Late Fusion Approach for Multimodal Emotion Recognition Based on Convolutional and Graph Neural Networks

Tomasz Wierciński
Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Gdansk, Poland
Gdansk, Poland, Country *tomaszwiercinski26@gmail.com*

Teresa Zawadzka
Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology
Gdansk, Poland, Country *tegra@eti.pg.edu.pl*

Abstract

The current trends in automatic emotion recognition encompass the application of deep learning techniques, as, if applied to a multimodal approach, give the most promising results. The study presented in the paper follows this trend - the objective of the research is to propose a deep learning-based solution allowing to recognize emotions in circumplex model with performance metrics on a par with the ones achieved by competitive solutions. The observation channels used are physiological signals i.e. electrocardiography, electroencephalography and electrodermal activity, while the applied technique is late fusion with Graph and Convolutional Neural Networks. The solution is validated for the AMIGOS dataset and the achieved results are comparable to the baseline methods. While already satisfactory, the results still leave a place for further investigations.

Keywords: affective computing, biosignals, neural network, emotion recognition, late fusion

1. Introduction

Affective computing is an area of study encompassing the recognition, processing, and interpretation of human emotions [21, 20]. The subject of this research is the detection and recognition of emotion. It is this area that concerns the creation of Automatic Emotion Recognition (AER) models capable of processing data gathered from a particular person or group of people and predicting the affective state of the individual person, or the group as a whole (in terms of the tone of a conversation between multiple people, for instance) [31]. The process of building such a model is part of a different but overlapping field called Machine Learning (ML). The AER models presented in this paper are artificial neural networks (ANNs) - a set of computational models consisting of interconnected artificial neurons. In particular, the Convolutional Neural Networks (CNN) and Graph Neural Networks (GNN) networks [1] are applied in the presented research.

The creation of an emotion recognition model requires an operationalization of human emotions. Operationalization can be understood as a process of defining measurement phenomena that often are not directly measurable and are often understood as "fuzzy concepts". In the case of emotions, it can be the definition of a set of classes that a given emotion can fall into or a set of continuous dimensions that an emotion can exist within. A handful of major approaches to emotion description can be distinguished based on different theories of emotion expression: the categorical approach, the dimensional approach, and the appraisal-based one [33]. In this research, the dimension approach is applied. It operates under the Dimensional Theory. The three most fundamental dimensions found to best encompass expressed emotions are valence, arousal, and dominance, oftentimes simplified to just valence and arousal. This notation is often referred to as the circumplex model. The appraisal-based approach assumes that emotions result from a complex process of continuous change in various subsystems understood as components forming emotion.

Given an emotion notation model, another issues concerns symptoms, which are used to recognize emotions. The symptoms encompass among others facial expressions, gestures, vocalization and bioelectrical signals. Bioelectrical signals refer to electric signals that can be measured from living tissue. They are a result of internal processes and can be measured via changes of electrical potential [29]. There's a wide range of signals that can be differentiated, including the electroencephalogram (EEG), the electricardiogram (ECG) and the electrodermal activity (EDA) also known as the galvanic skin response (GSR). Changes in these signals represent specific changes in the body and correlates with emotion.

The research goal of this study is to check the accuracy of the multimodal solution, which: (1) recognizes emotions in the two-dimensional model, (2) bases on bioelectrical signals, (3) implements Convolutional and Graph Neural Networks, and (4) incorporates late fusion approach.

The biological signals used in this study, are the ones mostly used in emotion recognition solutions i.e. EDA, ECG, and EEG. In the presented solution for EDA and ECG biosignals the Convolutional Neural Network was used and for EEG the Graph Neural network. The applied late fusion approach [11] assumes that the emotions recognized from each modality (here each separate biosignal) are further combined to achieve the final recognized emotion. This approach is contrary to the early fusion approach, which is a process of combining features from diverse types of modalities for further analysis.

The paper is organized as follows. Section 2 describes the incorporated methodology. Section 3 describes recent trends in automatic emotion recognition. In Section 4, the architectures of networks are presented. The consequent section - Section 5 is devoted to the chosen dataset used in the experiments as well as classification and model settings. The comparison of accuracies for each presented model and comparison of obtained results with baseline methods is done in Section 6. Finally, future works are discussed in Section 7.

