Multimodal Physiological Signal Emotion Recognition Based on Convolutional Recurrent Neural Network

Jinxiang Liao\textsuperscript{a}, Qinghua Zhong\textsuperscript{*}, Yongsheng Zhu\textsuperscript{b} and Dongli Cai\textsuperscript{c}

School of Physics & Telecommunication Engineering, South China Normal University, Guangzhou 510006, China

\textsuperscript{*}Corresponding author e-mail: zhongqinghua@m.scnu.edu.cn,
\textsuperscript{a}liaojinxiang@m.scnu.edu.cn, \textsuperscript{b}jerk_zhu@m.scnu.edu.cn, \textsuperscript{c}caidongli@m.scnu.edu.cn

Abstract. In order to solve the problem that the emotion recognition rate of single-mode physiological signals is not high in the physiological signals based emotion recognition, in this paper, we propose a convolutional recurrent neural network based method for multi-modal physiological signal emotion recognition task. The method used convolutional neural network to learn the spatial representations of multi-channel EEG signals and the Long Short-term Memory network to learn the temporal representations of peripheral physiological signals (EOG, EMG, GSR, RSP, BVP, and TMP). The two representations are combined for emotion recognition and classification. In the two emotion dimensions of Arousal and Valence, our experiments conducted on the open source dataset DEAP shows that, this method achieve 89.68\% and 89.19\% average accuracy in the EEG emotion classification, 63.06\% and 62.41\% average accuracy in the peripheral physiological signal emotion classification, 93.06\% and 91.95\% average accuracy in the combined feature emotion classification. The experimental results show that the convolutional recurrent neural network based method that we proposed efficiently extract multi-modal physiological signal feature to improve the emotion recognition performance.

1. Introduction

Emotion is a comprehensive state generated by whether objective things meet people's own needs. Different emotional states affect people's cognition, thinking, and decision-making. The recognition of different emotional states has a wide range of applications in distance education, medical treatment, intelligent systems, and human-computer interaction (HCI). In recent years, there has been an increasing interest in emotion recognition \cite{1}.

At present, the following approaches are mainly used to achieve emotion recognition: speech, facial expressions, comment, limb movements and physiological signals. However, speech, facial expressions, text, and body movements are non-physiological signals that are external manifestations of human emotions, except physiological signals. So it can be easily concealed by humans. Physiological signals, including electrocardiogram (ECG), my electricity (EMG), skin electricity (GSR), respiration (RSP), and electroencephalogram (EEG) produced through changes in the nervous system and endocrine system are not easily affected by human subjective consciousness control, so it can achieve higher emotion recognition objectivity and accuracy\cite{2}. Previous research mostly used single-modal data for
emotion recognition, such as, [3] [4] recognize emotions through facial expressions, and [5] [6] extract features from EEG for emotion recognition. [7] Has shown that certain emotional states because the same changes in individual physiological signals. For example, both "anger" and "surprise" cause heart rate to increase. The difference is that "anger" reduces heart rate variability and increases skin conductance, while "surprise" does not have significant changes. Therefore, it is possible to obtain more objective and accurate results by fusing the characteristics of multi-modal physiological signals. [8] used Hidden Markov Model (HMM) to extract physiological signal characteristics, and fused the three physiological signals of EEG, GSR, and heart rate (HR), proving that multiple physiological signals can form complementary advantages and improve emotion recognition effect. A lot of previous related researches have proved the effectiveness of traditional machine learning based emotion classification. However, the effect of machine learning methods on emotion classification is very dependent on the extraction of emotion-related features. Among them, the most commonly used features are Fourier transform (FT), power spectral density (PSD), and wavelet transform (WT). In recent years, the use of deep learning to extract abstract features has become a new trend. Deep learning not only reduce the workload of feature engineering but also improve recognition accuracy. With the rise of deep learning, convolutional neural networks have achieved good results in video image processing, but there are still rare application in feature extraction and classification of physiological signals. A number of studies have attempted to use Convolutional Neural Networks (CNN) to extract EEG emotional features directly, but did not achieve the desired classification effect. [9] Directly adopted CNN to recognize the electroencephalogram signals of motor imagination, and obtained only 45% classification accuracy. In [10], each electrode of the EEG and its nearest four electrode channels are combined into a new matrix. The accuracy of this method by directly using a convolutional neural network for feature extraction and classification has reached 70.11%. However, most CNN-based methods now rely heavily on complex preprocessing and feature engineering, or ignore the subtle spatial information contained in EEG signals. For example, in [11], the original EEG signal was converted into an image, and the ability of CNN to automatically learn features and share representations was not fully utilized.

