International Journal of Computational Intelligence Systems
Vol. **12(1)**; 2019, *pp.* 351–358
DOI: https://doi.org/10.2991/ijcis.2019.125905651; ISSN: 1875-6891; eISSN: 1875-6883
https://www.atlantis-press.com/journals/ijcis/

Multimodal Emotion Recognition Method Based on Convolutional Auto-Encoder

Jian Zhou[1,2], Xianwei Wei[1,2], Chunling Cheng[1,2,*], Qidong Yang[1,2], Qun Li[1,2]

[1] *College of Computer, Nanjing University of Posts and Telecommunications*
[2] *Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, No. 66 Xinmofan Road, Nanjing, Jiangsu, 210003, China*

ARTICLE INFO

Article History
Received 10 July 2018
Accepted 28 Oct 2018

Keywords
Emotion recognition
Convolutional auto-encoder
Fully connected neural network
EEG signals
EP signals

ABSTRACT

Emotion recognition is of great significance to computational intelligence systems. In order to improve the accuracy of emotion recognition, electroencephalogram (EEG) signals and external physiological (EP) signals are adopted due to their perfect performance in reflecting the slight variations of emotions, wherein EEG signals consist of multiple channels signals and EP signals consist of multiple types of signals. In this paper, a multimodal emotion recognition method based on convolutional auto-encoder (CAE) is proposed. Firstly, a CAE is designed to obtain the fusion features of multichannel EEG signals and multitype EP signals. Secondly, a fully connected neural network classifier is constructed to achieve emotion recognition. Finally, experiment results show that the proposed method can improve the accuracy of emotion recognition obviously compared with other similar methods.

© *2019 The Authors*. Published by Atlantis Press SARL.
This is an open access article distributed under the CC BY-NC 4.0 license (http://creativecommons.org/licenses/by-nc/4.0/).

1. INTRODUCTION

An emotion is a mental and physiological state which results from many senses and thoughts [1]. Usually, positive emotions include satisfaction and excitement whereas negative emotions include depression and wrath. Positive and negative emotions both have visible influence on individuals' behaviors. Consequently, emotion recognition can be used to assist computational intelligence systems. For example, in a medical emergency system [2], if the emotions of patients, especially of those with communication barriers, can be identified, the optimal emergency strategy can be selected, which makes the whole system more effective.

There are many researchers who pay their attention to EEG signals due to the ability of signals to reflect human emotions intuitively [3]. Compared with a single-channel EEG signal, multichannel EEG signals make the recognition accuracy higher and are preferred by researchers for emotion recognition. For example, Atkinson *et al.* [4] use the band-pass filter to extract features from multichannel EEG signals, then reduce the dimension of the extracted features by the minimum redundancy-maximum-relevance (mRMR) method. After that the features are used to identify emotions in terms of valence and arousal by the support vector machine (SVM). Ghare *et al.* [5] obtain four different frequency bands from multichannel EEG signals by the discrete wavelet transform. Then they extract statistical features and energy features from each of the four frequency bands. After that emotions are identified by SVM in terms of valence and arousal. Zhuang *et al.* [6] utilize the empirical mode decomposition (EMD) to extract features from multichannel EEG signals. EMD can represent a signal in the form of a sum of multiple intrinsic mode functions (IMFs) and a residual signal, which enables researchers to extract statistical features from every IMF. In the end, they carry out emotion classification by SVM in terms of valence. Compared with wavelet transform, EMD can extract features from signals in different frequencies more efficiently, however, the effective information of EEG signals is not completely extracted because of the existence of residual signal.

To further improve the accuracy, some researchers add EP signals as supplements to recognize emotions. [7] Accordingly, several multimodal feature fusion methods are proposed. Generally, EP signals consist of multiple types such as electrooculogram (EOG), galvanic skin response (GSR), skin temperature (ST), and respiratory belt (RB). The multimodal feature fusion methods can combine EEG signals with EP signals to produce fusion features. Liu *et al.* [8] adopt a single-type EP signal and a single-channel EEG signal for emotion recognition. An auto-encoder (AE) based on Restricted Boltzmann Machine (RBM) is used to fuse the two kinds of signals, and then SVM is applied to recognize emotion in terms of valence. Yin *et al.* [9] also conduct emotion recognition with a single-type EP signal and a single-channel EEG signal. The difference is that the signals are normalized at first and an AE based on fully connected neural networks (FCNNs) is used for feature fusion. In the end, they adopt the FCNN classifier to recognize emotions in terms of valence. However, the AE mentioned above can only perform feature fusion of a single-type EP signal and a single-channel EEG signal. In contrast, the convolutional auto-encoder (CAE) can achieve multimodal feature fusion of various inputs from the convolutional neural networks (CNNs) which has been successfully applied in image recognition [10–13]. In this paper, CAE is used for emotion recognition.

* *Corresponding author. Email:* chengcl@njupt.edu.cn

In order to obtain the fusion features of multichannel EEG signals and multitype EP signals to improve the accuracy of emotion recognition, this paper proposes a new multimodal emotion recognition method based on CAE. First of all, a CAE is designed to obtain the fusion features of multichannel EEG signals and multitype EP signals. After that, a FCNN classifier is constructed. Finally, the fusion features are fed to the FCNN classifier to obtain emotion classification results.

