Dominance	Number of peaks of LF zEMG signal Numbers of peaks ratio for LF SCR and VLF SCR Number of peaks in the LF SCR
Liking	Mean of first derivative of normalized EEG on electrode C4 in the beta band Mean of first derivative of normalized EEG on electrode Cz Numbers of peaks ratio for LF tEMG and VLF tEMG

The highest informational value for valence is in features extracted from EEG signals – particularly F8 electrode (left temporal region) and alpha and beta band on CP1 electrode (parietal region).

The most important feature for arousal is the number of myoresponses in low-passed zEMG signal. Other high-ranking features include the ones extracted from low-passed SCR signal and breathing rate variability. Most important EEG feature is the power spectral density of the theta band on FC5 electrode (right temporal region).

For the dominance parameter, important features are similar to the ones for arousal, with EEG features ranking lower.

The most important features for liking are coming from C4 and CZ electrodes (central region) as well as tEMG.

B. Classification evaluation

Table 3 shows accuracy and F1 scores reached using SVM, CatBoost and ANN methods for each emotional parameter. All shown accuracies and F1 scores are mean values calculated on 10-fold stratified cross-validation (train-test splitting ratio of 90:10).

All three classifiers reached similar scores. Best general results are achieved in the liking category (71.1% accuracy and 82.5% F1 score when using ANN approach), while the valence category performs the worst (66.2% accuracy using the SVM classifier and 78.3% F1 score using ANN). This is aligned with the results in literature which indicate the valence axis is the hardest one to predict [11]. These results outperform the classification described in Koelstra et al. [12] on the same dataset. Valence classification accuracy outperforms the one given by [11] (66% vs. 58%) but for arousal it is lower (67% vs 75%).

TABLE III. CLASSIFIER EVALUATION

	Arousal (%)		**Valence (%)**		**Dominance (%)**		**Liking (%)**	
	Acc	***F1***	***Acc***	***F1***	***Acc***	***F1***	***Acc***	***F1***
SVM	**67.1**	75.4	**66.2**	74.3	**71.2**	79.4	70.9	80.1
Boost	64.2	**78.0**	63.3	77.5	67.2	**80.4**	69.5	82.1
ANN	64.3	77.9	64.6	**78.3**	66.0	79.3	**71.1**	**82.5**

IV. Conclusion

In this paper, we have compared the results of binary classification (high vs low) using different machine learning approaches (SVM, CatBoosting, ANN) in case of four typical emotional parameters (arousal, valence, dominance and liking) on the publicly available DEAP dataset. Large pool of features (1490) was extracted. However, the classification accuracy was less than 72%. This might be due to the general nature of cross-subject emotion classification or an inherent problem within the reliability of data labeling according to the subjective criteria.

Analysis of the ratings given by each subject shows great discrepancies in how the videos used in the experiment are perceived. This might be remedied by conducting an experiment choosing a different set of videos, particularly ones that have low rating deviations. Another option is using a different threshold for distinguishing between high and low values, as scores are mostly concentrated in the higher range.

Feature analysis has shown that EEG signals carry the most information, but other modalities are not to be discarded, especially when doing classification of arousal and dominance. One limitation of this analysis was the small number of subjects, influenced by the complexity and nonconformity of the experiment. Further research improvements could be made by applying the hybrid approach into the analysis, based on physiological and non-physiological (face expression) data.

References

[1] D. B. Lindsley, "Emotion", Handbook of experimental psychology, Oxford, England: Wiley, pp. 473–516, 1951.

[2] M. Li, H. Xu, X. Liu, and S. Lu, "Emotion recognition from multichannel EEG signals using K-nearest neighbor classification", Technology and Health Care, vol. 26, no. S1, pp. 509–519, Jan. 2018.

[3] Y. L. Lin and G. Wei, "Speech emotion recognition based on HMM and SVM", 2005 International Conference on Machine Learning and Cybernetics, vol. 8, pp. 4898–4901, Aug. 2005.

[4] Z. Liu et al, "A facial expression emotion recognition based human-robot interaction system," IEEE/CAA Journal of Automatica Sinica, vol. 4, no. 4, pp. 668–676, Sept. 2017.

[5] S. Jerritta, M. Murugappan, K. Wan, and S. Yaacob, "Emotion recognition from facial EMG signals using higher order statistics and principal component analysis," Journal of the Chinese Institute of Engineers, vol. 37, no. 3, pp. 385–394, Apr. 2014.