This paper proposes a convolutional recurrent neural network based method for multi-modal physiological signal emotion recognition. Aiming at the problem that EEG emotion recognition relies on complex preprocessing and feature engineering, CNN is used to extract the subtle spatial features of EEG, and Long Short-term Memory (LSTM) is used to extract peripheral physiological signals (eye electricity, my electricity, skin electricity, breathing Rhythm, blood volume pulse, and body temperature) temporal representations, ready for emotional recognition. In order to solve the problem of emotional recognition limitation of single-mode physiological signals, the spatial characteristics of EEG and the temporal characteristics of peripheral physiological signals (PERI) are combined to form multi-modal physiological signal feature for emotional recognition, thereby achieving the purpose of improving emotion classification accuracy.

2. METHOD

2.1. Convert One-dimensional EEG Signal to Two-dimensional EEG Matrix

The EEG-based BCI system uses an international 10-20 system to obtain EEG signals. Depending on recording channels number, different BCI system was used for obtaining EEG data. As shown in Fig.1, the circles in red indicate the test points used in the DEAP data set. In general, the original data collected by the EEG signal acquisition system at time $t$ is a one-dimensional data vector $v_t = [s^1_t, s^2_t, \cdots, s^n_t]$, while $s^n_t$ is the data recorded by the nth electrode channel at time $t$. The EEG acquisition system has a total of $n$ channels for reading EEG data, and $n$ equals 32 in this paper. From the EEG electrode map, it is observed that each electrode is physically neighboring multiple electrodes which records the EEG signals in a certain area of brain, while the one-dimensional EEG vector sequence are restricted to two neighbors. Furthermore, different brain regions correspond to different EEG signals in different emotional states. In this paper, based on the spatial information of the electrode distribution of the
acquisition system, one-dimensional EEG data sequences are converted into a two-dimensional EEG data matrix, as shown in Fig.1.

Figure 1. Conversion of one-dimensional EEG sequences into two-dimensional EEG matrices.

In this paper, we generalized the International 10-20 System with test electrodes used in the DEAP dataset to form a matrix \((h \times w)\), where \(h\) is the maximum point number of the vertical test points and \(w\) is the maximum point number of the horizontal test points. With the DEAP dataset, \(h\) equals \(w\) equals 9. The one-dimensional EEG data sequence at time \(t\) corresponds to the transformation function of the two-dimensional EEG data matrix \(f_t\), as shown in following eq.1:

\[
f_t = \begin{bmatrix}
0 & 0 & 0 & s_{t}^1 & 0 & s_{t}^{17} & 0 & 0 & 0 \\
0 & 0 & 0 & s_{t}^2 & 0 & s_{t}^{18} & 0 & 0 & 0 \\
0 & 0 & 0 & s_{t}^3 & 0 & s_{t}^{19} & 0 & s_{t}^{20} & s_{t}^{21} \\
0 & s_{t}^4 & 0 & s_{t}^5 & 0 & s_{t}^{22} & 0 & s_{t}^{23} & 0 \\
0 & s_{t}^6 & 0 & s_{t}^{10} & 0 & s_{t}^{24} & 0 & s_{t}^{25} & 0 \\
0 & s_{t}^7 & 0 & s_{t}^8 & 0 & s_{t}^{26} & 0 & s_{t}^{27} & 0 \\
0 & s_{t}^9 & 0 & s_{t}^{11} & 0 & s_{t}^{28} & 0 & s_{t}^{29} & 0 \\
0 & s_{t}^{12} & 0 & s_{t}^{13} & 0 & s_{t}^{30} & 0 & s_{t}^{31} & 0 \\
0 & 0 & 0 & s_{t}^{14} & 0 & s_{t}^{32} & 0 & 0 & 0 \\
\end{bmatrix}
\]

