User-independent classification of emotions in a mixed arousal-valence model

Mauro Nascimben[1,2], Thomas Zoëga Ramsøy[1], Luis Emilio Bruni[2]

[1]*Neurons Inc, Taastrup, Denmark*
[2]*Augmented Cognition Lab, Aalborg University Copenhagen, Copenhagen, Denmark*

***Abstract*—In this work we classified EEG features connected with emotions elicited by musical videos. To detect emotions, we used a user-independent approach with data coming from multiple participants in order to test the "peak-end rule". Participant's video ratings were processed to create a mixed valence-arousal labelling. Input features were refined using a combination of feature ranking and data reduction based on intrinsic dimensionality search. Compared to previous literature, our results show that the proposed mixed arousal-valence classification is compatible with previous works applying a distinct arousal or valence classification.**

Keywords—EEG, emotion recognition, human-computer interaction

I. INTRODUCTION

Affective computing is an already stablished area of Human-Computer Interaction (i.e. HCI) [1]. In this field, electroencephalography (EEG) is commonly selected by researchers for measuring emotional states because it has a high temporal resolution, it is not invasive and it is wearable.

Emotion recognition through EEG could be used in various HCI scenarios like entertainment, e-learning, virtual worlds [2], or e-healthcare applications [3]. A person's inner emotional state is a complex summation of different components connected with subjective experience and physiological response to a stimulus. In a general view, emotions are a large category that encloses mood, feelings and affects. In the present study, we analyze emotions under the theoretical framework of the "Circumplex model of affect" [4] that maps human emotions using only two indexes: valence and arousal. Arousal relates to subjective alertness levels, valence is a way to sort a stimulus as pleasant or unpleasant.

In literature, EEG based approaches for emotion classification are defined as either "user dependent" or "user independent". User dependent means that a new model is generated for each user and trained/tested on the same user data. This kind of models reach higher accuracy of detection but lack on generalization. User independent models try to extract features from different subjects. They achieve better generalization, but their classification accuracy is lower when compared to user dependent models.

In this study, we propose an EEG user independent approach to classify emotions. We analyzed data from a public domain dataset extracting labels from self-reported emotional levels together with EEG features. After this step, we ranked the extracted features (feature selection) and classified them using several classifiers. We payed special attention to the class imbalance problem to avoid overfitting and ran cross-validation multiple times in order to compare the classifiers' performance.

II. MATERIALS AND METHODS

A. Dataset

We applied our implementation to the DEAP dataset [5]. EEG data comes from 32 subjects (16 males and 16 females) while watching 40 musical videos lasting 60 seconds. Each subject rated the stimuli in terms of valence and arousal. Every recording contains 32 EEG channels preprocessed with a sampling rate of 128Hz and ocular/electromyographic artifacts already removed. Other preprocessing steps included band-pass filtering in 4-45Hz and re-referencing in common average mode. Preprocessing also included EEG baseline correction using 3 sec of free running EEG recorded before video start. All further analysis was carried out in Matlab environment under academic license.

B. Channel selection and time window

Analysis for emotions with EEG requires the identification of a small subset of electrodes to target relevant features for classification. Previous studies on "affective EEG" [6, 7, 8, 9] report that frontal and parietal lobes are the most informative ones about the emotional states, while the alpha, gamma and beta waves appear to be the most discriminative. Table I was compiled comparing different papers in the attempt to find the mostly used electrodes for emotion detection. We did not use a subject-dependent channel selection in order to maintain a completely user-independent model.

Each signal from the reported channels in Table I was analyzed in the time window between 49- and 59-seconds during video screening (equivalent to an analysis window of 10s). We decided for this approach because during musical videos viewers emotions fluctuate but they could define an objective positive or negative emotional judgement connected with that video at the end of it. In the last ten seconds of each recording, the feelings that the musical video inspired in the audience are conscious and truly aware. This method is in accordance with the "peak-end rule" [10], a psychological theory stating that people don't judge an emotional experience on the total sum or average of every moment of the experience but only during the most intense point and at its end. In this way, we tried to capture the relation between the subjective valence/arousal score and the EEG recorded closer to that moment. The window length also takes in account observations from [11]: on the DEAP dataset, the authors found that the optimal arousal window is between 3 and 12 seconds.