[6] Y. L. Hsu, J. S. Wang, W. C. Chiang, and C. H. Hung, "Automatic ECG-Based Emotion Recognition in Music Listening", IEEE Transactions on Affective Computing, vol. 11, no. 1, pp. 85–99, Jan. 2020.

[7] C. K. Lee et al, "Using Neural Network to Recognize Human Emotions from Heart Rate Variability and Skin Resistance", IEEE Engineering in Medicine and Biology 27th Annual Conference , pp. 5523–5525, Jan. 2005.

[8] Y. Huang, J. Yang, P. Liao, and J. Pan, "Fusion of facial expressions and EEG for multimodal emotion recognition", Computational intelligence and neuroscience, vol. 2017, Sept. 2017.

[9] Y. Tan, Z. Sun, F. Duan, J. Solé-Casals, and C. F. Caiafa, "A multimodal emotion recognition method based on facial expressions and electroencephalography", Biomedical Signal Processing and Control, vol. 70, 2021

[10] R. Li et al, "MindLink-Eumpy: An Open-Source Python Toolbox for Multimodal Emotion Recognition", Frontiers in Human Neuroscience, vol. 15, 2021.

[11] C. A. Torres-Valencia, H. F. García-Arias, M. A. Álvarez López, and A. A. Orozco-Gutiérrez, "Comparative analysis of physiological signals and electroencephalogram (EEG) for multimodal emotion recognition using generative models", Signal Processing and Artificial Vision 2014 XIX Symposium on Image, pp. 1–5, Sep. 2014.

[12] S. Koelstra et al, "DEAP: A Database for Emotion Analysis ;Using Physiological Signals", IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 18–31, Jan. 2012.

[13] H. Yang, J. Han, and K. Min, "A Multi-Column CNN Model for Emotion Recognition from EEG Signals", Sensors (Basel, Switzerland), vol. 19, Oct. 2019.

[14] C. R. Haris et al, "Array programming with NumPy", Nature, 585, pp. 357–362, 2020.

[15] P. Virtanen et al, "SciPy 1.0: fundamental algorithms for scientific computing in Python", Nature Methods, vol. 17, no. 3, pp. 261–272, 2020.

[16] A. Bizzego, A. Battisti, G. Gabrieli, G. Esposito, and C. Furlanello, "Pyphysio: A physiological signal processing library for data science approaches in physiology", SoftwareX, vol. 10, Jul. 2019.

[17] B. Hjorth, "EEG analysis based on time domain properties", Electroencephalography and Clinical Neurophysiology, vol. 29, no. 3, pp. 306–310, 1970.

[18] R. Champseix, "Aura-healthcare/hrv-analysis: Package for Heart Rate Variability analysis in Python", Accessed on: Jul. 14, 2021. [Online], Available: https://github.com/Aura-healthcare/hrv-analysis

[19] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 12, pp. 2067–2083, Dec. 2008.

[20] R. Jenke, A. Peer, and M. Buss, "Feature Extraction and Selection for Emotion Recognition from EEG," IEEE Transactions on Affective Computing, vol. 5, no. 3, pp. 327–339, Jul. 2014.

[21] A. Bartolomé-Tomás, R. Sánchez-Reolid, A. Fernández-Sotos, J. M. Latorre, A. Fernández-Caballero, "Arousal Detection in Elderly People from Electrodermal Activity Using Musical Stimuli", Sensors 2020, vol. 20, Aug. 2020.

[22] X. Chen and J. C. Jeong, "Enhanced recursive feature elimination", Sixth International Conference on Machine Learning and Applications, pp. 429–435, 2007.

[23] L. Paninski, "Estimation of Entropy and Mutual Information", Neural Computation, vol. 15, pp. 1191–1253, Jun. 2003.

[24] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python", Journal of Machine Learning Research, vol. 12, no. 85, pp. 2825–2830, 2011.

[25] L. Prokhorenkova, G.Gusev, A. Vorobev, A. Veronika Dorogush, and A. Gulin , "CatBoost: unbiased boosting with categorical features", Advances in Neural Information Processing Systems, vol. 31, 2018.

[26] A. Paszke et al, "PyTorch: An Imperative Style, High-Performance Deep Learning Library", Advances in Neural Information Processing Systems, vol. 32, 2019.