2. Methodology

The methodology applied for the research is experiment-based, which means that a set of experiments were designed to find the answer to the research question. The experiments were planned following an approach based on the agile methodologies for Data Science projects [25]. As a consequence the experiments were designed iteratively based on results from previous tasks. Said experiments were therefore not known a'priori, but planned flexibly throughout the entire project. For example in EEG prepossessing the two approaches were examined: the extraction of the list of features and using almost raw signal - the latter was chosen. For the sake of clarity, the paper presents only the experiments that resulted in the architecture, prepossessing or model settings applied in the final model.

The first two experiments (*CNN for EDA* and *CNN for ECG*) concern the design of the single channel models for the EDA and ECG bioelectrical signals. Both of the experiments are an attempt at designing an optimal AER model based on one signal with performance scores comparable to those found in literature that inspired their designs. CNN for EDA involved feature extraction using domain specific methods for the two signals - extraction of features unique to the signals such as the QRS complex and the SCR peaks (a model to represent waveforms observed in ECG and Skin Conductance Response peaks). CNN for ECG involved minimal preprocessing of the data, with feature extraction primarily taking place through the convolutional layers of the neural network. Both experiments used Convolutional Neural Networks with the second experiment expanding on the convolutional layers in terms of kernel size and the amount of said layers. The third experiment (GNN for EEG) pertains to the EEG model trained using the Graph Learning method. The adjacency matrix is learned from training data and its features are extracted through spatial convolution. The neural network architecture is based on the work of Jia et al. [10]. The model has been rewritten to work with the chosen technologies and adjusted for use in emotion recognition - the original work regarded the use of EEG signals for sleep stage classification. The aim of this experiment is the refitting of the original architecture for emotion recognition in order to achieve similar or higher performance scores than other recent models using EEG signals for emotion classification. The fourth experiment (late fusion model) is the final primary experiment concerning the full model. The model joins all three previously trained modalities - EDA, ECG, and EEG - using the late fusion approach.

The presented models of emotion recognition were appraised based on resulting classification metrics. Accuracy and F1 scores were calculated. The model validation was performed using a k-fold method with 10 folds with metrics calculated from left out folds from all iterations. For the late fusion model, additionally, the validation metrics were compered with the first three models i.e. the unimodal approaches. Also, the late fusion model was compared with baseline research, which was chosen as the ones which present multimodal approach and are validated for AMIGOS dataset [18].

3. Related Work

The research in the field of automatic emotion recognition varies with respect to numerous aspects. To the most important ones belong the incorporated emotion model, used symptoms, applied techniques, number of modalities used, and the characteristics o people for which the solution is developed. The number of research in the field is so huge, that the section is organized focusing on the most important differences between research and does not indicate the particular research but the surveys analyzing solutions regarding specific aspects of the research.

Automatic emotion recognition solutions are based on one or more modalities. When only one modality is processed to recognize emotion it is said that the approach is unimodal [30]. Unimodal approaches concern different symptoms. As it was previously mentioned, apart from facial expressions [19] physiological signals are one of the most often used. Among them researchers are very interested in EEG, which reveals in numerous works in the last years [16, 3, 9]. The other two physiological signals commonly used in automatic emotion recognition are ECG [7] and EDA [34]. When more than one modality is used to recognize emotion the approach is called multimodal [6]. In this approach often face expression symptoms are combined with other ones [15].

The other aspect of the emotion recognition methods is the applied technique to recognize emotions. Current research concentrates mainly on deep-learning solutions, as the achieved results are the best for these techniques [12, 5].

The last aspect discussed here is connected with the characteristics of the people, for whom the solution is developed. The research can be divided into two groups. The first group is related to typically developed people and the second one is to people suffering from specific diseases. In the second group the most often diseases which are analysed are autism [14], epilepsy [23], or Huntington's Disease [13].

