Spatial-Temporal Preserving Multimodal Algorithm for EEG-based Emotion Recognition

Heng Lu *

Department of Computer Science, Metropolitan College, Boston University, 1010 Commonwealth Ave, Boston, MA USA 02215

* Corresponding Author Email: hengl@bu.edu

Abstract. Emotion recognition has been the focus of some human computer interaction researchers. Designing an efficient algorithm for emotion recognition is crucial for its potential use in HCI. Many algorithms have been developed for emotion recognition and a majority of them is based on electroencephalography (EEG). Despite the success of spatial–temporal recurrent neural network (STRNN) and multimodal deep neural networks, they can still be improved. This paper proposes an algorithm which tries to combine STRNN with multimodal deep neural network, where EEG signals are classified using STRNN and eye movement signals are classified using SVM. The results from the two classifiers are fused by max fusion. This paper shows that by taking spatial and temporal information of EEG signals into consideration and using multimodal fusion, the performance of emotion recognition can be improved.

Keywords: Emotion Recognition, Human-computer Interaction, Machine Learning.

1. Introduction

Emotion plays an important role in our daily activities, such as studying and working, and is arousing increasing researchers' interest due to its potential use in human computer interaction [1]. Moreover, some mental diseases have close relationship with human emotions such as depression and autism spectrum disorder [2]. For example, patients who have difficulty expressing their feelings can benefit from emotion recognition [3], and psychologists may perform better diagnosis with the help of emotion recognition.

Various emotion recognition systems have been built with different types of signals, and they have different benefits and limitations [3]. For instance, facial-recognition-based systems are convenient to train and use since the needed data can be easily collected without any specific sensors and even some simple algorithm can achieve high accuracy, while they might suffer from culture differences [4]. M. Egger et al. [3] analyzes the advantages and disadvantages of different systems, such as electromyography (EMG)-based systems and electrodermal-activity-based (EDA) systems. Among these systems, electroencephalography (EEG)-based systems are believed to be reliable since its high accuracy and stable neural patterns [5].

However, EEG-based systems still face some challenges [6], such as its inevitable noises due to eye blinks or the participants' distraction. Moreover, the connections between EEG signals and brain activities remain unclear, which introduces difficulties in interpreting the signals. Also, the preprocessing of EEG signals might be hard and tricky [10].

In recent years, researchers have developed various algorithms for EEG-based emotion recognition, which varies from simple SVM and k nearest neighbors [7] to complex neural networks. T. Zhang et al. [8] propose a spatial–temporal recurrent neural network (STRNN) to integrate both spatial and temporal information of EEG signals, which gains a great improvement in classification accuracy. Moreover, due to the complexity of emotions, it is difficult to describe emotions with single-modality signals such as EEG. Some researchers then propose emotion recognizing methods with multimodal signals like the combination of EEG signals and eye movement signals [9]. Although these algorithms achieve notable performance, they can still be improved.

First, the STRNN framework [8] requires careful definition of scanning orders in spatial RNN layers, which requires serious consideration by humans. However, the handcrafted scanning orders are not always available because of the possible lacking in the layout of electrodes. Second, some advanced algorithms have been developed for EEG-based multimodal emotion recognition, while the spatial-temporal dependencies are ignored, introducing difficulty in the interpretability of the system.

2. Methods

2.1. Datasets and feature extraction

The dataset used is SEED-V [5]. SEED-V contains EEG signals and eye movements of 16 subjects, who were required to watch emotional film clips for five emotions (happy, sad, neutral, fear, and disgust).

For EEG signals, differential entropy features are extracted for emotion recognition, which are calculated using short-term Fourier transform with a non-overlapping Hanning window in five frequency bands (delta: 1-4 Hz, theta: 4-8 Hz, alpha: 8-14 Hz, beta: 14-31 Hz, and gamma: 31-50Hz). Finally, a sliding window of 8s is applied to scan the sequences and the sequences in the window are used as the representation of the point which is in the center of the window. For eye movement signals, the same 33 features are extracted as W. Liu et al. [5].

After feature extraction, an EEG sample has a dimension of T×n×d. T denotes the number of EEG segments in one EEG sample and is set to 8 since the slicing window is of 8s. n denotes the number of electrodes and is set to 62 since the EEG signals in SEED-V dataset were collected using a 62-channel ESI NeuroScan System [5]. d denotes the number of features associated with each electrode and is set to 5 since five DE features are calculated for each EEG sequence.