The rest of the paper is structured as follows: the dataset for emotion analysis and the concept of CAE are introduced in Section 2. In Section 3, the multimodal emotion recognition method based on CAE is proposed. In Section 4, experiments and analysis are presented, including parameters selection, emotion recognition result, and comparison analysis. Conclusion are given in Section 5.

2. RELATED WORK

2.1. Database for Emotion Analysis Using Physiological Signals

Koelstra *et al.* [14] adopt music videos to induce participants to generate emotion variations, then use electrode caps to collect EEG signals of different cerebral cortex channels. They arrange those signals and then publish a database for emotion analysis using physiological signals (DEAP). Firstly, they select 32 participants to get involved in the data collecting experiments and select 40 videos which induce participants to generate emotion variations. Each video is one minute long. Second, each participant is required to wear a data collecting device which is able to collect EP signals of 8 types and EEG signals from 32 different channels. Three seconds before the video is played, the device starts to record EP signals and EEG signals when the participant is still in a calm mood. Additionally, the researchers collect each EEG signal for 63 seconds for down-sampling at a frequency of 128 Hz. Finally, they label each participant's emotion in terms of four indicators such as valence, arousal, dominance, and liking as the feedback from a specific participant who has watched the videos. Table 1 is an example of data for one participant.

Table 1 DEAP representation for each participant.

Name	Shape	Contents
Data	$40 \times (32 + 8) \times 63 \times 128$	Video × (channel +T type) × seconds × sampling frequency
Labels	40×4	Video × (valence, arousal, dominance, liking)

DEAP, database for emotion analysis using physiological signals.

2.2. Convolutional Auto-Encoder

Rumelhart *et al.* [15] first propose the concept of AE and employ it to process data with large dimensions. AE mainly consists of two components, the encoder and the decoder, which can be respectively realized by different neural network. On the basis of AE, Masci *et al.* [16] realize the encoder with CNN and propose CAE which adopts an unsupervised learning algorithm for encoding. CAE has achieved good results for unsupervised feature extraction in recent years [17, 18]. As shown in Figure 1 which presents the structure of CAE, the encoder codes the input signals to extract and fuse their features. The decoder does the opposite reconstruction work to reconstruct the input signals. Usually, it is the trained encoder which is actually used and the decoder is only used for joint training with the encoder [6, 16, 19]. Similarly in this paper, the decoder is only used for training while the encoder is used to obtain the fusion features of EEG signals and EP signals.

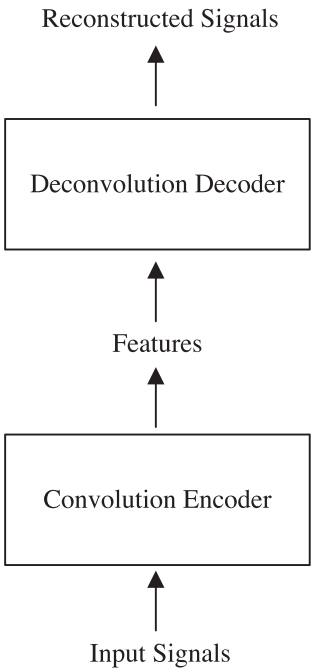

Figure 1 Structure of convolutional auto-encoder (CAE).

3. MULTIMODAL EMOTION RECOGNITION METHOD BASED ON CAE

Firstly, the framework of the multimodal emotion recognition method based on CAE is introduced. Then the two important stages of the proposed method are presented respectively: feature fusion based on CAE and emotion classification based on FCNN.

3.1. Framework of the Proposed Method

The framework of the proposed method is shown in Figure 2. In Figure 2, $Input = \{S_{EEG}, S_{EP}\}$ consists of EEG signals from n channels and EP signals of m types, where $S_{EEG} = \{sig_{1st\ EEG}, \cdots sig_{ith\ EEG}, \cdots sig_{nth\ EEG}\}$ and $S_{EP} = \{sig_{1st\ EP}, \cdots sig_{jth\ EP}, \cdots sig_{mth\ EP}\}$, $Input \epsilon R^{t \times f \times (n+m)}$. To be specific,

$$sig_{ith\ EEG} = \begin{Bmatrix} S_{EEG_1^1}, S_{EEG_2^1}, \cdots, S_{EEG_f^1} \\ S_{EEG_1^2}, S_{EEG_2^2}, \cdots, S_{EEG_f^2} \\ \cdots \quad \cdots \quad \cdots \quad \cdots \\ S_{EEG_1^t}, S_{EEG_2^t}, \cdots, S_{EEG_f^t} \end{Bmatrix}$$ represents the EEG signal from the *ith* channel and $$sig_{jth\ EP} = \begin{Bmatrix} S_{EP_1^1}, S_{EP_2^1}, \cdots, S_{EP_f^1} \\ S_{EP_1^2}, S_{EP_2^2}, \cdots, S_{EP_f^2} \\ \cdots \quad \cdots \quad \cdots \quad \cdots \\ S_{EP_1^t}, S_{EP_2^t}, \cdots, S_{EP_f^t} \end{Bmatrix}$$ represents the EP signal of the *jth* type. $s_{EEG_f^t}$ and $s_{EP_f^t}$, respectively represent the EEG signal value and the EP signal value when at the time t and the sampling point f.