4. Models' Architectures

This section presents the architectures applied in all four experiments. The CNN for EDA and ECGconsists of a sequence of one-dimensional convolution layers and pooling layers followed by fully connected dense layers for emotion classification. The convolutional layers act as feature extractors and the following dense layers decide the assigned emotion class. The alternating max-pooling layers are present to reduce the overfitting of the model. The use of a CNN enables the use of bioelectrical signals as with minimal preprocessing, allowing for emotion detection without the need for prior feature extraction or expert knowledge. Similar models have been used in previous works for ECG and EDA signals [26, 22], where the CNN detects the patterns of SCR peaks within the EDA signal and R-peaks within the ECG signal.

The Graph Neural Network structure for EEG is based on an existing architecture used for classification of EEG signal into sleep phases. GraphSleepNet [10] makes use of Graph Learning to build pairwise relationships between nodes by optimizing for a given loss function and consequently create an adjacency matrix meant to minimize the value of a given criterion. Furthermore, the provided data was segmented into equal length windows across time, resulting in a series of constructed graphs representing consecutive segments of time. This allowed to use not only spatial or graph convolution based on constructed graph edges but also temporal convolution using corresponding nodes across consecutive graphs. Previous research has been done regarding the networks fitness for emotion recognition. For the purposes of this work the network was rewritten for use with the TensorFlow package in version 2.6.0 and used as a part of a larger network architecture to allow for late fusion of multiple bioelectrical signals (as opposed to the earlier approach using early fusion).

The late fusion model is a combination of the three previously described models (i.e. CNNs for EDA and EEG, GNN for EEG) for individual bioelectrical signals. It combines the three modalities - EEG, EDA and ECG - using late fusion. With all of the previous "partial" models already trained to maximize performance metrics, transfer learning was applied when building the late fusion model.

5. Experimental Settings

In this section, the dataset, classification settings and model settings in our experiments are presented. The dataset selected for use with the model is AMIGOS [26]. The dataset is publically available for research purposes and gatheres bioeletrical data of participants. The dataset also operationalizes emotion using the dimensional approach, making use of the valence and arousal dimensions for annotation of emotion. The annotations were performed externally by multiple annotators. Additionally, the dataset offers personality information. Originally, the processing of the data involved feature extraction for the EDA and ECG data. The data was divided into overlapping segments of constant length using a sliding window technique before having domain-specific features extracted. For EDA these were the rise and decay time, latency, amplitude, half amplitude, and width of detected SCR peaks. The detection of SCR peaks and subsequent feature extraction was done using the pysiology Python package [4].

In the presented research the classification problem is defined as a classification of one of the four quadrants in the dimensional model. In this model, personality-based clustering has been applied. For AMIGOS dataset the limitations for the values of valence and arousal enforced through the used annotation software can be expressed as $[1, 9]$ for both valence and arousal. Therefore, $z_{min} = \{1, 1\}$ and $z_{max} = \{9, 9\}$ resulting in $\tau = \{5, 5\}$. In AMIGOS dataset personality is denoted using the OCEAN notation (Openness to Experience, Conscientiousness, Extraversion, Agreeableness, Neuroticism) [27]. The valence and arousal values refer to the difference between the threshold found the personality-based clustering and the midpoint equal to 5. Of note is the fact that much criticism has been attributed to the trait concept, particularly regarding the ability to extrapolate a person's behavior across various environments, as more steps are being taken to create more testable hypothesis regarding the person-situation interaction [8].

Analysis of correlation reveals the highest correlations existing between pairs of personality traits, such as the correlation of 0.54 between agreeableness and extroversion. From correlation values between emotion annotations and aspects of personality, the highest values are the positive correlation of 0.28 between arousal and conscientiousness and the negative correlation of -0.31 between valence and conscientiousness. All of the other correlation values from this group specifically do not exceed an absolute value of 0.25. In regards to the p-values between emotion annotations and aspects of personality, none have reached the often assumed threshold of significance of $p < 0.05$. The lowest among them is the value of 0.062 between valence and conscientiousness.

