2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

EEG Emotion Recognition via Graph-based Spatio-Temporal Attention Neural Networks

Shadi Sartipi[1,*], Mastaneh Torkamani-Azar[2], and Mujdat Cetin[1,3]

Abstract— **Emotion recognition based on electroencephalography (EEG) signals has been receiving significant attention in the domains of affective computing and brain-computer interfaces (BCI). Although several deep learning methods have been proposed dealing with the emotion recognition task, developing methods that effectively extract and use discriminative features is still a challenge. In this work, we propose the novel spatio-temporal attention neural network (STANN) to extract discriminative spatial and temporal features of EEG signals by a parallel structure of multi-column convolutional neural network and attention-based bidirectional long-short term memory. Moreover, we explore the inter-channel relationships of EEG signals via graph signal processing (GSP) tools. Our experimental analysis demonstrates that the proposed network improves the state-of-the-art results in subject-wise, binary classification of valence and arousal levels as well as four-class classification in the valence-arousal emotion space when raw EEG signals or their graph representations, in an architecture coined as GFT-STANN, are used as model inputs.**

I. INTRODUCTION

Emotions and their corresponding affective states play an important role in human life and behavior [1]. Automatically extracting information about emotions could enhance human-machine interactions and assist healthcare workers and caregivers to communicate with patients suffering from expression and speech problems. Therefore, emotion recognition using physiological signals, with the potential to improve the performance of brain-computer interface (BCI) systems, has received significant amount of attention lately. Multichannel electroencephalography (EEG) carries spectral and rhythmic brain signals that provide information about neural activity in specific cortical regions [2]. Ease of use and high temporal resolution - in comparison with other non-invasive recording techniques - make EEG a desirable modality to study emotions.

EEG-based emotion recognition consists of two main stages: extracting discriminative features and performing classification. The commonly used features in these tasks are Hjorth parameters, fractal dimension, high order statistics, differential entropy, power spectral density, rational and differential asymmetry, and differential causality [3], [4], [5].

This work has been partially supported by the National Science Foundation (NSF) under grants CCF-1934962 and DGE-1922591. *Corresponding author.

[1]S. Sartipi and M. Cetin are with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627, USA (ssartipi@ur.rochester.edu, mujdat.cetin@rochester.edu).

[2]M. Torkamani-Azar is with the School of Computing, University of Eastern Finland, Joensuu 80110, Finland (mastaneh.torkamani@uef.fi).

[3]M. Cetin is with the Goergen Institute for Data Science, University of Rochester, Rochester, NY 14627, USA.

Deep learning methods have been shown to outperform traditional methods in different fields, including computer vision [6] and biomedical signal processing [7]. Several architectures and methodologies have been proposed to deal with EEG emotion recognition based on deep learning methods [8], [9]. For instance, in [8], authors assign different weights to the EEG channels by applying a channel-wise attention mechanism. They then use a convolutional neural network (CNN) and a recurrent neural network (RNN) to extract the spatio-temporal features. Authors in [10] have applied a CNN on the frequency and time domain features and have shown that the combination of the raw EEG data with temporal and frequency-based features outperforms shallow networks. Authors of [11] have introduced the multi-column convolutional neural network (MCNN) for emotion classification. They evaluate their method in a subject-independent scheme by considering five participants as the test data. Separable EEGNet based on the Hilbert-Huang transform has been proposed in [9]. The data are transformed into the time-frequency domain, and feature extraction is performed by the combination of point-wise and depth-wise network elements.

It has been shown that CNNs are effective in extracting spatial information while RNNs capture the time dependencies well. EEG data are recorded from multiple electrodes that form a spatial structure. In order to process these structured time series effectively, both spatial and temporal information need to be accounted for. We propose the parallel spatio-temporal attention neural network (STANN) that takes into account these two aspects of the data within a unified architecture. This new architecture constitutes the main technical contribution of our paper. STANN also utilizes the advantage of time scaling as offered by bidirectional attention networks. Focus of the attention mechanism on specific time scales - by multiplication of hidden state outputs by trainable weights - can be physiologically interpreted by language-related components of event-related potentials (ERPs) and the time it takes for the brain to perceive and react to emotionally-loaded stimuli. Due to the complex structure of brain signals and their time-varying nature, besides using raw EEG signals as the input, we also propose the idea of using the graph Fourier transform (GFT) [12] of those signals as the input to the proposed network. To that end, we consider EEG electrodes as graph nodes and form the graph based on the Euclidean distances among them. Unlike the traditional common spatial pattern (CSP) filtering approach that is dependent on individual participants or tasks, in this work, we only benefit from the positions of the scalp EEG electrodes that are constant across all participants.