Moreover, F in Figure 2 denotes the fusion features of EEG signals and EP signals, which is obtained by feeding S_{EEG} and S_{EP} to the trained encoder of CAE. $\overline{C}$ in Figure 2 denotes the emotion recognition result, which is obtained by feeding F to the trained FCNN classifier. In this paper, emotions are divided into four categories, denoted as $\overline{C} \epsilon \{c_1, c_2, c_3, c_4\}$ which are satisfaction, depression, excitement, and wrath, respectively.

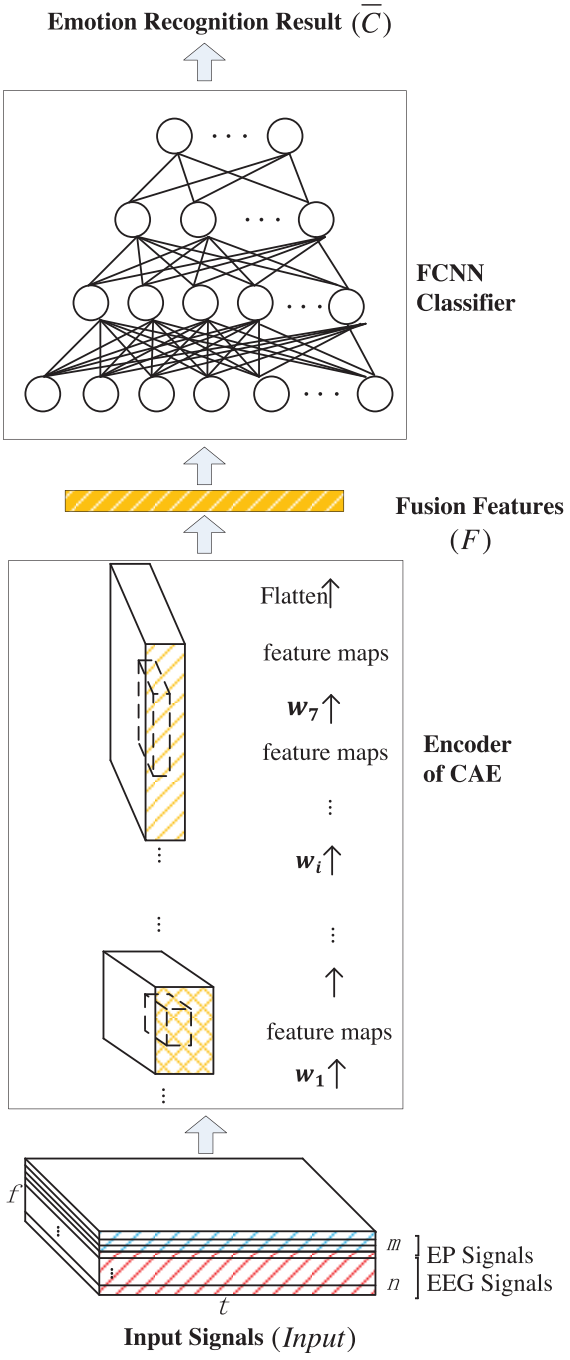

Figure 2 | Framework of multimodal emotion recognition method based on convolutional auto-encoder (CAE).

3.2. Feature Fusion Based on CAE

The procedure of CAE used in this paper is shown in Figure 3. It can be seen from Figure 3 that a seven-layer CNN is adopted in the encoder and $n + m$ FCNNs are used in the decoder. The encoder is used to extract and obtain the fusion features from the input signals, while the decoder is used to obtain the reconstructed signals. By comparing the original input signals and the reconstructed signals, CAE can get trained. The trained encoder is actually used to obtain the fusion features of multichannel EEG signals and multitype EP signals. CAE used in this paper is introduced with respect to its encoder, decoder, and training.

3.2.1. Encoder

The encoder of CAE is constructed by a seven-layer CNN. The output of each layer is a feature map which is calculated through convolution operations as Formula (1) and Formula (2).

$$map_1 = Input * w_1 \tag{1}$$

$$map_i = map_{i-1} * w_i, \quad i\epsilon\,[2, 7] \tag{2}$$

In particular, * represents a convolution operation. $W = \{w_1, w_2, \cdots w_i \cdots w_7\}$ is the weight set of CNN, where w_i represents the weight of the *ith* layer. $Map = \{map_1, map_2, \cdots map_i \cdots map_7\}$ is the output set, where map_i represents the output of the *ith* layer. The output of the last layer, map_7, is converted into the encoded sequence via Formula (3).

$$F = \text{Flatten}\,(map_7) \tag{3}$$

The convolution operation depends on the size and number of convolution kernel. Consequently, different size and number of convolution kernel may lead to different computational complexity and feature fusion effect. The 3×3 convolution kernel is adopted in the first layer and second layer to extract features from EEG signals and EP signals, for that the 3×3 kernel can obtain the same feature extraction effect with fewer parameters [20, 21] comparing to larger convolution kernels such as the 5×5 or the 9×9 convolution kernel. The 1×1 convolution kernel is accordingly adopted in the third layer to fuse EEG signals and EP signals to produce the preliminary fusion features, as shown in Figure 4, for that the 1×1 convolution kernel is frequently used to fuse features [22, 23]. In Figure 4, the blue balls and red balls, respectively, represent the extracted features of EP signals and EEG signals, while the yellow balls represent the preliminary fusion features. After that,