5.1. Model settings

The full multimodal neural network could be understood as a joining of three submodels, as described in previous subsections. These submodels - the two CNNs for ECG and EDA signals and the GNN for EEG data - were initially trained separately. For every training session a k-fold split was applied to divide the data into a training and validation set. Stratification was performed during the split to ensure each subset of the data contained an identical proportion of emotion classes. Five equal folds were created and the training process was repeated five times, each time with a different validation set. The accuracy of the model was assessed using all samples in the dataset following this process.

CNNs for EDA and ECG were trained on a single channel, understood as a single time series of values detected by a bioelectrical signal sensor. The ECG signal in the AMIGOS dataset was recorded as two channels - the left and the right ECG channel. For the purposes of the experiment, the left channel was selected (ECGL). Integration of the two channels was considered, however since comparative results shown in the literature have given no significant improvement over the ECGL channel by itself [26], the integration was not applied during the experiment. The kernel sizes of the convolutional layers were different for the ECG and EDA models. The ECG model had kernels of sizes 15, 10, 5, and 1 in the given order. The EDA model had kernels of sizes 10, 3, 1, and 1. The model used for training on the data with previously extract domain-specific features consisted of two instead of three convolution blocks with kernels of sizes 5 and 3 used for temporal convolutions on the series of features extracted from consecutive segments of the signal.

The RMSprop [24] algorithm was used for training GNN for EEG with a learning rate of 0.001 and categorical cross-entropy was used as the loss function. The GNN model used a single Graph module followed by two dense layers of 128 units and an output layer. A 0.5 dropout rate was used. The model was trained for upwards of 200 epochs - an early stopping mechanism was applied based on loss values of the validation set with a generous 20 epochs of patience. The model with the best validation loss was picked as basis for the later full multimodal neural network.

The final - late fusion model was trained in three different configurations. In all configurations a k-fold stratified split was applied, just as it was performed on the individual submodels. The different configurations were assessed based on the volatility of loss on the validation set during training, accuracy of predictions on the all validation sets - calculated on all samples by exploiting the k-fold split - and assessment of the confusion matrix.

The first training configuration to be tested was the training of the complete model from randomized initial weights. The other two configurations are variations on a transfer learning approach - with weights transferred from the partial submodels. The transfer learning approach was tested both with training fully enabled and disabled on the transferred weights - the training occurred only on the last connecting dense layers. The training approach making use of transfer learning with all layers being trained yielded the most promising results out of the three variants. The results of this training are reported in the following sections.

6. Performance Evaluations and Discussion

This section summarizes the research presenting the comparison of the performance of the final - late fusion model with respect to unimodal solutions and baseline methods. The architecture of each of the unimodal models was tuned for maximization of performance metrics, such as F1 score and accuracy scored on a validation set, they were combined into the late fusion model. The graph convolution and temporal convolution layers are operating in parallel on separate bioelectrical signals and their output is concatenated and processed into class probabilities using fully connected layers. In each case the late fusion model has better performance metrics than the unimodal models. The comparison is presented in Table 1.

Table 1. Training results for each of the sub-models and the full emotion classification model

Modality	**Valence**		**Arousal**		**All**	
	Accuracy	**F1-score**	**Accuracy**	**F1-score**	**Accuracy**	**F1-score**
EDA	0.6797	0.6654	0.6103	0.6096	0.4537	0.6383
ECG	0.6993	0.651	0.6199	0.6411	0.4764	0.6474
EEG	0.7128	0.6627	0.6655	0.6972	0.5101	0.6836
Multimodal	0.8221	0.8008	0.7616	0.7774	0.6406	0.7902

The list of performance metrics in order of valence classification accuracy can be seen in Table 2. All of the collated works made use of at least one bioelectrical signal and classified valence and arousal into "high" and "low" classes, resulting in four emotion classes in total - the same as the presented approach. The model achieves comparative results to other multimodal solutions making use of the same bioelectrical signals.

Table 2. Classification accuracy for AMIGOS dataset in literature. Highest accuracy marked in bold.