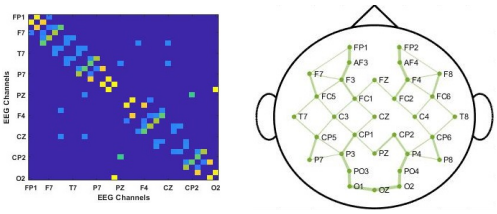

Fig. 1. Illustration of (**left**) the adjacency matrix, and (**right**) corresponding graph from a sample montage of the 10-20 electrode placement system.

In this way, our graph Fourier transform spatio-temporal attention neural network (GFT-STANN) captures the spatial information along with discriminative time dependencies. The proposed method is evaluated on the publicly available DEAP dataset [13]. We provide comprehensive experimental results to show the benefits of STANN with raw and graph-based representations of EEG data.

II. Graph-based EEG data representation

EEG data are recorded from multiple electrodes over the scalp which results in a two-dimensional (2D) graph signal $\mathbf{X}_{tr} \in \mathbb{R}^{N \times T}$, where N is the number of electrodes and T is the number of time points. Due to the structural and functional connectivity of the brain [14], exploring relative spatial locations of these electrodes helps with decoding responses elicited from sensory stimuli [15]. Here, we model the scalp structure as an undirected weighted graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathbf{A})$, where $\mathcal{V} = \{1, 2, ..., N\}$ is the set of nodes or channels, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges, and $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the adjacency matrix. The weight value, $\mathbf{A}_{ij}$, between two nodes i and j is calculated based on the inverse Euclidean distance d_{ij} as follows:

$$\mathbf{A}_{ij} = d_{ij}^{-1}, \ \mathbf{A}_{ii} = 0, \quad \text{for } i, j = 1, 2, ..., N. \tag{1}$$

K nearest neighbors (KNN) are computed for each node to construct the symmetric adjacency matrix [16]. Here, K is set to 2. Fig 1 illustrates the 2-NN scalp topology for a 10-20 electrode placement system.

Besides this graph layout, the spectral representation of spatial EEG signals could provide information regarding their characteristics. GFT is used to perform spatial frequency analysis of the signals over the graph. Let $\mathbf{D}$ be the diagonal matrix of the node degrees, $\mathbf{D}_{ii} = \sum_k \mathbf{A}_{ik}$, and the combinatorial graph Laplacian be $\mathbf{L} = \mathbf{D} - \mathbf{A}$ [12]. The GFT of signal $\mathbf{X}_{tr}$ with respect to $\mathbf{L}$ is calculated as follows:

$$\tilde{\mathbf{X}}_{tr} = \mathbf{V}^{\mathbf{T}} \mathbf{X}_{tr} \tag{2}$$

where $\mathbf{V}$ is the the orthonormal matrix of eigenvectors of the matrix $\mathbf{L}$.

III. Spatio-temporal Attention Neural Network

In this section, details of the proposed STANN shown in Fig. 2 are presented. The MCNN, recurrent attention network, and the proposed STANN architecture that combines these components are described in this section.

MCNN. The MCNN architecture follows the structure introduced in [17]. In this framework, there are several independently acting columns that are essentially functioning as

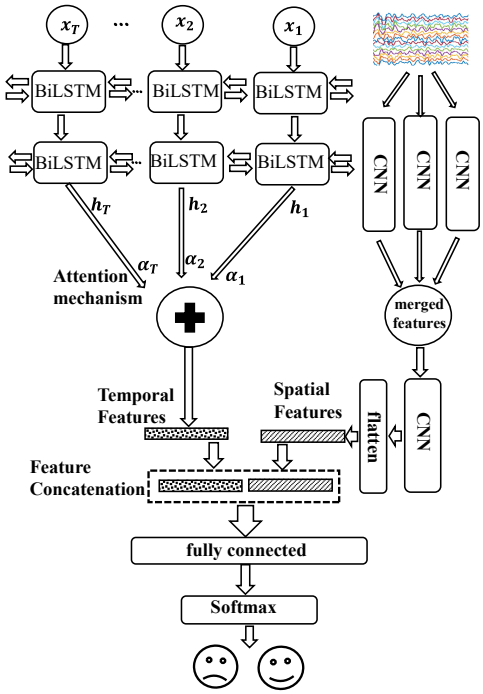

Fig. 2. Proposed parallel spatio-temporal attention neural network (STANN) architecture.

deep networks. The weights of all columns are initialized randomly and all columns start to train on the same input. The model output is the average of all the columns' outputs. In the present study, each column consists of a network with 2D CNN, batch normalization, and pooling layers.

