Neural Network Based Multimodal Emotion Estimation

A study on Modalities used in Autonomous Systems

Strahil Sokolov
University of Telecommunications and Post
Faculty of Telecommunications, Department of Information Technologies
Sofia, Bulgaria
e-mail: strahil.sokolov@gmail.com

***Abstract*—This paper presents recent research on approaches in autonomous systems for combining multiple modalities for emotion estimation based on neural networks. Emotion recognition is an active area of research especially in the field of autonomous systems in automotive industry, Human Computer Interaction (HCI), multimedia indexing and video surveillance. Both invasive and non-invasive acquisition methods are part of this study. ECG and EEG, Voice and Facial Analysis with modern approaches such as Deep Neural Networks are presented.**

Keywords-deep learning, emotion analysis, multimodal biometrics.

I. INTRODUCTION

Estimating human emotion in Autonomous Systems is an active field of research nowadays. Autonomous Systems in the automotive industry are becoming more and more popular because of the self-driving vehicles. The research in this field enables researchers to seek combination of different biometric modalities through Human-Computer Interface (HCI). HCI is considered the main carrier of human intelligence for the digital world of Autonomous Systems [1]. For the purpose of this research the two main types of acquisition techniques will be considered: invasive and non-invasive. Below are the two modality groups and the respectful modalities which belong to them:

- invasive: EEG, ECG
- non-invasive: face, voice

Invasive HCI depends on the interaction of the following adaptive components: the user generating brain signals with their encoded intent and the interface system translating the signals into instructions to complete the user’s intent. The effect of this is that both the user and the sensing autonomous system must have the ability adapt to and complete each other. The encoded user intent in signal features that the interface can measure. The HCI measures these features and translates them into system commands.

In [2] the authors propose a complete end-to-end multimodal emotion estimation system based on voice and facial analysis with deep neural networks. It is proposed to capture the emotional content for various styles of speaking, via robust features. The authors are utilizing convolutional neural network (CNN) to extract features from the speech; the visual modality a deep residual network of 50 layers is used. Then, long short-term memory (LSTM) networks are utilized to make the model more invariant to outliers. The system is trained by taking advantage of the existing correlations of each of the managed streams streams to improve the approach based on auditory and visual handcrafted features for the prediction of spontaneous and natural emotions on the RECOLA database of the AVEC 2016 research challenge on emotion recognition. The reported combined performance for valence is ~0.71 and for arousal ~0.61.

In [3] the authors present a new emoF-BVP database of multimodal (face, body gesture, voice and physiological signals) recordings of actors enacting various expressions of emotions. The database contains audio and video sequences of actors enacting three different intensities of expressions of 23 different emotions along with facial feature tracking, skeletal tracking and the corresponding physiological data. Four deep belief network (DBN) models are proposed and show that these models generate robust multimodal features for emotion classification in an unsupervised manner. The experimental results show that the DBN models are promising for emotion recognition. A convolutional deep belief network (CDBN) models are proposed to learn salient multimodal features of expressions of emotions. The CDBN models have been evaluated on most modern emotion recognition databases and the accuracy reported ranges from 58,5 to 97.3 per cent.

The HCI in this research uses Electroencephalography (EEG) activity or other electrophysiological measures of brain functions as new non-muscular channels for communication and control with smart devices such as wearables. The research aims developing for improvement of autonomous systems in smart mobile applications, based on processing of recorded electrophysiological signals at execution of different mental tasks [7], [8], [9].

Estimation of the basic emotional states is a basic component of intelligent interface of an autonomous systems. The recorded brain signals with experimental setup for two basic emotional states after noise filtering are estimated by clustering and classification using Convolutional Deep Neural Network (DNN), Neural Network (CNN), and statistical features. Classification is performed with Support Vector Machines (SVM). Since the human emotions are modelled as combinations from physiological elements such as arousal, valence, dominance, liking, etc., these quantities are the classifier's outputs. In this research the recorded with experimental setup electrophysiological signals for two emotional states after noise filtering are estimated on the base of clustering and classification with Deep Convolutional Neural Network.

Neural networks have been around for the most part of our era and during the past few years they have been rediscovered. Not only do they solve quite a few computer vision challenges ranging from face recognition to face obfuscation [12] and further to facial emotion recognition as well as challenges in other areas [13].