Reference	**Modality**	**Valence acc.**	**Arousal acc.**
Santamaria-Granados et al. 2019 [26]	ECG, EDA	0.75	0.76
Wang et al. 2018 [32]	EEG, ECG, EDA	0.801	0.684
Chang et al. 2019 [2]	EEG, ECG, EDA	0.832	0.701
Siddharth et al. 2022 [28]	EEG, ECG, EDA	0.8394	**0.8276**
Menon et al. 2022 [17]	EEG, ECG, EDA	**0.871**	0.805
Presented approach	EEG, ECG, EDA	0.8221	0.7616

7. Conclusion

While the performance metrics of the final emotion classification model of this thesis placed it within the range of leading multimodal models from recent years (making use of the same emotion labeling and bioelectrical signals), it is believed that these results could be improved given enough refinement. Other approaches to the fusion of the three signals could be investigated. Currently, the resulting features from each modality are merely concatenated for further processing, however more advanced methods could be investigated, even such experimental methods as the application of another graph learning layer with the assumption of each of the modalities as a node within the graph. Additionally, more attention could be placed on the single-channel models - the ECG and EDA ones. While the results for the individual models were satisfactory, the model architectures could be improved upon based on the more cutting edge approaches to emotion recognition from those signals developed in recent years. Any improvement done in terms of performance to any of the sub-models is believed to yield better results for the full model.

References

1. Naveed Ahmed, Zaher Al Aghbari, and Shini Girija. A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications*, 17:200171, 2023.
2. En Jui Chang, Abbas Rahimi, Luca Benini, and An Yeu Andy Wu. Hyperdimensional computing-based multimodality emotion recognition with physiological signals. *Proceedings 2019 IEEE International Conference on Artificial Intelligence Circuits and Systems, AICAS 2019*, pages 137–141, 3 2019.
3. Didar Dadebayev, Wei Wei Goh, and Ee Xion Tan. Eeg-based emotion recognition: Review of commercial eeg devices and machine learning techniques. *J. King Saud Univ. Comput. Inf. Sci.*, 34(7):4385–4401, jul 2022.
4. Giulio Gabrieli, Atiqah Azhari, and Gianluca Esposito. *PySiology: A Python Package for Physiological Feature Extraction*. Springer, Singapore, 2020.
5. Divya Garg, Gyanendra Verma, and Awadhesh Singh. A review of deep learning based methods for affect analysis using physiological signals. *Multimedia Tools and Applications*, page 46, 01 2023.
6. Xin Gu, Yinghua Shen, and Jie Xu. Multimodal emotion recognition in deep learning:a survey. In *2021 International Conference on Culture-oriented Science Technology (ICCST)*, pages 77–82, 2021.
7. Muhammad Anas Hasnul, Nor Azlina Ab. Aziz, Salem Alelyani, Mohamed Mohana, and Azlan Abd. Aziz. Electrocardiogram-based emotion recognition systems and their applications in healthcare—a review. *Sensors*, 21(15), 2021.
8. Kai T. Horstmann and Matthias Ziegler. Situational perception. *The Wiley Handbook of Personality Assessment*, pages 31–43, 4 2016.
9. Essam Houssein, Asmaa Hamad, and Abdelmgeid Ali. Human emotion recognition from eeg-based brain–computer interface using machine learning: a comprehensive review. *Neural Computing and Applications*, 34, 05 2022.
10. Ziyu Jia, Youfang Lin, Jing Wang, Ronghao Zhou, Xiaojun Ning, Yuanlai He, and Yaoshuai Zhao. Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification. *IJCAI International Joint Conference on Artificial Intelligence*, 2:1324–1330, 7 2020.
11. Ruhina Karani and Sharmishta Desai. Review on multimodal fusion techniques for human emotion recognition. *International Journal of Advanced Computer Science and Applications*, 13, 01 2022.
12. Amjad Rehman Khan. Facial emotion recognition using conventional machine learning and deep learning methods: Current achievements, analysis and remaining challenges. *Information*, 13(6), 2022.
13. Catarina Kordsachia, Izelle Labuschagne, and Julie Stout. Beyond emotion recognition deficits: A theory guided analysis of emotion processing in huntington's disease. *Neuroscience Biobehavioral Reviews*, 73, 11 2016.
14. Agnieszka Landowska, Aleksandra Karpus, Teresa Zawadzka, Ben Robins, Duygun Erol Barkana, Hatice Kose, Tatjana Zorcec, and Nicholas Cummins. Automatic emotion recognition in children with autism: A systematic literature review. *SENSORS*, 22(4), FEB 2022.
15. Sze Chit Leong, Yuk Ming Tang, Chung Hin Lai, and C.K.M. Lee. Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing. *Computer Science Review*, 48:100545, 2023.
16. Xiang Li, Yazhou Zhang, Prayag Tiwari, Dawei Song, Bin Hu, Meihong Yang, Zhigang Zhao, Neeraj Kumar, and Pekka Marttinen. Eeg based emotion recognition: A tutorial and review. *ACM Comput. Surv.*, 55(4), nov 2022.