2471-7819/19/$31.00 ©2019 IEEE
DOI 10.1109/BIBE.2019.00086

TABLE I. ELECTRODES SELECTION

Selected channels in each frequency band	
Frequency band	***EEG channels***
Theta (4-7Hz)	'F7','F8','FC2','Fp1','Fp2','Fz','O1','P3', 'P7','P8','T7','T8'
Alpha (8-13Hz)	'F8','Fp1','Fp2','Fz','O2','P4','P7','PO3', 'T7','T8'
Beta (14-30Hz)	'CP5','F8','FC5','Fp1','Fp2','Oz','P8','T7', 'T8'
Gamma (31-45Hz)	'AF4','CP5','F8','FC5','Fp1','Fp2','P8', 'PO4','T7','T8'

C. *Feature extraction: frequency domain*

In frequency domain, EEG researchers usually subdivide the frequency spectrum in bands called theta (4-7Hz), alpha (8-13Hz), beta (14-30Hz) and gamma (31-45Hz). For each electrode in each frequency band (Table I), we extracted the signal power as the area under the power spectral density curve of every second, averaged and normalized the results dividing by the total power in 4-45Hz.

D. *Feature extraction: time domain*

In time domain we normalized EEG signals in amplitude with z-scores and calculated kernel density in one dimension using a Gaussian kernel with optimized bandwidth and the cumulative density function. From the cumulative density function, values at 25%, 50% and 75% percentile were taken. From the kernel density, mean and standard deviation (sigma) were collected. Other time domain parameters included skewness, kurtosis, mean envelope and standard deviation of signals amplitude.

E. *Feature extraction: complexity measures*

In EEG analysis Fractal Dimension (i.e. FD) is applied to determine the chaotic dynamics of the brain [12]. While Higuchi FD algorithm works iteratively over the time, Katz method calculates the sum of the Euclidean distances between successive points in the temporal series, divided by the maximum distance from each point and the initial point. Another complexity measure is the Hurst exponent, which is estimated by breaking the time series into chunks and a rescaled range is calculated on each chunk before averaging over all chunks. Hjorth mobility and complexity are two normalized slope descriptors (NSDs) used since the seventies in EEG analysis [13]. Mobility is the square root of variance of the first derivative of the signal divided by variance of the signal. Complexity compares the signal's similarity to a pure sine wave, where the value converges to 1 if the signal is more similar.

In total, we extracted 4 features in frequency domain, 9 features in time domain and 5 complexity features. In frequency domain, we obtained the relative power for each frequency band averaging the power spectral density over channels of Table I. For the time domain and complexity, we averaged data of the four different electrode subsets proposed in Table I, gathering all the results in two matrices of 36 and 20 features respectively. In each matrix we ranked all features independently using the absolute value two-sample t-test (e.g. |t|) with pooled variance estimate, using the cut-off value of 1.96 for feature exclusion. In this way, we avoided to explicit pre-select features, automatically choosing inputs among those with higher rank. The final feature matrix comprehended 16 features.

F. *Creation of labels from subjective data*

All subjects rated the musical videos in terms of valence, arousal. Valence and arousal scores were assigned directly after each trial on continuous scales with values between 1 and 9 (1 meaning "negative" valence/unpleasantness or calm/bored for arousal and 9 meaning "positive" valence/pleasantness or excited/engaged for arousal). We scaled data by standard deviation to bring the subjective ratings to a comparable metric. This transformation scales absolute values to relative scores that reflect each answer's rank in comparison to the ranks of all responses in that sample. Figure 1 reports the standardized ratings in terms of arousal and valence for each musical video.

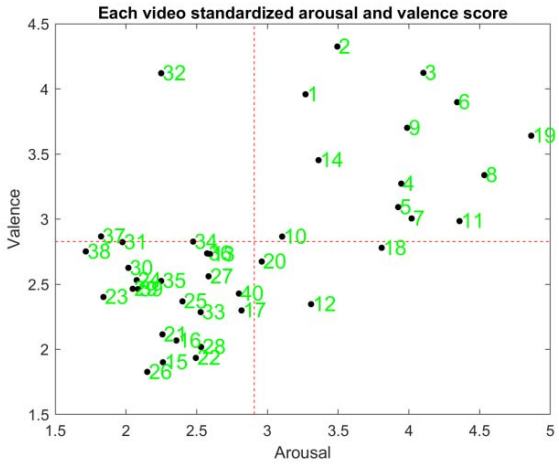

Fig.1. Standardized scores of arousal and valence for each musical video. Red dashed lines are mean values of Valence and Arousal