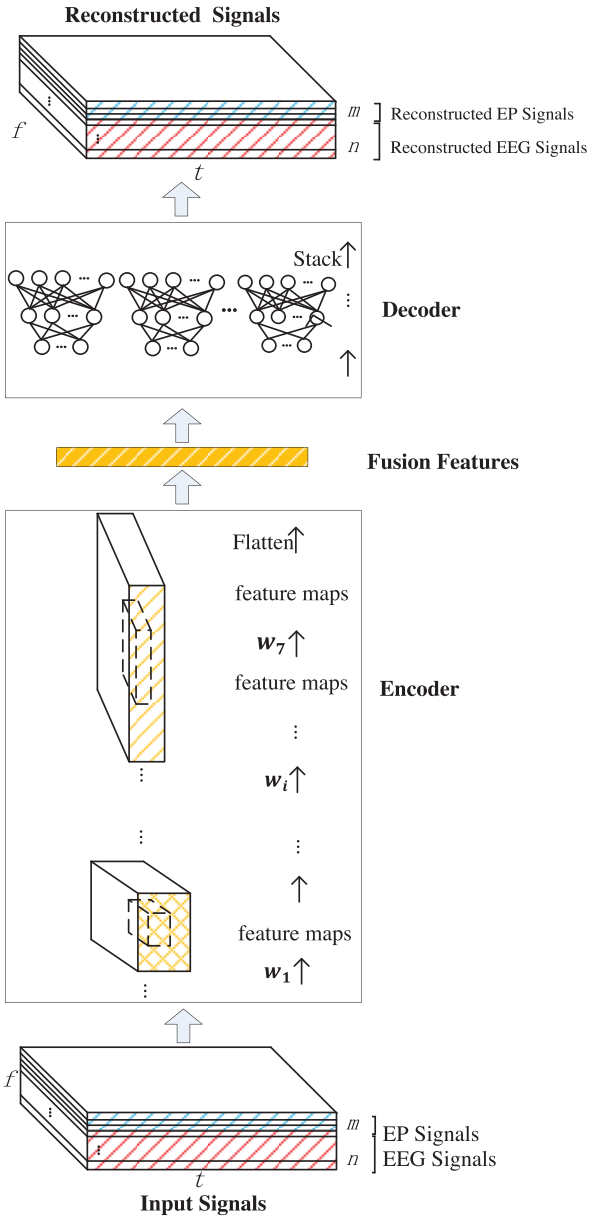

Figure 3 | Procedure of convolutional auto-encoder (CAE).

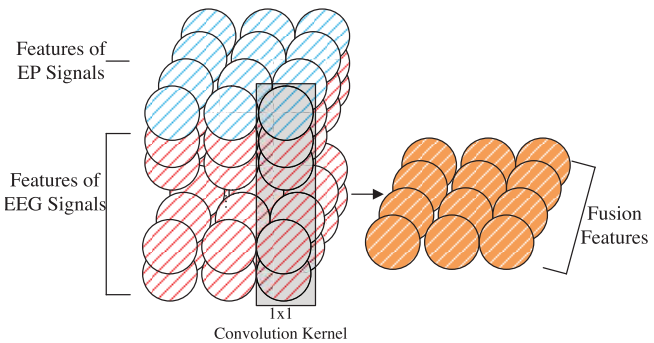

Figure 4 | Example of feature fusion.

the fourth to sixth layers perform the same operation to produce the final fusion features as the first to third layers to produce the preliminary fusion features. Finally, based on the actual size of feature map outputted by the upper layer, the 2×7 convolution kernel, which is conducive to transform the final fusion features into the form of encoded sequence, is used in the seventh layer. The size of convolution kernel adopted in this paper is shown in Table 2, and the number of convolution kernel will be discussed in Section 4.1.1.

Table 2 | Size of convolution kernel.

Layer	Convolution Kernel size
First layer	3×3
Second layer	3×3
Third layer	1×1
Fourth layer	3×3
Fifth layer	3×3
Sixth layer	1×1
Seventh layer	2×7

3.2.2. Decoder

The decoder of CAE is constructed by $n + m$ FCNNs. Each FCNN with F as its input is used to output a reconstructed single-channel EEG signal or a reconstructed single-type EP signal.