Recurrent Attention Network. RNNs capture the dependencies across time steps from time-series data. These networks exploit the temporal information by establishing connections between subsequent layers [18]. This property makes RNNs perfect for learning short-term dependencies. Furthermore, Long Short-Term Memory (LSTM) resolves the vanishing gradient problem by maintaining the gradient back-propagation to earlier time steps and keeping long-term temporal dependencies.

Let x_t and h_t denote the input data and the hidden state at time t, respectively. Three gates control the LSTM performance. The input gate (i_t) controls the flow of the input, the forget gate (f_t) selects which information should be kept or forgotten, and the output gate (o_t) computes the output of the given updated cell. More details regarding the operations within LSTM cells can be found in [18].

Bidirectional LSTM (BiLSTM) entails two LSTM blocks in a single layer, which simultaneously process the information in two opposite directions. The output of each layer is the concatenation of the outputs of two LSTM blocks, i.e, $h_i = [\overrightarrow{h_f}, \overleftarrow{h_b}]$ where $\overrightarrow{h_f}$ and $\overleftarrow{h_b}$ correspond to the forward and backward hidden states, respectively [18].

Certain time steps might carry the most discriminative information, and attention mechanism serves the purpose of emphasizing those steps [19]. The output of the attention mechanism is the multiplication of outputs of hidden states by trainable weights. Given h_i as the output of the i^{th} LSTM

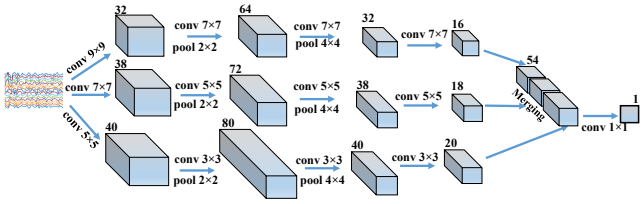

Fig. 3. Details of the MCNN model implemented in this paper. Each column has a 2D CNN structure.

cell and letting W and b be the trainable parameters, the output of the attention layer, v, is found as follows:

$$v = \sum_i \alpha_i h_i, \quad \alpha_i = \frac{\exp(Wh_i + b)}{\sum_j \exp(Wh_j + b)}. \tag{3}$$

Proposed STANN architecture. STANN involves parallel operation of MCNN and attention-based BiLSTM. The network consists of two parts: MCNN for encoding the spatio-temporal information within a temporal slice and the recurrent attention network for exploiting the attentive temporal dependencies across different time steps. As shown in Fig. 2, spatial and temporal features are computed in parallel and concatenated at the outputs of MCNN and BiLSTM networks. The MCNN part of the model consists of three columns. The same network structure is used for each column except kernel sizes and number of filters. Different kernel sizes explore a variety of short and long range dependencies across nearby EEG channels. In the context of EEG signals, the input of MCNN has the shape of number of EEG channels×number of time steps×1. Fig. 3 shows details of the proposed MCNN. Each column consists of four convolution (conv) layers followed by batch normalization. An average-pooling layer followed by a dropout layer is applied after the first and second conv layers. The dropout probability rates are set to 0.5 and 0.4, respectively. In every conv operation, the same zero-padding technique is used to prevent losing the edge information of the input data. The widely used activation function in CNNs, rectified linear unit (ReLU), is adopted in this network. After merging the outputs of all columns, a 1×1 conv filter is applied to calculate the spatial feature maps.

The LSTM part of the network entails two BiLSTM layers with the same hidden layer size, 80, and 128 time steps. The forward and backward LSTM layers are each followed by a dropout layer with probability rates of 0.2 and 0.1, respectively. BiLSTM outputs are fed to the attention mechanism. The hyperbolic tangent (tanh) activation function is used for both BiLSTM layers.

As shown in Fig. 2, the extracted spatial and temporal features are flattened and concatenated. Finally, the feature vector is passed through the fully connected layer with 128 hidden units before going through a SoftMax operation. The model is implemented in Python with Tensorflow and Keras libraries.

IV. Experiments and Results

DEAP Dataset. The proposed architecture is evaluated using the DEAP dataset [13] recorded from 32 individuals each having rated 40 one-minute long music videos. The physiological recordings consist of 32 and 8 channels related to EEG and peripheral physiological signals, respectively. In this paper, we only use the EEG signals and refer to each one-minute EEG recording as a trial. Participants were asked to rate the level of valence, arousal, liking, and dominance in each video from 1 to 9. The recorded signals were downsampled from 512 Hz to 128 Hz, ocular artifacts were removed, and a bandpass filter from 4.0 to 47.0 Hz was applied. Each EEG recording thus contains 60 s trial data, in addition to the 3 s baseline data.