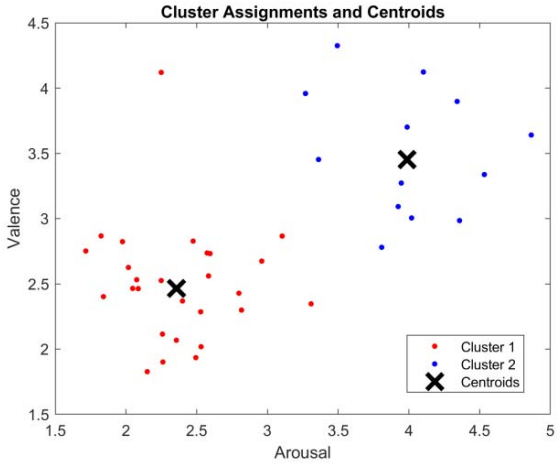

Fig. 2. Unsupervised clustering of video ratings

To subdivide the dataset in groups and to assign a label for each video, we used k-mean as unsupervised method to determine the classes. We decided for a division in two classes because the data appears to be naturally grouped in two clusters (Fig. 1). Five attempts have been made with k-

means++ algorithm and different distance initialization. The best sum of distances between points and centroids was selected (Fig.2). Except video number 32, all other videos appear to be part of a "High Arousal-High Valence" (label "2") or a "Low Arousal-Low Valence" (label "1") group. Another reason we preferred binary classification was the reduction of additional class imbalance problems.

III. RESULTS

The dataset comprises 32 subjects and 40 videos. Consequently, the feature matrix size resulted in 1280 samples by 16 columns. All values in each column were normalized in range from zero to one.

A common problem to be addressed in machine learning is class imbalance, i.e.: when the total number of one class is far less than the number of samples of the other. The major part of machine learning algorithms perform better when classes are nearly equal. Common solutions for class imbalance are through sampling or applying a cost function. Cost function approaches introduce a penalization parameter for instances of the bigger class. Sampling based methods delete instances from the over-represented class or add copies of instances to the under-represented class. In the present study, class "1" had 864 instances and class"2" 416. We solved the class imbalance problem using under-sampling: 416 randomly selected samples from Class 1 were collected and used to build a new dataset together with all instances of Class 2.

A. Dimensionality reduction

Before classification, the feature matrix was inspected with a dimensionality reduction technique to eliminate redundant information. Initially, we investigated dimensionality with principal component analysis (i.e. PCA). PCA eigenvalues estimated that 6 dimensions could summarize adequately the actual feature matrix (the percentage of the total variance explained by 6 principal components was 91.8289%). However, discriminative information in the data might not be necessarily captured by components with largest variance. We preferred to detect the intrinsic dimensionality (i.e. ID) of the dataset that could be estimated using geometric methods like maximum likelihood estimation between neighboring points as reported in [14]. An issue with this kind of ID estimator based in maximum likelihood is that it can underestimate the correct ID in certain situations. In fact, for our dataset, the estimated ID was zero. To overcome this problem, we selected a more robust algorithm as proposed in [15] that computes the ID considering both the normalized nearest neighbor distances and the angles computed on couples of neighboring points. With this method, the estimated ID was 1. We also tested our feature matrix with an improved algorithm for maximum likelihood estimation [16] and it returned 1 ID for our dataset. Given one as target dimension of our dataset, we applied different dimensionality reduction techniques and selected the outputs helpful for class separability. Twelve dimensionality reduction techniques were tested with one as target dimension: linear discriminant analysis, generalized discriminant analysis, stochastic neighbor embedding, stochastic proximity embedding, deep autoencoders (using denoising autoencoder pretraining), t-distributed stochastic neighbor embedding, laplacian eigenmaps, neighborhood preserving embedding, classical multidimensional scaling, neighborhood components analysis, linearity preserving projection, landmark isomap and diffusion maps. These algorithms have been coded following the instructions found in [17]. We run a two-sample t-test with pooled variance estimate at significance level of 0.05 to determine the dimensionality reduction methods able to enhance classification. Statistical test returned a significant absolute value ($|t|>1.96$) for 1D vectors obtained by generalized discriminant analysis (gaussian kernel), linear discriminant analysis and neighborhood components analysis. Normality of data distribution for all vectors was ensured by Jarque-Bera test as prerequisite for running t-tests. The new feature space comprising vectors from dimensionality reduction has 832 samples and 3 columns.

B. Classification

The three-dimensional feature space was the input to cross-validate different models: growing a single classification tree, an ensemble of 100 classification trees using bootstrap aggregating, support vector machine with linear kernel, k-nearest neighbor classifier. All these classifiers had their hyperparameters tuned with iterative search. Cross-validation was computed in 10 K-fold fashion with data partition in 90% for training and 10% for validation (Table II). We also included a feedforward neural network based with Bayesian regularization (two layers of 10 and 5 neurons, max learning epochs 5000, learning rate 0.01) and an ensemble of 100 neural networks with the same characteristics. The average accuracy of each run of cross-validation is displayed in Table III.