The authors [14] discuss an approach for emotion recognition "in the wild" based on combination of deep neural networks and Bayesian classifiers. The neural network were used in bottom-up approach, analyzing emotions expressed by isolated faces. The Bayesian classifier estimates a global emotion integrating top-down features obtained through a scene descriptor. According to the validations on dataset released for the Emotion Recognition in the Wild Challenge 2017 the method has been reported to achieve an accuracy of 64.68% on the test set, where the 53.62% was the competition baseline.

The rest of the paper is organized as follows: In Section 2, human emotion model is described. Section 3 presents a proposed schema for EEG acquisition and processing. In Section 4 facial features extraction and processing is presented. At the end of the paper there are the Conclusion and Acknowledgements.

II. Two-Dimensional Human Emotion Model

The human emotion is a highly subjective phenomenon: it has been accepted by psychologists that multiple dimensions or scales can be used to categorize emotions. The two–dimensional model of emotion, shown in Figure 1 is introduced in [10]. The valence axis represents the quality of an emotion ranging from unpleasant to pleasant. The arousal axis refers to the quantitative activation level ranging from calm to excited state. This approach for recognition of EEG-based emotions use time and frequency features. The time features are in fact some statistical quantities such as means and standard deviations of the raw signals and its first and second derivatives as well. The classified emotions are: joy, relax, sad and fear.

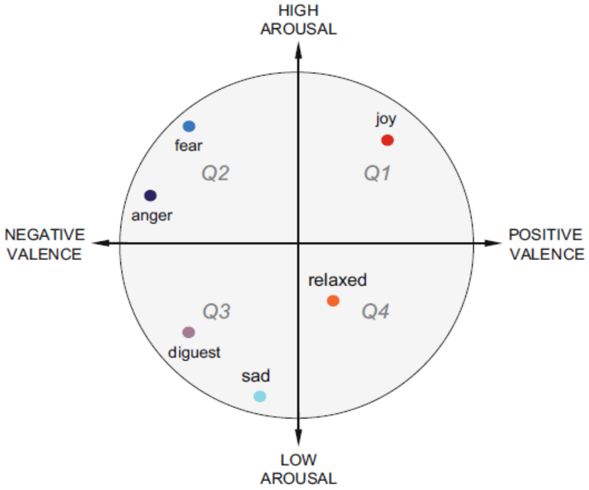

Figure 1: Two–dimensional human emotion model

In [11] the authors use the EEG signal to classify two basic emotions: happiness and sadness. These emotions are evoked by showing subjects pictures. The authors propose a frequency band searching method to choose an optimal band. For this frequency band the recorded EEG signal is filtered. The classification of these two emotions is realized with linear Support Vector Machine (SVM) and Common Spatial Patterns (CSP). Two kinds of trials was used with lengths of 3*s* and 1*s*. Classification accuracies on 10 subjects of 93.5% was achieved for 3 *s* trails and 93% was achieved for 1*s* trials. Their experimental results show that the gamma band (10*Hz* to 30 *Hz*) is suitable for EEG-based emotion classification.

Another technique is shown in [10], where the reported accuracy of the emotional valence is about 71%. This technique relies on changes in the power spectrum of short-time stationary oscillatory EEG processes within the standard EEG frequency bands. The features are extracted from very limited set of electrodes and the dimensionality is further reduced with Principal Component Analysis (PCA). The performance is evaluated for arousal, valence and modality separately. With highest classification accuracy over 90% is arousal.

III. EEG Processing and Emotion Estimation

The research conducted on EEG preprocessing and Emotion recognition is described in more detail in previous work[15].

For EEG processing the spherical spline method was applied, where the human head is modelled as a sphere. The parameter for spline flexibility is set to its default value of 4. The process of analysis of recorded EEG data is connected with the spectral power of the signal in a set of following standard frequency bands: θ (theta - frequency range from 4 *Hz* to 7*Hz*), α (alpha - 8 *Hz* to 13 *Hz*), β (beta-low - 14 *Hz* to 29 *Hz*) and γ (gamma - 30 *Hz* to 45*Hz*).

Feature Vectors need to be calculated as follows: for a given EEG channel *c* and band *b* a feature vector is composed as follows:

$$\mathbf{f}_{c,b} = [Act_{c,b}, Mob_{c,b}, Cpl_{c,b}] \tag{1}$$

where

$$Act_{c,b} = var\left(j_{c,b}\left(t\right)\right) \tag{2}$$

$$Mob_{c,d} = \sqrt{\frac{\mathrm{var}\left(\frac{dj_{c,b}}{dt}\right)}{Act_{c,b}}} \tag{3}$$

$$Cpl_{c,b} = \sqrt{\frac{\mathrm{var}\left(\frac{d^2 j_{c,b}}{dt^2}\right)}{\mathrm{var}\left(\frac{dj_{c,b}}{dt}\right)} - Mob_{c,b}^2} \;. \tag{4}$$