Each FCNN has a hidden layer H and obtains its output via Formula (4). $\Psi = \{\Psi_1, \Psi_2\}$ is the weight set of FCNN in the decoder, where Ψ_1 represents the weight from the input layer to the hidden layer and Ψ_2 represents the weight from the hidden layer to the output layer. After that, the output of FCNN is transformed via Formula (5), where $Rsig_{ith\ EEG}$ represents the reconstructed EEG signal from the *ith* channel and $Rsig_{jth\ EP}$ represents the reconstructed EP signal of the *jth* type. Finally, the reconstructed signals $Rec = \{RS_{EEG}, RS_{EP}\}$ can be obtained by appending the outputs of all FCNNs, where $RS_{EEG} = \{Rsig_{1st\ EEG}, \cdots Rsig_{ith\ EEG}, \cdots Rsig_{nth\ EEG}\}$ and $RS_{EP} = \{Rsig_{1st\ EP}, \cdots Rsig_{jth\ EP}, \cdots Rsig_{mth\ EP}\}$, $Rec \in \mathrm{R}^{t \times f \times (m+n)}$.

$$\begin{cases} H = F\Psi_1 \\ output = H\Psi_2 \end{cases} \tag{4}$$

$$Rsig_{ithEEG} | Rsig_{jthEP} = \mathrm{Stack}\left(output\right) \tag{5}$$

3.2.3. Training

In the CAE training phase, the mean square error (MSE) of *Input* and *Rec* is used as the loss function, denoted as Formula (6). By optimizing the loss function with the stochastic gradient descent (SGD) [24], the optimal weight set W of CNN in the encoder and the optimal weight set Ψ of FCNN in the decoder can be finally obtained.

$$\min_{\mathrm{W},\Psi} \mathrm{E}\left(Input - Rec\right)^2 \tag{6}$$

The fusion features F, which will be used as the input of the classifier to perform emotion recognition, can be obtained by the trained encoder of CAE.

3.3. Emotion Classification Based on FCNN

In order to accurately identify emotions, a FCNN classifier is designed in this paper. The structure of FCNN for emotion classification is shown in Figure 5. It can be seen from Figure 5 that the FCNN classifier has an input layer, two hidden layers and an output layer. The input of FCNN classifier is the fusion features F which is obtained by the encoder of CAE. The output of the FCNN classifier is the emotion recognition result $\overline{C}$. The FCNN classifier used in this paper is introduced with respect to its forward propagation and training.

3.3.1. Forward propagation

$\overline{C}$ is obtained through Formula (7), where $\Phi = \{\varphi_1, \varphi_2, \varphi_3\}$ represents the weight set of the FCNN classifier. The hidden layers, $\mathcal{H}_1$ and $\mathcal{H}_2$, use *sigmoid* () as their activation functions and the output layer uses s*oftmax* () as its activation function.

$$\begin{cases} \mathcal{H}_1 = sigmoid\,(F\varphi_1) \\ \mathcal{H}_2 = sigmoid\,(\mathcal{H}_1\varphi_2) \\ \overline{C} = softmax\,(\mathcal{H}_2\varphi_3) \end{cases} \tag{7}$$

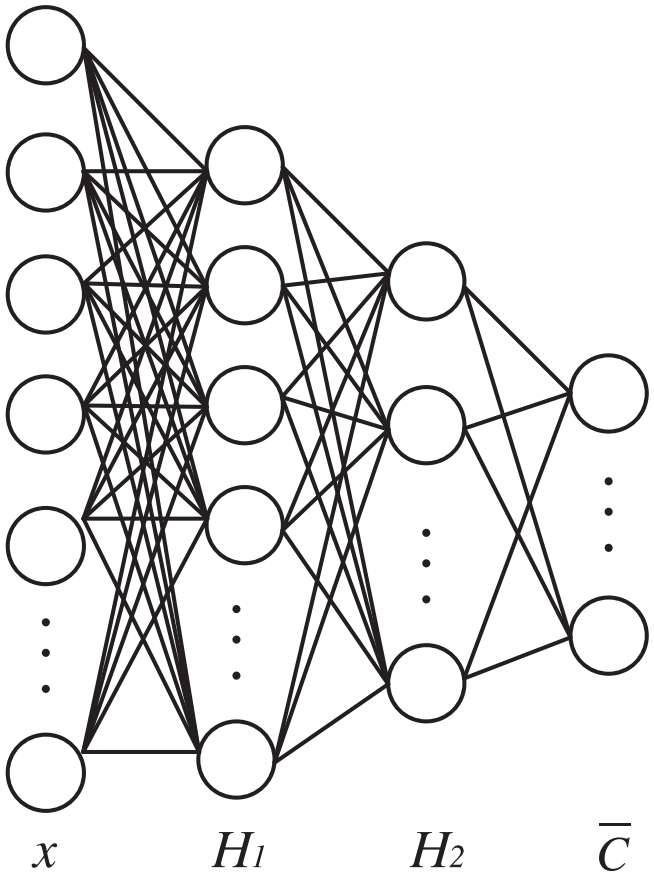

Figure 5 | Structure of fully connected neural network (FCNN) for emotion classification.

3.3.2. Training

In the FCNN classifier training phase, suppose there are m samples, where $\{(x_1, c_1), (x_2, c_2), \ldots (x_i, c_i) \ldots (x_m, c_m)\}$. The cross-entropy loss is used as the loss function, denoted as Formula (8). Particularly, $\overline{c_i}$ is the output of the FCNN classifier with x_i as its input, and c_i is the actual emotion corresponding to x_i. The loss function is optimized by SGD with the fixed parameters of CAE, then the optimal weight set Φ of the FCNN classifier can be obtained. The trained FCNN classifier is used to classify emotions.

$$\min_{\Phi} -\frac{1}{m}\sum_{i=1}^{m} c_i \log \overline{c_i} + (1 - c_i) \log\left(1 - \overline{c_i}\right) \tag{8}$$

4. EXPERIMENTS AND ANALYSIS

The proposed method is implemented with neural network framework Keras and TensorFlow in Python, and is tested on DEAP.