TABLE II. CROSS VALIDATION RESULTS

Results of each run of 10-Fold Cross Validation (accuracy mean±std %)					
Classifier	1st run	2nd run	3rd run	4th run	5th run
Classification tree	61.30±5.28	62.03±6.30	62.28±7.58	62.99±3.63	61.54±4.44
Ensemble of trees	58.52±8.40	60.34±4.11	60.09±6.73	57.33±5.69	60.69±4.15
Support vector machines	64.55±4.19	64.04±3.45	64.32±4.62	64.67±3.38	63.48±5.43
k-Nearest Neighbor	64.17±6.52	64.30±4.65	64.06±6.18	64.18±5.24	64.30±3.33
Feedforward NN	55.48±0.83	71.31±1.73	69.52±0.61	54.17±1.16	68.69±1.13
Ensemble 100 Feedforward NN	63.12±2.19	67.08±2.69	68.51±2.06	64.64±1.41	63.50±1.51

TABLE III. PROPOSED MODELS COMPARISON

10-Fold Cross Validation results (average of 5 runs)		
Classifier	***CV Accuracy***	***Standard deviation***
Classification tree	62.08%	±0.67%
Ensemble of trees	59.39%	±1.42%
Support vector machines	64.21%	±0.47%
k-Nearest Neighbor	64.20%	±0.10%
Feedforward NN	63.91%	±8.38%
Ensemble 100 Feedforward NN	65.37%	±2.33%

Feedforward neural network despite an acceptable accuracy has higher variability compared to the other machine learning methods. For example, if we only cross-validated one time the feedforward NN, we could find an accuracy of 71.31±1.73% (2nd CV run) but this value doesn't appear to be realistic. In order to report a fair comparison between methodologies and to avoid overfitting we preferred judging each classifier by the mean cross validation accuracy of all five runs. Performance is higher in case of an ensemble of neural networks: cross-validation accuracy is around 65.37% considering all five runs. Methods with stable outcomes across cross-validation runs are support vector machines and k-nearest neighbor.

C. Comparison with previous works

We compared our outcomes to previous works on emotion recognition. We selected those that apply a user-independent approach with similar stimuli (musical videos): heterogeneity of stimuli applied in different articles reduces the number of possible comparisons. Another difference between some papers found in the literature and the present study is the lack of standardization in subjective scores of valence and arousal. A further difference between articles is the method used to represent emotions. There are two mainstream viewpoints to differentiate between emotions: "categorical" or "dimensional" perspectives of emotions. Categorical representation of emotions divides feelings in six basic emotions: anger, disgust, fear, happiness, sadness, and surprise. Dimensional models of emotions use three dimensions: valence, arousal, and dominance. In the present paper we applied the model called "Circumplex Model of Affect", which only uses valence and arousal. For this reason, we looked for papers applying the same perspective as we did in the DEAP dataset.

- In [18] authors analyze emotions elicited by presentation of DEAP dataset. They achieved the average accuracies of 57.6% and 62% for two classes of valence (high/low) and arousal (high/low).
- Authors of [18] used DEAP dataset in their study without standardization of valence and arousal subjective scores. They used a separate two-class classification for valence (high/low) and arousal (high/low): in valence they achieved 61.17% ± 4.18% while on arousal detection they reported 64.84% ± 9.56%. Their results are obtained with a single cross-validation run with 80% and 20% data division.
- In [19] the proposed methodology of emotion classification uses a mixed model of arousal and valence dividing the valence-arousal plane in quadrants. The authors selected DEAP videos corresponding to the high valence-high arousal (i.e. HVHA) and those of the low valence-low arousal (i.e. LALV) quadrants (two classes in total). They trained a Multilayer Perceptron with one hidden layer on 30 subjects and tested the model on the remaining two. Test set accuracy resulted is 58.5%.
- Authors of [20] applied a Deep Belief Network to distinct arousal (high/low) and valence (high/low) scores of the DEAP dataset. They trained the model on 31 participants and tested on the last one in a "leave-one-out" fashion. This methodology suffers of large discrepancy in the classification outcomes as shown by the standard deviation of the test set accuracy (69.84±11.69% for arousal and 66.88±11.22% for valence). In the same paper, the authors also examined the DEAP data with a SVM classifier including a radial basis function as kernel and pre-selecting significant features with Analysis of Variance. They reached test set accuracy of 56.72±13.45% on arousal and 55.08±13.19% on valence.