TABLE I

Average classification accuracies (%) for different scenarios.

Method	Theta	Alpha	Beta	Gamma	Wide-band
Binary valence classification[1]					
BiLSTM	84.6	85.5	87.0	83.5	89.0
MCNN	86.1	86.0	**91.4**	83.2	93.3
STANN	88.1	88.5	91.2	**86.6**	94.4
GFT-STANN	**89.8** ±4.7	**89.0** ±6.0	91.3 ±5.0	85.2 ±6.9	**94.8** ±**2.9**
Binary arousal classification[2]					
BiLSTM	87.0	87.2	88.1	85.0	90.3
MCNN	88.5	87.0	91.6	84.2	93.9
STANN	90.2	89.7	**92.5**	**86.7**	94.9
GFT-STANN	**91.7** ±4.1	**90.5** ±5.1	92.3 ±4.7	86.2 ±6.7	**96.1** ±**2.2**
Four-class valence-arousal classification					
BiLSTM	80.9	81.2	84.2	**85.0**	77.7
MCNN	81.7	81.3	87.5	73.0	89.7
STANN	84.1	84.1	86.5	77.6	90.9
GFT-STANN	**87.6** ±5.3	**86.7** ±7.0	**88.6** ±5.5	78.2 ±7.7	**92.7** ±**4.2**

Average number of positive and negative samples per participant: [1](1327,1073) and [2](1382,1018).

Classification Results. In this study, we segment the 60-s long trials into 1 s data samples. The size of data samples is equal to 32×128 where 32 is the number of the EEG channels and 128 is the number of time samples. The trial data are baseline corrected. Thus, the data for each participant consist of $40 \times 60 = 2400$ data samples. Each data sample is filtered into four subbands as theta (4-8 Hz), alpha (8-12 Hz), beta (12-29 Hz), and gamma (30-47 Hz). We validate the performance of our proposed framework by considering two classification schemes. The first scenario involves binary classifications of high-versus-low valence and high-versus-low arousal. To obtain a binary problem, the 9-level ratings of valence and arousal are quantized into two levels using a threshold of 5. In the second scenario, the valence-arousal (VA) space [20] is divided into four sub-spaces, i.e., low valence-low arousal, low valence-high arousal, high valence-high arousal, and high valence-low arousal [15]. For each scenario, the subject-wise 10-fold cross-validation (CV) is repeated 10 times, and the average classification accuracy is reported. The model is trained by the Adam optimizer [21] to minimize the cross-entropy between the predicted and true labels. The batch size and epochs are selected as 300 and 35, respectively. All the parameters are selected using a grid search paradigm until the highest average classification accuracy is achieved.

For comparison, we consider MCNN and BiLSTM ar-

TABLE II

COMPARISON OF THE PROPOSED GFT-STANN WITH STATE-OF-THE-ART METHODS FROM RECENT LITERATURE.

Method	Valence (%)	Arousal (%)	Four-class (%)
Proposed method	**94.8**	**96.1**	**92.7**
Tao *et al.* [8]	93.7	93.4	-
Huang *et al.* [9]	89.9	88.3	-
Chen *et al.* [10]	88.8	86.9	-
Soroush *et al.* [24]	-	-	89.8
Li *et al.* [15]	-	-	62.0

chitectures with raw data as baseline models and evaluate their performance separately using the aforementioned parameters. This evaluation demonstrates how the proposed parallelization is beneficial in improving the classification accuracy. In order to assess the effect of the graph-based EEG data representation, we calculate the GFT coefficients of frequency subbands and compare results of the proposed model with two different input modalities, i.e. raw EEG and EEG-based GFT coefficients. Since computation of GFT is independent from subjects and tasks, it would not hurt the automated operation of our model. Table I depicts the binary valence, binary arousal, and four-class classification accuracies for the baseline methods and proposed approach based on features from various frequency bands. The average classification accuracies for binary valence and arousal and four class classification problems based on GFT-STANN are 94.8%, 96.1%, and 92.7%, respectively. These results show that not only the proposed network outperforms the baseline methods, but also the GFT improves the overall performance. Furthermore, results in Table I demonstrate that wide-band EEG and beta band features outperform other spectral features in binary classifications of high-versus-low valence and arousal states. These results are in line with the role of frequency bands in characterizing emotional processes [22], [23].

Table II presents a comparison of the proposed method with several methods from the recent literature and demonstrates the superiority of our proposed approach. All the results reported here are subject-dependent with a 10-fold CV except for [9] which involves a 4-fold CV.