In \(f_t\), “0” indicates an unused electrode channel in the DEAP data set, which has no effect on the neural network. Through the transformation of the transformation function \(f_t\), the one-dimensional EEG data vector sequence \([v_t, v_{t+1}, \cdots, v_{t+L}]\) is converted into a two-dimensional EEG data matrix sequence \([f_t, f_{t+1}, \cdots, f_{t+L}]\). During the duration \([t, t+L]\), the dimension of the two-dimensional EEG data matrix is \(L+1\). After performing the transformation function \(f_t\), the Z-score normalization is performed on non-zero elements for each EEG data matrix, as shown in eq.2:
\[ z = \frac{x - \mu}{\sigma} \] (2)

Among them, \( x \) is non-zero elements in the EEG data matrix, \( \mu \) denotes the mean of all non-zero elements, and \( \sigma \) denotes the standard deviation of all non-zero elements.

Finally, to divide the two-dimensional matrix sequence into multiple sets of two-dimensional matrix sequences a 1s sliding window is used. As shown in Fig.2, each matrix sequence has a fixed dimension and there is no overlap between two consecutive matrix sequences. The matrix sequence segment \( S_j \) is expressed as shown in eq.3:

\[ S_j = [f_t, f_{t+1}, \cdots, f_{t+L}] \] (3)

Among eq.3, \( S \) is the matrix sequence dimension, and the subscript \( j \) is the matrix sequence segment number.

![Sliding Window](image)

**Figure 2.** Segmentation of data.

### 2.2. Using 2D-CNN to Extract the Spatial Features of the EEG Matrix

In order to extract the spatial features of each two-dimensional EEG matrix, this paper applies three consecutive 2D convolutional layers (2D-CNN), as shown in Fig.3. The \( j \)th input matrix sequence segment is defined as eq.4:

\[ S_j = [f_t, f_{t+1}, \cdots, f_{t+S-1}] \in R^{S \times h \times w} \] (4)

There are \( S \) data matrices of size \( h \times w \). Each data matrix are input to 2D-CNN and parsed into a spatial feature vector \( V_S \), as shown in eq.5:

\[ V_S = \text{CNN}_{2D}(S_j), V_S \in R^{1215} \] (5)

Each layer of 2D-CNN has the same kernel size of 3x3 for the correlation between EEG channels extraction. The kernel of this size is widely applied in the field of computer vision. To prevent missing the information at the edge of the input data mesh we use zero-padding techniques in each convolutional operation. This method uses 64 feature maps for the first layer of convolution, and the second and third layers of convolution have 128 and 256 feature maps, respectively. In the classical CNN architecture of
computer vision, the convolutional layer is often followed by a pooling layer to reduce the data dimension. But the size of the data matrix in this paper is much smaller than that in computer vision, in order to reduce the loss of data information, no pooling operation is performed in the CNN model of this paper.

**Figure 3.** CNN learn the spatial representations of multi-channel EEG signals.

In order to speed up model convergence and prevent overfitting, a batch normalization layer (BN) is introduced after each layer of convolution operation. After performing three layers of 2D-CNN on each data matrix $f_t$, in order to obtain the EEG spatial characteristics during the duration $[t, t+S-1]$, each convolutional feature matrix $f_t$ are fused in depth. After that, the fused $9 \times 9 \times (256 \times S)$ matrix sequence is convolved with 15 convolution kernels with size of $1 \times 1$ to obtain the matrix sequence size of $15 \times 9 \times 9$, and it is expanded into a spatial feature vector $V_S \in \mathbb{R}^{1215}$ to prepare for feature fusion.