In order to obtain clear emotion labels, this paper uses k-means algorithm [25] which is a simple and fast clustering algorithm, to cluster the samples in DEAP in terms of valence and arousal. For a given sample set, k-means algorithm divides the sample set into k clusters according to the distance between samples. The goal of this algorithm is to make the distance between samples in clusters as small as possible and the distance between clusters as large as possible. The clustering result in this paper is shown in Figure 6, where the vertical axis represents valence and the horizontal axis represents arousal. These clustering results, $\{c_1 =$ Satisfaction, $c_2 =$ *Depression*, $c_3 =$ *Excitement*, $c_4 =$ *Wrath* $c_4 =$ *"Wrath"*$\}$, can describe emotion states in a more objective and clear way as the labels of samples.

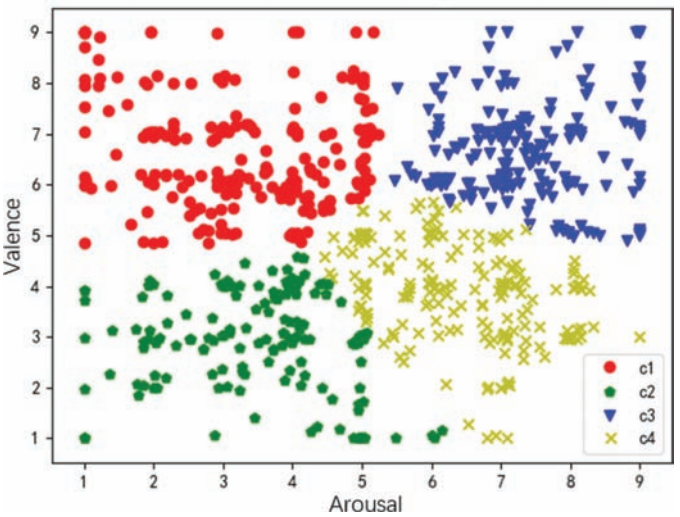

Figure 6 | Clustering result in terms of valence and arousal.

4.1. Parameters Selection

The proposed method requires a process of parameter adjustment to achieve better results. Therefore, different experiments are carried out to select the appropriate parameters. The number of convolution kernel in the encoder of CAE and the learning rate in the FCNN classifier are mainly discussed in the following parts:

4.1.1. Number of convolution kernel

Different number of convolution kernel can extract different granularity features from EEG signals and EP signals. However, more convolution kernels doesn't mean better feature extraction effect. Too many convolution kernels result in too many input features of the FCNN classifier, which makes it difficult for the classifier to learn useful or important features. On the contrary, too few convolution kernels may lead to insufficient input features, which decreases the ability of the classifier to recognize emotions.

In this paper, several experiments are carried out on the number of convolution kernel. The results of the five experiments are shown in Table 3. It can be seen that the accuracy of training set and the accuracy of test set are on the rise in the first three cases. This is because that as the number of convolution kernel increases, the number of effective extracted features increases. In the last two cases, the accuracy of training set still increases, but the accuracy of test set decreases. This is because as the number of convolution kernel increases, the number of input features of the FCNN classifier still increases, resulting in over fitting of the classifier. Therefore, the number of convolution kernel in Case 3 is adopted as an optimization in this paper.

Table 3 Influence of the number of convolution kernels on the accuracy.

	Case 1	Case 2	Case 3	Case 4	Case 5
First layer	16	32	**64**	64	128
Second layer	16	32	**128**	128	256
Third layer	32	64	**128**	256	512
Fourth layer	32	64	**256**	512	1024
Fifth layer	64	128	**256**	512	1024
Sixth layer	64	128	**512**	1024	2048
Seventh layer	128	256	**1024**	1024	2048
Accuracy of training set	0.703	0.934	**0.965**	0.99	0.99
Accuracy of test set	0.659	0.874	**0.920**	0.679	0.503

4.1.2. Learning rate

The learning rate is a parameter in the FCNN classifier training phase. The size of learning rate can directly affect whether FCNN converges. If the learning rate is too high, FCNN will be unable to converge. Conversely, a too low learning rate can cause FCNN to converge very slowly and fall into a local optimum. Therefore, choosing an appropriate learning rate can make the loss decrease rapidly in the initial stage of training, and make FCNN obtain the global optimal solution.

This paper uses the exponentially decayed learning rate method [26] for training the FCNN classifier. Table 4 shows the influence of initial learning rate and the decay on both the steps and the loss. It can be seen from Table 4 that when the initial learning rate is 0.1, FCNN cannot converge due to the high learning rate. When the initial learning rate is 0.001, the optimal decay is 0.96. In this case, FCNN converges around 3000 training steps, and the loss fluctuates within 30–50 range. When the initial learning rate is 0.01, the optimal decay is still 0.96. This is the best case in which FCNN converges around 2000 training steps and the loss is the minimum.

Table 4 Influence of learning rate on the steps and loss.