2.2. Spatial-temporal RNN

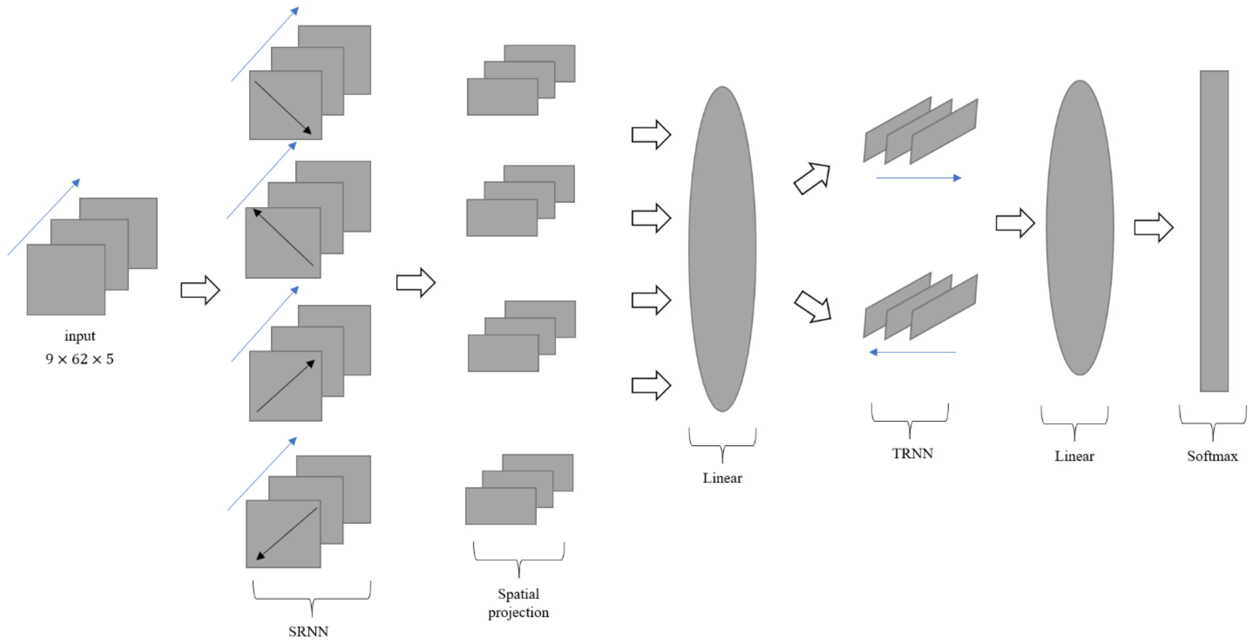

Figure 1. Structure of spatial–temporal recurrent neural network

The STRNN framework is proposed by ZHANG et al. [8]. In the spatial layer, four directional RNNs are used to extract spatial features by traversing the region from four specific angles, which is intended to reduce the effect of noises. After spatial projection, the representations learned from the SRNN layer are sequentially concatenated at each time slice and thus form a temporal sequence. Then, a bi-directional RNN are applied to capture temporal representations. Finally, the output of TRNN layer is put into the softmax layer for emotion recognition.

In the SRNN layer, the numbers of input node, hidden node and output node are respectively set to be 5, 30, 30. The spatial projection is simply set to be taking the last 10 spatial hidden states. The

TRNN layer has 30 hidden nodes and 5 output nodes which correspond to the five categories of emotions.

2.3. Multimodal emotion recognition

The EEG signals are classified to five emotions using STRNN and the eye movement signals are classified using a simple support vector machine. After this, a max fusion method is applied to perform multimodal emotion recognition, where the maximum values of all probabilities that a sample belongs to each category in the two classifiers and make the prediction by the highest probabilities.