2.3. Using LSTM to Extract the Temporal Features of Peripheral Physiological Signals

In the DEAP data set, the signals collected on the 33-40 channels are 2 EOG, 2 EMG, 1 GSR, 1 RSP, 1 BVP, and 1 TMP, then a one-dimensional vector composed of peripheral physiological signals at a time $t$ can be expressed as eq.6:

$$ r_t = [S_t^{33}, S_t^{34}, S_t^{35}, S_t^{36}, S_t^{37}, S_t^{38}, S_t^{39}, S_t^{40}]' $$

(6)

After segmenting the data by using a 1s sliding window, we can obtain $j^{th}$ inputted data segment of RNN, as shown in $R_j$:

$$ R_j = [r_t, r_{t+1}, \cdots, r_{t+S-1}] $$

(7)

Where $S$ represents the sliding window size.

As shown in Fig.4, the peripheral physiological signal data vector $R_j$ is input to a Recurrent Neural Network (RNN) to learn the temporal representations. In this paper, LSTM units are used to construct two stacked the RNN layer. Each layer of RNN has $S$ LSTM units, and the input to the second RNN
layer is the output of the previous RNN layer. The hidden state of the LSTM unit of the first RNN layer at current time step \( t \) is denoted as \( h_t \), and the \( h_{t-1} \) is the hidden state of the previous time step \( t-1 \). The hidden state from the previous time \( t-1 \) as the input of the current time \( t \) affects the output of the LSTM unit. The input sequence of the second layer LSTM is the hidden state sequence \( [h_t, h_{t+1}, \ldots, h_{t+S-1}] \) of the first layer LSTM. The output \( h_{i+S-1} \) at the last time step \( t+S-1 \) is the temporal representations \( V_T \) of the peripheral physiological signal, as shown in eq.8 and eq.9:

\[
\begin{align*}
   h_{i+S-1} &= \text{RNN}_{LSTM}(R_f), h_{i+S-1} \in R^v \\
   V_T &= h_{i+S-1}
\end{align*}
\]

Where \( v \) is the hidden state size of the LSTM unit. The fully connected (FC) layer before LSTM is used to enhance the representation of temporal information.

Figure 4. LSTM learn the temporal representations of PERI.

2.4. Feature Fusion of Multi-modal Physiological Signals

After the operations of feature extraction, the spatial representations of EEG and the temporal representations of PERI can be obtained. As shown in Fig.5, the extracted EEG spatial features are fused with the temporal representations of the PERI. Finally, the fused features are fed into the softmax layer to predict the emotional state, as shown in eq.10:

\[
P_j = \text{softmax} \left( \left[ V_S, V_T \right] \right), P_j \in R^n
\]
Among them, \( V_s \) denotes EEG spatial feature vector, \( V_r \) is PERI temporal feature vector, and \( n \) represents emotion category. In this paper, \( n \) equals 2.

![Multimodal feature fusion and emotion recognition](image)

**Figure 5.** Multimodal feature fusion and emotion recognition.

3. Experiment and result analysis

3.1. DEAP Dataset

| Table 1. Description of DEAP dataset. |
|--------------------------------------|
| Number of Subjects | 32 |
| Number of Experiments | 40/Subject |
| Stimulate | Music Video |
| Label | Valence (1-9); Arousal (1-9); Dominance (1-9); Liking (1-9) |
| Signal | 32-channels EEG; 8-channels PERI (2EOG; 2EMG; 1GSR; 1RSP; 1BVP; 1TMP) |

DEAP is a multi-modal physiological signal dataset for analyzing human emotions [12]. DEAP stimulated the emotions of the experimental participants through a one-minute music video, which recorded the 32-channels EEG and 8-channels PERI corresponding to emotions in 40 different experiments of 32 experimental participants. The data were cut into segments of 63s. The first 3s are the pre-trial baseline. The original physiological signal in the data set was collected at a sampling rate of 512 Hz, and the final signal was down-sampled to 128 Hz. The EEG signal is band-pass filtered at 4.0-45.0 Hz, and the EOG (eye movement) artifacts are removed. The specific description is shown in Table 1.