Hjorth parameters, known as activity, mobility and complexity are correspondingly (2), (3) and (4). Arousal estimation the feature vector is organized as augmentation of $\mathbf{f}_{CP6,\theta}$, $\mathbf{f}_{Cz,\alpha}$, $\mathbf{f}_{FC2,\beta}$ and $\frac{Act_{Fz,\beta}}{Act_{Fz,\alpha}}$. Valence estimation is via the feature vector that consists of $\mathbf{f}_{Oz,\theta}$, $\mathbf{f}_{PO4,\alpha}$, $\mathbf{f}_{CP1,\beta}$, $\mathbf{f}_{FC6,\gamma}$, $\mathbf{f}_{Oz,\beta}$, $\mathbf{f}_{Cz,\beta}$, $\mathbf{f}_{T8,\gamma}$ and $\mathbf{f}_{FC6,\gamma}$.

For features selection was chosen the Minimum Redundancy and Maximum Relevance (mRMR) criterion, presented in [15]. The relevance RL of the set of selected features $F = \{f_1, f_2,..\}$ and target classes C was defined as:

$$RL = \frac{1}{|F|} \sum_{f_i \in F} I(f_i, C) \tag{5}$$

where I denotes the mutual information. The redundancy RD of the features was defined as follows:

$$RD = \frac{1}{|F|^2} \sum_{f_i f_j \in F} I(f_i, f_j) \tag{6}$$

For incremental search max[I(F, C)] is equivalent to max[RL(F, C) – RD(F).]

The features selection from all possible sets is mRMR selection of ratios $\frac{Act_{c,b}}{Act_{k,b}}, \quad c \neq k$ and $\frac{Mob_{c,b}}{Mob_{k,b}}, \quad c \neq k$ where c and k denote the EEG channel and b is the activity (θ, α or β).

With the acquired EEG signals was formed sequence of Multidimensional Feature Images (FMI). For this purpose the raw EEG signals are measured or extracted from DEAP [8], where for each subject has 40 trails and each trail includes the EEG signals of 32 channels with duration of 60 *s*. The next step is extraction of power spectrum density as EEG frequency domain feature.

In the current research we join the rapid cascaded classifier with the accurate monolithic one within the two-level combined cascade of classifiers instead of using them independently. This is realized in order to achieve higher detection and lower false alarm rates. The proposed approach for face detection and validation is based on our previous research. It utilizes the OpenCV face detection algorithm [6]. The two-level cascade of classifiers is called “combined” since it combines different types of classifiers: the first level is represented by the Haar-like features’ cascade of weak classifiers, which is responsible for the face-like objects detection, and the second level is a CNN for the objects’ verification.

For classification was used two different Neural Models. The first is Deep Neural Network comprising of 4 Neural layers. The model contains an initial neural layer of 4000 nodes, followed by layers of 400 and 800 neurons, before the output neural layer of 3 nodes. These nodes are inputs for CNN. All layers are fully connected with Softmax [4] acting as the Activator, and use Dropout [5] technique. The second is a Convolutional Neural Network model designed to classify images effectively. The model uses 2 Convolutional layers.

In this phase the fairly fast face detector is also able to deliver faces in frontal pose. This depends on the training set of images for the CNN. In our approach this has proven to be useful since we are using short-length videos of the subjects’ faces may have slight fluctuations off the frontal pose.

The EEG signals were downsampled to 128 *Hz*, bandpass filtered (4 *Hz* to 45 *Hz*). The all data was averaged to the common reference. The performance is validated and evaluated using the k-fold technique. The testing part is extracted from the whole dataset. The rest of the dataset is used to train the classifier. This procedure repeats 20 times and the accuracy is calculated as an average of the accuracies in the iterations.

The Deep Neural Model achieves accuracies for Valence of 73.48% and accuracies for Arousal of 70.35% respectively for classification on high and low classes. For classification on 3 classes - high, normal and low the achieved accuracy is 54.51% for Valence and 51.63% for Arousal.

The Arousal and Valence scores in dataset are given as fractional numbers ranging from 1 to 9. We have quantized these scores to 3, 5 and 7 levels and the testing was performed for each case. The calculated classification accuracies versus dimensionality of the feature vectors is seen on Figure 2.