3. Result

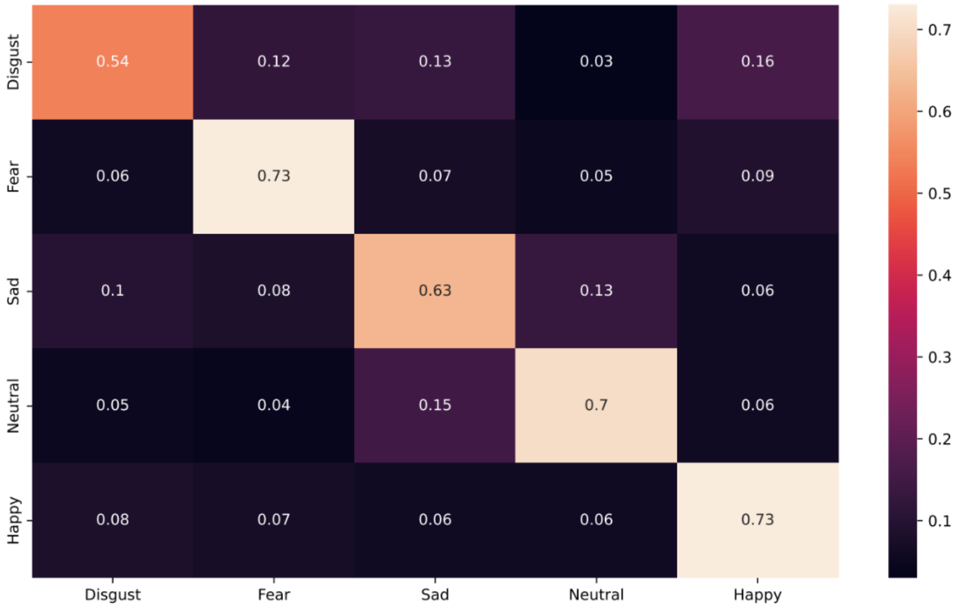

Figure 2. Confusion matrix for STRNN on SEED-V dataset

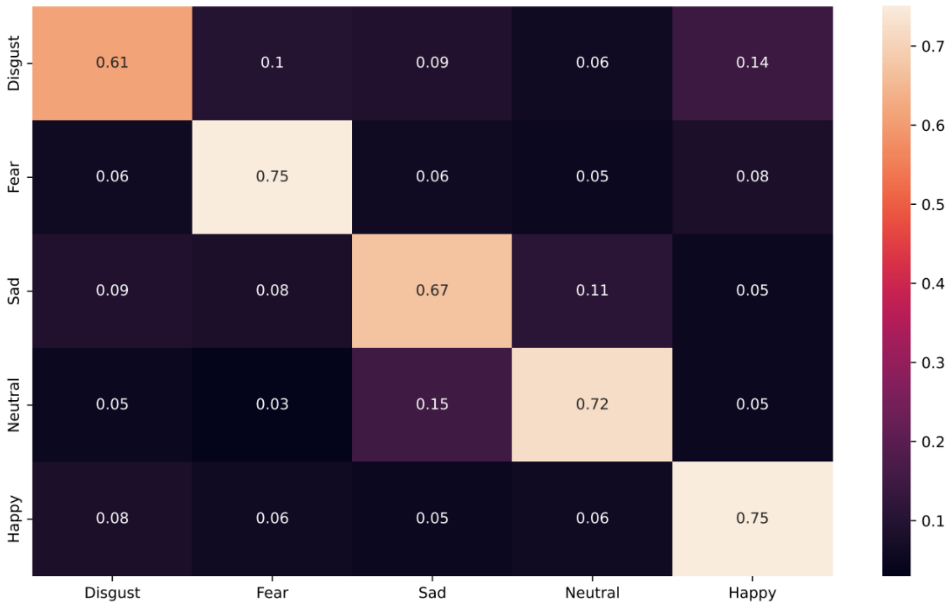

Figure 3. Confusion matrix for multimodal emotion recognition

The classification result of STRNN on EEG signals is shown in Figure 2. The algorithm performs worst in the recognition of the disgust emotion and only has an accuracy of 54%. The average accuracy on the five emotions is 66.6%. As Figure 3 shows, the classification accuracy increases when eye movement signals are taken into consideration and max fusion is used to produce final predictions. The final average accuracy is 70.0% for multimodal emotion recognition.

4. Conclusion

In this paper, a multimodal algorithm for emotion recognition is proposed, where a STRNN is used to capture the spatial and temporal information of EEG signals and max fusion is used to fuse classifiers for multimodal emotion recognition. The paper has shown that a STRNN is useful for emotion recognition based on EEG signals, and even a simple multimodal fusion method called max fusion can improve the classification performance. However, the result of paper is not satisfying since the accuracy is lower than the classification accuracy of some advanced techniques such as DCCA [5], which achieves an average accuracy of 85.6% on SEED-V. The low accuracy may result from the simple spatial projection since some important spatial information may be lost in this process, and a careful reconsideration of the multimodal fusion method may also help.

References

[1] Mauss, I. B., and Robinson, M.D., “Measures of emotion: A review, Cognition and Emotion”, 23 (2), 209 - 237 (2009).

[2] Zheng, W. -L., Liu, W., Lu, Y., Lu, B. -L., and Cichocki, A., “EmotionMeter: A Multimodal Framework for Recognizing Human Emotions”, IEEE Transactions on Cybernetics, 49 (3), 1110 - 1122 (2019).

[3] Egger, M., Ley, M., Hanke, S., “Emotion Recognition from Physiological Signal Analysis: A Review”, Electronic Notes in Theoretical Computer Science, 343, 35 - 55 (2019).

[4] Ekman, P., “Universals and cultural differences in facial expressions of emotion”, Nebraska Symposium on Motivation, 19, 207 - 283 (1971).

[5] Liu, W. et al., “Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition”, IEEE Transactions on Cognitive and Developmental Systems, 14 (2), 715-729 (2022).

[6] Zhang, D. et al., “Making Sense of Spatio-Temporal Preserving Representations for EEG-Based Human Intention Recognition”, IEEE Transactions on Cybernetics, 50(7), 3033-3044 (2020).

[7] Mehmood, R. M., Lee, H. J., “A novel feature extraction method based on late positive potential for emotion recognition in human brain signal patterns”, Computers and Electrical Engineering, 53, 444 - 457 (2016).

[8] Zhang, T. et al., “Spatial–Temporal Recurrent Neural Network for Emotion Recognition”, IEEE Transactions on Cybernetics, 49 (3), 839 - 847 (2019).

[9] Poria, S. et al., “A review of affective computing: from unimodal analysis to multimodal fusion”, Information Fusion, 37, 98 - 125 (2017).

[10] Nakisa, B. et al., “Evolutionary computation algorithms for feature selection of EEG-based emotion recognition using mobile sensors”, Expert Systems with Applications, 93, 143 - 155 (2018).