IV. CONCLUSIONS

In the present paper, we report an user-independent method applied on the DEAP dataset that uses standardized subjective emotional scores to create a series of labels summing up arousal and valence elicited by musical videos. We extracted 16 features from EEG (time domain, frequency domain and complexity measures) and we reduced the dimensionality of the input dataset to avoid redundant data using an intrinsic dimensionality search. Our results on mixed arousal-valence classification are compatible with those on distinct arousal or valence classification found in previous literature. Comparisons with previous literature were limited to the works that used the same dataset. In our results we report different cross-validation runs: an ensemble of 100 NN reached the peak accuracy but SVM and k-NN showed more stable predictive ability over time.

REFERENCES

[1] Picard, R. W. (2000). Affective computing. MIT press.

[2] C. Hondrou and G. Caridakis, "Affective, natural interaction using EEG: sensors, application and future directions," in SETN, 2012, pp. 331–338.

[3] M. Ali, A. H. Mosa, F. Al Machot, and K. Kyamakya, "EEG-based emotion recognition approach for e-healthcare applications," in ICUFN, 2016, pp. 946–950.

[4] J. Posner, J. a. Russell, and B. S. Peterson, "The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology," Development and Psychopathology, pp. 715–734, 2005.

[5] Koelstra S, Muhl C, Soleymani M, Lee JS, Yazdani A, Ebrahimi T, et al. IEEE transactions on affective computing. DEAP: A database for emotion analysis. Using Physiol Signals. 2012; 3:18–31.

[6] Alarcao, S. M., & Fonseca, M. J. (2017). Emotions recognition using EEG signals: a survey. IEEE Transactions on Affective Computing.

[7] Li, M., & Lu, B. L. (2009, September). Emotion classification based on gamma-band EEG. In 2009 Annual International Conference of the IEEE Engineering in medicine and biology society (pp. 1223-1226). IEEE.

[8] Murugappan, M., Ramachandran, N., & Sazali, Y. (2010). Classification of human emotion from EEG using discrete wavelet transform. Journal of biomedical science and engineering, 3(04), 390.

[9] Lin, Y. P., Wang, C. H., Jung, T. P., Wu, T. L., Jeng, S. K., Duann, J. R., & Chen, J. H. (2010). EEG-based emotion recognition in music listening. IEEE Transactions on Biomedical Engineering, 57(7), 1798-1806.

[10] Kahneman, D. (2000). Evaluation by moments: Past and future. Choices, values, and frames, 693-708.

[11] Candra, H. et al., Investigation of window size in classification of EEG-emotion signal with wavelet entropy and support vector machine. In Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society, Milan, Italy, 25–29 August 2015; pp. 7250–7253.

[12] Lutzenberger, W., Preissl, H., & Pulvermüller, F. (1995). Fractal dimension of electroencephalographic time series and underlying brain processes. Biological Cybernetics, 73(5), 477-482.

[13] Hjorth, Bo; Elema-Schönander, AB (1970). "EEG analysis based on time domain properties". Electroencephalography and Clinical Neurophysiology. 29: 306–310.

[14] E. Levina and P.J. Bickel. Maximum likelihood estimation of intrinsic dimension. Proc. of NIPS 17, 1:777–784, 2005

[15] Ceruti, C., Bassis, S., Rozza, A., Lombardi, G., Casiraghi, E., & Campadelli, P. (2014). Danco: An intrinsic dimensionality estimator exploiting angle and norm concentration. Pattern recognition, 47(8), 2569-2581.

[16] Campadelli, P., Ceruti, C., Casiraghi, E., Lombardi, G., and Rozza, A.: Minimum neighbor distance estimators of intrinsic dimension. In D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis (editors), European Conference on Machine learning and Knowledge Discovery in Databases, Proceedings, Part II, pp. 374–389. Springer Berlin Heidelberg, 2011

[17] L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik. Dimensionality Reduction: A Comparative Review. Tilburg University Technical Report, TiCC-TR 2009-005, 2009.

[18] N. Kumar, K. Khaund, and S. M. Hazarika, "Bispectral Analysis of EEG for Emotion Recognition," Procedia Computer Science, vol. 84, pp. 31–35, 2016.

[19] Pandey, P., & Seeja, K. R. (2019). Emotional state recognition with eeg signals using subject independent approach. In Data Science and Big Data Analytics (pp. 117-124). Springer, Singapore.

[20] Xu, H., & Plataniotis, K. N. (2016, July). EEG-based affect states classification using deep belief networks. In 2016 Digital Media Industry & Academic Forum (DMIAF) (pp. 148-153). IEEE.