Initial Learning Rate	Decay	Steps	Loss
0.1	0.7	No fitting	370–450
0.1	0.8	No fitting	300–400
0.1	0.9	No fitting	300–400
0.1	0.96	No fitting	100–230
0.01	0.7	50 000	170–200
0.01	0.8	50 000	80–100
0.01	0.9	40 000	30–75
0.01	**0.96**	**20 000**	**10–20**
0.001	0.7	40 000	370–450
0.001	0.8	40 000	200–300
0.001	0.9	40 000	100–150

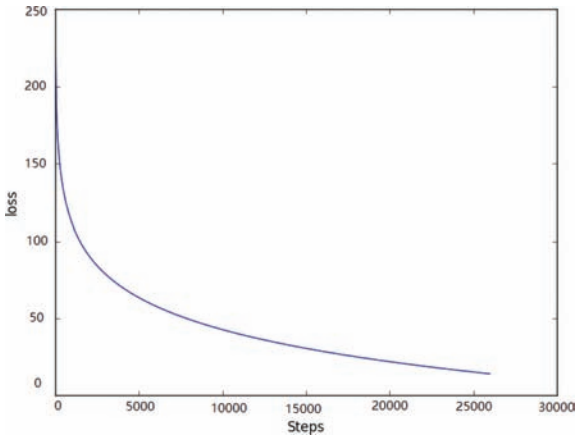

Figure 7 Loss varies with steps.

The initial learning rate and the decay in the FCNN classifier are set to 0.01 and 0.96, respectively. The loss of FCNN is smoothed and then plotted in Figure 7. Figure 7 shows the loss decreases rapidly at the first 5000 training steps. This is because the relatively high initial learning rate can make FCNN converge rapidly. As the number of steps increases, the learning rate gradually decreases. From 10 000 to 20 000 training steps, the loss decreases slowly and finally converges to the range of 10–20. Therefore, the parameters selected in this paper enable FCNN to converge quickly to the optimal solution.

4.2. Emotion Recognition Result

The experiment settings of multimodal emotion recognition method based on CAE are shown in Table 5. The bottom row is the input signals. The seven rows above it are the settings of the seven-layer CNN used as the encoder of CAE. It should be noted that the decoder is only used for the training of CAE, so the settings of decoder are not in Table 5. The first three rows are the settings of the FCNN classifier.

Each sample used in this experiment is denoted as $(Input, c)$. $Input = \{S_{EEG}, S_{EP}\}$ consists of EEG signals from 14 channels, $S_{EEG} = \{sig_{1st\ EEG}, \cdots sig_{ith\ EEG}, \cdots sig_{14th\ EEG}\}$, and EP signals of 3 types including GSR, ST and RP, $S_{EP} = \{sig_{1st\ EP}, sig_{2nd\ EP}, sig_{3rd\ EP}\}$. $Input \epsilon R^{t \times f \times (n+m)}$, specifically $t = 60$, $f = 128$, $n = 14$ and $m = 3$. $c \epsilon \{c_1, c_2, c_3, c_4\}$ is the emotion label. A total of 1280 samples are used in the experiment.

Table 5 Experiment settings of multimodal emotion recognition method based on CAE.

Type	Configurations
Output emotion	Hidden unit:4, activation:softmax
Fully connection	Hidden unit:64, activition:sigmoid
Fully connection	Hidden unit:512, activition: sigmoid
Convolution	Kernel size:2 × 7, stride:2 × 2, padding:1
Convolution	Kernel size:1 × 1, stride:1 × 1, padding:0
Convolution	Kernel size:3 × 3, stride:2 × 2, padding:1
Convolution	Kernel size:3 × 3, stride:2 × 2, padding:1
Convolution	Kernel size:1 × 1, stride:1 × 1, padding:0
Convolution	Kernel size:3 × 3, stride:2 × 2, padding:1
Convolution	Kernel size:3 × 3, stride:2 × 2, padding:1
Input signals	60 × 128 × (14 + 3)

CAE, convolutional auto-encoder.

The samples are separated into eight parts at random in this paper. The cross-validation is performed eight times. For each time, one part of the samples is selected in turn as the validation set with the rest as the training set. The accuracy of each validation is shown in Figure 8. It can be seen from Figure 8 that the highest accuracy of emotion recognition is 98.4%, the lowest accuracy of emotion recognition is 84.7%. In summary, the average accuracy of emotion recognition is 92.07%.

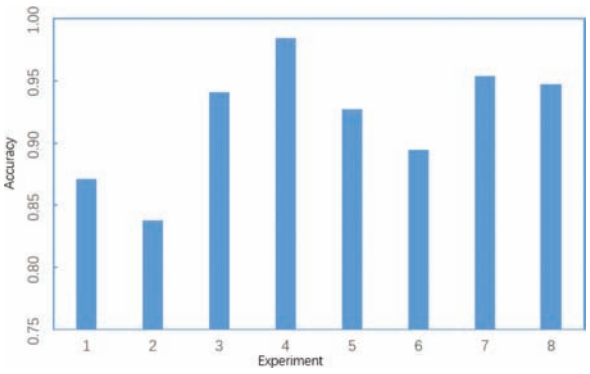

Figure 8 Accuracies of 8 validations.