17. Alisha Menon, Anirudh Natarajan, Reva Agashe, Daniel Sun, Melvin Aristio, Harrison Liew, Yakun Sophia Shao, and Jan M. Rabaey. Efficient emotion recognition using hyperdimensional computing with combinatorial channel encoding and cellular automata. *Brain Informatics*, 9:1–13, 12 2022.
18. Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 12:479–493, 4 2021.
19. Soumya Mohanta and Karan Veer. Trends and challenges of image analysis in facial emotion recognition: a review. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 11, 09 2022.
20. Rosalind W Picard. *Affective computing*. MIT press, 2000.
21. Rosalind W. Picard and Stood Marie Curie. Affective computing. *Affective Computing*, 1997.
22. B. Pyakillya, N. Kazachenko, and N. Mikhailovsky. Deep learning for ecg classification. *Journal of Physics: Conference Series*, 913, 10 2017.
23. Liang Qi, Jing Zhao, PanWen Zhao, Hui Zhang, JianGuo Zhong, PingLei Pan, GenDi Wang, ZhongQuan Yi, and LiLi Xie. Theory of mind and facial emotion recognition in adults with temporal lobe epilepsy: A meta-analysis. *FRONTIERS IN PSYCHIATRY*, 13, OCT 6 2022.
24. Sebastian Ruder. An overview of gradient descent optimization algorithms. *ArXiv*, abs/1609.04747, 2016.
25. Jeffrey S. Saltz and Iva Krasteva. Current approaches for executing big data science projects—a systematic literature review. *PeerJ Computer Science*, 8:e862, February 2022.
26. Luz Santamaria-Granados, Mario Munoz-Organero, Gustavo Ramirez-González, Enas Abdulhay, and N Arunkumar. Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos). *IEEE Access*, 7:57–67, 2019.
27. Ralf Schulze and Richard D. Roberts. Openness conscientiousness extraversion agreeableness neuroticism index condensed. 2018.
28. Siddharth, Tzyy Ping Jung, and Terrence J. Sejnowski. Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing. *IEEE Transactions on Affective Computing*, 13:96–107, 2022.
29. Yogendra Narain Singh, Sanjay Kumar Singh, and Amit Kumar Ray. Bioelectrical signals as emerging biometrics: Issues and challenges. *ISRN Signal Processing*, 2012:1–13, 7 2012.
30. Pragya Singh Tomar, Kirti Mathur, and Ugrasen Suman. Unimodal approaches for emotion recognition: A systematic review. *Cognitive Systems Research*, 77:94–109, 2023.
31. Emmeke A. Veltmeijer, Charlotte Gerritsen, and Koen V. Hindriks. Automatic emotion recognition for groups: A review. *IEEE Transactions on Affective Computing*, 14(1):89–107, 2023.
32. Sheng Hui Wang, Huai Ting Li, En Jui Chang, and An Yeu Andy Wu. Entropy-assisted emotion recognition of valence and arousal using xgboost classifier. *IFIP Advances in Information and Communication Technology*, 519:249–260, 2018.
33. Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, and Wenqiang Zhang. A systematic review on affective computing: Emotion models, databases, and recent advances, 03 2022.
34. Dian Yu and Shouqian Sun. A systematic exploration of deep neural networks for eda-based emotion recognition. *Information*, 11(4), 2020.