V. CONCLUSION

We have proposed an end-to-end deep learning framework for EEG emotion recognition. The presented GFT-STANN approach captures the spatial and temporal information over the graph-based input data in a parallel format. Graph-based representation of EEG signals provides a concise set of structural and graph-spectral domain information without being dependent on individual differences or conducted experiments. Moreover, the attention mechanism helps to find the most discriminative time steps. The fused spatio-temporal features achieve higher accuracy compared to state-of-the-art methods in valence and arousal classification on the DEAP dataset.

REFERENCES

[1] W. A. Cunningham and T. Kirkland, "Emotion, cognition, and the classical elements of mind," *Emotion Review*, vol. 4, no. 4, pp. 369–370, 2012.

[2] E. Niedermeyer and F. L. da Silva, *Electroencephalography: basic principles, clinical applications, and related fields*, Lippincott Williams & Wilkins, 2005.

[3] H. Sanggarini, I. Wijayanto, and S. Hadiyoso, "Hjorth descriptor as feature extraction for classification of familiarity in EEG signal," in *2019 International Conference on Information and Communications Technology (ICOIACT)*. IEEE, 2019, pp. 306–309.

[4] L.-C. Shi, Y.-Y. Jiao, and B.-L. Lu, "Differential entropy feature for EEG-based vigilance estimation," in *Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*. IEEE, 2013, pp. 6627–6630.

[5] Y.-P. Lin, C.-H. Wang, T.-P. Jung, T.-L. Wu, S.-K. Jeng, J.-R. Duann, and J.-H. Chen, "EEG-based emotion recognition in music listening," *IEEE. Trans. Biomed. Eng.*, vol. 57, no. 7, pp. 1798–1806, 2010.

[6] M. Jaderberg, K. Simonyan, A. Zisserman, et al., "Spatial transformer networks," in *Annu. Conf. Neural Inform. Process. Syst. (NIPS)*, 2015, pp. 2017–2025.

[7] S. K. Goh, H. A. Abbass, K. C. Tan, A. Al-Mamun, N. Thakor, A. Bezerianos, and J. Li, "Spatio–spectral representation learning for electroencephalographic gait-pattern classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 9, pp. 1858–1867, 2018.

[8] W. Tao, C. Li, R. Song, J. Cheng, Y. Liu, F. Wan, and X. Chen, "EEG-based emotion recognition via channel-wise attention and self attention," *IEEE Trans. Affective Comput.*, 2020.

[9] W. Huang, Y. Xue, L. Hu, and H. Liuli, "S-EEGNet: Electroencephalogram signal classification based on a separable convolution neural network with bilinear interpolation," *IEEE Access*, vol. 8, pp. 131636–131646, 2020.

[10] J. Chen, P. Zhang, Z. Mao, Y. Huang, D. Jiang, and Y. Zhang, "Accurate EEG-based emotion recognition on combined features using deep convolutional neural networks," *IEEE Access*, vol. 7, pp. 44317–44328, 2019.

[11] H. Yang, J. Han, and K. Min, "A multi-column CNN model for emotion recognition from EEG signals," *Sensors*, vol. 19, no. 21, pp. 4736, 2019.

[12] S. S. Saboksayr, G. Mateos, and M. Cetin, "Online discriminative graph learning from multi-class smooth signals," *Signal Processing*, vol. 186, pp. 108101, 2021.

[13] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 18–31, 2011.

[14] H.-J. Park and K. Friston, "Structural and functional brain networks: from connections to cognition," *Science*, vol. 342, no. 6158, 2013.

[15] P. Li, H. Liu, Y. Si, C. Li, F. Li, X. Zhu, X. Huang, Y. Zeng, D. Yao, Y. Zhang, et al., "EEG based emotion recognition by combining functional connectivity network and local activations," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 10, pp. 2869–2881, 2019.

[16] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nature reviews neuroscience*, vol. 10, no. 3, pp. 186–198, 2009.

[17] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2016, pp. 589–597.

[18] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," *arXiv preprint arXiv:1506.00019*, 2015.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[20] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161, 1980.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] W. J. Ray and H. W. Cole, "EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes," *Science*, vol. 228, no. 4700, pp. 750–752, 1985.

[23] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Mental Develop.*, vol. 7, no. 3, pp. 162–175, 2015.

[24] M. Z. Soroush, K. Maghooli, S. K. Setarehdan, and A. M. Nasrabadi, "Emotion recognition using EEG phase space dynamics and poincare intersections," *Biomed Signal Process Control*, vol. 59, pp. 101918, 2020.