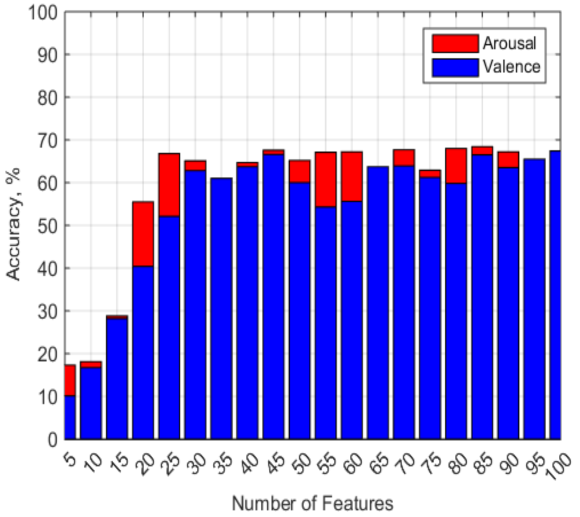

Figure 2: Classification accuracies for arousal and valence versus dimensionality of the feature vectors for CNN

Face emotion estimation runs in parallel and contributes to the improvement of the scores generated by the EEG analysis module. The previously proposed DNN-based face emotion estimation provides an additional improvement for the EEG signal analysis. More experiments need to be conducted on DEAP and other datasets for facial emotion recognition with other DNN structures.

IV. CONCLUSION

An approach for automated multimodal EEG and face-based estimation of human emotions was presented. Estimation of the basic emotional states is a step of building up HCI for disabled persons. The recorded brain signals with experimental setup for two basic emotional states after noise filtering are estimated on the base of clustering and classification with Deep Neural Network, Convolutional Neural Network and statistical features. Classification is performed with support vector machines. Since the human emotions are modelled as combinations from physiological elements such as arousal, valence, dominance, liking, etc., these quantities are the classifier's outputs.

The improvement of the classification is delivered via parallel face emotion analysis system based on DNN. In the ongoing research we are building on top of the achieved framework by experimenting with various Deep Neural Network types and adding other modalities like the ear. The research will be presented in an upcoming study.

REFERENCES

[1] Serbedzija, N. B. "Service components and ensembles: Building blocks for Autonomous Systems." In International Conference on Autonomic and Autonomous Systems (ICAS), Lisbon, Portugal. 2013.

[2] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1301-1309, Dec. 2017.

[3] H. Ranganathan, S. Chakraborty and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, 2016, pp. 1-9.

[4] R. A. Dunne, and N. A. Campbell, ”On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function,” Proc. 8th Aust. Conf. on the Neural Networks, Melbourne, vol. 185, pp. 181-191, 1997.

[5] N. Srivastava, and R. Salakhutdinov, “Multimodal learning with deep boltzmann machines,” Journal of Machine Learning Research, vol. 15, no.1, pp. 2949–2980, 2014.

[6] P. Viola, and M. J. Jones, "Robust real-time face detection." International journal of computer vision,” vol. 57, no. 2 pp. 137-154, 2004.

[7] H. Peng, F. Long, and C. Ding, “Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226–1238, 2005.

[8] S. Koelstra, C. Mühl, 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 Transactions on Affective Computing, vol. 3, no. 1, pp. 18–31, 2012.

[9] X.-W. Wang, D. Nie, and B.-L. Lu, “EEG-Based Emotion Recognition Using Frequency Domain Features and Support Vector Machines,” ICONIP’11 Proc. of the 18th int. conf. on Neural Information Processing, part I, pp. 734–743, 2011.

[10] M. Li, and B.-L. Lu, “Emotion classification based on gamma-band EEG,” Engineering in Medicine and Biology Society, 2009. Proc. of. The Annual Int. Conf. EMBC pp. 1223–1226, 2009.

[11] P. Lahane and M. Thirugnanam, "A novel approach for analyzing human emotions based on electroencephalography (EEG)," 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), Vellore, 2017, pp. 1-6.

[12] Emeršič Z, V. Štruc P. Peer, “Ear Recognition: More Than a Survey”. Neurocomputing. 10.1016/j.neucom.2016.08.139, 2017.

[13] Blaž M, et al. “k-Same-Net: k-Anonymity with Generative Deep Neural Networks for Face Deidentification. Entropy”. vol.20. p.60, 2018.

[14] L. Surace et al., “Emotion Recognition in the Wild using Deep Neural Networks and Bayesian Classifiers”, sep. 2017.

[15] S Sokolov, Y Velchev, S Radeva, D Radev,“ Human emotion estimation from EEG and face using statistical features and SVM” Computer Science & Information Technology, pp37-41, 2017