3.2. Emotion Model

Emotions should be defined and represented quantitatively, for emotion recognition. Relevant scholars first proposed basic definitions of emotions decades ago, but precise definitions have not been widely accepted by psychologists. Scholars are more inclined to simulate emotions in the following two ways. One is to divide emotions into different basic categories. Ekman and Friesen consider emotions is discrete and propose six basic emotions: happiness, sadness, anger, fear, surprise, and disgust [13]. The other is to use multiple dimensions to express emotions, such as the emotion wheel of Plutchik [14] and the titer-wake scale of Russell [15]. The emotion model in this article is Russell's titer-wake scale, as shown in Fig.6.
3.3. Data Pre-processing

In order to improve the accuracy of emotion recognition, 60s experimental data minus pre-trial baseline data for the first 3s to remove the baseline interference. The 3s pre-trial baseline data is divided into 3 segments with a sliding window of 1s size, and then the 3s pre-trial baseline data is averaged. Finally, the 60s experimental data divided into 60 segments are subtracted from the pre-trial baseline data average, as shown in eq.11:

$$m'_j = m_j - \frac{\sum_{i=1}^{n} b_i}{n}$$

Among them, $m'_j$ is the $j^{th}$ pre-processed data obtained by subtracting the pre-trial baseline from the raw EEG signals, $m_j$ is the $j^{th}$ 60s experimental raw EEG data, and $b_j$ is the $i^{th}$ data of the pre-trial baseline data.

3.4. Results and Analysis

This section mainly applies the method of this paper to the DEAP dataset for emotion recognition of multi-modal physiological signals. In this paper, we propose a convolutional recurrent neural network based method to extract the spatial features of EEG signals and the temporal representations of peripheral physiological signals, and then, both of the two features are integrated to recognize emotional states. Finally, the accuracy of emotion recognition is used to show the difference between multi-modal physiological signal emotion recognition and single-modal physiological signal recognition.

Fig.7 describe the accuracy of emotion recognition that use multi-modal physiological signals, EEG, and peripheral physiological signals from 32 participants in the two emotion dimensions of Valence and Arousal. Among them, we recorded the accuracy of emotion recognition of multi-modal physiological signals, EEG signals and PERI in square-point curve, circular-point curve and triangular-point curve. In
the two sentiment dimensions of Arousal and Valence, our experiments conducted on the open source dataset DEAP show that, this method achieve 89.68% and 89.19% average accuracy in the EEG emotion classification, 63.06% and 62.41% average accuracy in the peripheral physiological signal emotion classification, 93.06% and 91.95% average accuracy in the multi-modal physiological signal emotion classification. The experimental results show that in the process of emotion recognition, peripheral physiological signals can supplement the EEG signal to achieve better emotional recognition.

Figure 7. (a) Classification results of Valence emotional dimension; (b) Classification results of Arousal emotional dimension.
Figure 8. Performance comparison between relevant approaches.

As shown in Fig.8, this paper compares related studies, which are based on DEAP datasets. **García.** Uses a Gaussian process hidden variable model (GP-LVM) to learn hidden space representation [16]. This model maps high-dimensional data in low-dimensional hidden space, and finally uses support vector machine (SVM) to identify emotions; **Chen.** proposed a novel multimodal feature fusion method, which uses Hidden Markov Model (HMM) and multimodal feature set to identify emotions [17]; **Tang.** developed the Bimodal-LSTM model to perform emotion recognition tasks using EEG signals and surrounding physiological signals, and obtained the latest results with an average accuracy rate of 83.53% [18]; **Liu.** Applied a Bimodal Deep Auto Encoder (BDAE) to classify emotions on EEG and eye movement signals, with an average accuracy rate of 82.85% [19].

4. Conclusion

In this paper, we propose a convolutional recurrent neural network based method for multi-modal physiological signal emotion recognition task. The method used CNN to learn the subtle spatial representations of multi-channel EEG signals and the LSTM to learn the temporal representations of PERI. The two features are combined for emotion recognition and classification. The method was tested on DEAP dataset. In the Arousal emotion dimension, the average emotion recognition accuracy is 93.06%; In the Valence emotion dimension, the average emotion recognition accuracy is 91.95%. The experimental results show that compared with single-modal EEG signal emotion recognition, the accuracy rate of multimodal physiological signal emotion recognition of the model is improved, which proves the reliability of the model.

In the process of multimodal physiological signal emotion recognition, the signals of all 32 channels EEG and 8 channels PERI were all selected. The experimental results show that the emotional recognition accuracy of 8 channels PERI is low. Choosing the appropriate EEG channels and peripheral physiological signals for emotion recognition requires further research.

Acknowledgments

This work was supported by National Natural Science Foundation of China [grant No.61871433], Guangdong Excellent Young Teacher Training Program of China [grant No.YQ2015046] and Pearl River Nova Program of Guangzhou [grant No.201610010199].
References

[1] NIE Dan, WANG Xiaowei and DUAN Ruonan, et al, “A Survey on EEG Based Emotion Recognition”, Chinese Journal of Biomedical Engineering, 31, 4, 595 - 606 (2012).

[2] Kim J and André E, “Emotion recognition based on physiological changes in music listening”, IEEE transactions on pattern analysis and machine intelligence, 30, 12, 2067 - 2083 (2008).

[3] Fan Y, Lu X and Li D, et al, “Video-based emotion recognition using CNN-RNN and C3D hybrid networks”, in the 18th ACM International Conference on Multimodal Interaction (2016), pp. 445 - 450.

[4] Valstar M F, Gunes H and Pantic M, “How to distinguish posed from spontaneous smiles using geometric features”, in the 9th international conference on Multimodal interfaces (2007), pp. 38 – 45.

[5] Petrantonakis P.C and Hadjileontiadis L.J, “Emotion recognition from EEG using higher order crossings”, IEEE Transactions on Information Technology in Biomedicine, 14, 2, 186 - 197 (2009).

[6] Li K, Li X, Zhang Y, et al, “Affective state recognition from EEG with deep belief networks”, in International Conference on Bioinformatics and Biomedicine (IEEE, 2013), pp. 305 - 310.

[7] Kreibig S D, “Autonomic nervous system activity in emotion: a review”, Biological Psychology, 84, 3, 394 - 421 (2010).

[8] Torres-Valencia C A, Garcia-Arias H F and Lopez M A A, et al, “Comparative analysis of physiological signals and electroencephalogram (eeg) for multimodal emotion recognition using generative models”, in XIX Symposium on Image, Signal Processing and Artificial Vision (IEEE, 2014), pp. 1 - 5.

[9] Mingli Fan, “Feature Extraction and Classification of Motor Imagery EEG Signals based on Convolutional Neural Network”, M.A. thesis, Zhengzhou University, 2017.

[10] WANG Yanna and SUN Bingyu, “Cigarette Craving EEG Classification Based on Convolution Neural Networks”, Computer Systems & Applications, 26, 6, 254 - 258 (2017).

[11] Bashivan P, Rish I and Yeasin M, et al, “Learning representations from EEG with deep recurrent-convolutional neural networks”, arXiv:1511.06448 (2015).

[12] Koelstra S, Muhl C and Soleymani M, et al, “Deap: A database for emotion analysis; using physiological signals”, IEEE transactions on affective computing, 3, 1, 18 - 31 (2011).

[13] Ekman P, Friesen W V and O'sullivan M, et al, “Universals and cultural differences in the judgments of facial expressions of emotion”, Journal of personality and social psychology, 53, 4, 712 (1987).

[14] Plutchik R, “The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice”, American scientist, 89, 4, 344 - 350 (2001).

[15] Russell J A, “A circumplex model of affect”, Journal of personality and social psychology, 39, 6, 1161 (1980).

[16] García H F, Álvarez M A and Orozco Á A, “Gaussian process dynamical models for multimodal affect recognition”, in 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE, 2016), pp. 850 - 853.

[17] Chen J, Hu B, Xu L, et al, “Feature-level fusion of multimodal physiological signals for emotion recognition”, in International Conference on Bioinformatics and Biomedicine (IEEE, 2015), pp. 395 - 399.

[18] Tang H, Liu W and Zheng W L, et al, “Multimodal emotion recognition using deep neural networks”, in International Conference on Neural Information Processing (Springer, Cham, 2017), pp. 811 - 819.

[19] Liu W, Zheng W L and Lu B L, “Emotion recognition using multimodal deep learning”, in International conference on neural information processing (Springer, Cham, 2016), pp. 521 - 529.