4.3. Comparison Analysis

Considering the average accuracy of emotion recognition as the performance indicator, the proposed method is compared with other five similar methods. The result is shown in Table 6. Method 1 uses mRMR to extract features from multichannel EEG signals, then SVM is used to perform binary classification in terms of valence and arousal. Method 2 adopts EMD to extract features from multichannel EEG signals and SVM serves as a classifier in terms of valence and arousal. Method 3 uses AE to extract features from an EOG signal and a single-channel EEG signal, after that uses SVM to classify emotions in terms of valence. Method 4 uses RBM to extract features from an EOG signal and a single-channel EEG signal, then uses an FCNN classifier to identify emotions. Method 5 uses CNN to extract features from multichannel EEG signals and various EP signals, then uses a FCNN classifier to perform binary classification.

From Table 6, it is obvious that mRMR used in Method 1 and EMD adopted in Method 2 have constrained the increase of accuracy because the features extracted from EEG signals are insufficient and there are no EP signals used for supplement. Compared with the first two methods, Methods 3 and 4 conduct feature fusion of a single-channel EEG signal and a single-type signal by using AE or RBM so that the accuracy is improved. The method proposed in this paper distinguishes from these two types of methods by using multichannel EEG signals and multitype EP signals simultaneously. In our previous work [27], Method 5 extracts features from multichannel EEG signals and various EP signals directly via CNN without fusing the features. Compared with that, the encoder of CAE can fuse multimodal signals to generate fusion features through different convolution kernels. Therefore, using CAE, the proposed method obtains the best emotion recognition results.

In conclusion, the multimodal emotion recognition method based on CAE proposed in this paper can effectively fuse multichannel EEG signals and multitype EP signals to acquire fusion features, thereby improving the accuracy of emotion recognition.

5. CONCLUSION

Emotion recognition is able to assist computational intelligence systems. In this paper, a new multimodal emotion recognition method based on CAE is proposed to improve the accuracy. Firstly, the fusion features of multichannel EEG signals and multitype EP signals are obtained by the trained encoder of CAE. Secondly, the

Table 6 Comparison with other methods.

	Signal Type	Feature Extraction	Classifier	Accuracy (%)
1. Atkinson *et al.* [4]	EEGs	mRMR	SVM	60.7
2. Zhuang *et al.* [6]	EEGs	EMD	SVM	70.9
3. Liu *et al.* [8]	EOG, EEG	Auto-encoder	SVM	83.25
4. Yin *et al.*[9]	EOG, EEG	Restricted Boltzmann machine	FCNN	77.19
5. Cheng *et al.* [27]	EEGs, GSR, ST, RB	Convolution neural network	FCNN	83.45
6. The proposed method	EEGs, GSR, ST, RB	Convolution auto-encoder	FCNN	92.07

EEg, electroencephalogram; mRMR, minimum redundancy-maximum-relevance; EOG, electrooculogram; SVM, support vector machine; GSR, galvanic skin response; FCNN, fully connected neural network.

fusion features are fed to the trained FCNN classifier to obtain emotion classification results. Finally, experiment results on DEAP show that the average accuracy of emotion recognition is improved to 92.07% with the proposed method. In the future, the proposed method is required to test on other datasets to improve the stability.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No. 71301081, 61373139, 61572261), Natural Science Foundation of Jiangsu Province (No. BK20130877, BK20150868), Natural Science Foundation of the Higher Education Institutions of Jiangsu Province (No.17KJB520027).

REFERENCES

[1] W. James, What is an emotion?, Mind. 9(34) (1884), 188–205.
[2] D.A. Alexander, S. Klein, Ambulance personnel and critical incidents: impact of accident and emergency work on mental health and emotional well-being, Br. J. Psychiatry J. Ment. Sci. 178(1) (2001), 76–81.
[3] W.B. Cannon, The James-Lange theory of emotions: a critical examination and an alternative theory, Am. J. Psychol. 3(4) (1987), 567–586.
[4] J. Atkinson, D. Campos, Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers, Expert Syst. Appl. 47 (2016), 35–41.
[5] P.S. Ghare, A.N. Paithane, Human emotion recognition using non linear and non stationary EEG signal, in Proceedings of International Conference on Automatic Control and Dynamic Optimization Techniques, Pune, 2016, pp. 1013–1016.
[6] N. Zhuang, Y. Zeng, L. Tong, C. Zhang, H. Zhang, B. Yan, Emotion recognition from EEG signals using multidimensional information in EMD domain, BioMed Res. Int. (2017), 1–9.
[7] Q. Zhang, X. Chen, Q. Zhan, T. Yang, Respiration-based emotion recognition with deep learning, Comput. Ind. 92 (2017), 84–90.
[8] W. Liu, W.L. Zheng, B.L. Lu, Emotion recognition using multimodal deep learning, in Proceedings of International Conference on Neural Information Processing, Kyoto, 2016, pp. 2816–2831.
[9] Z. Yin, M. Zhao, Y. Wang, J. Yang, J. Zhang, Recognition of emotions using multimodal physiological signals and an ensemble deep learning model, Comput. Methods Programs Biomed. 140 (2017), 93–110.
[10] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in Proceedings of International Conference on Learning Representations, San Juan, 2016, pp. 2092–2096.
[11] R. Girshick, Fast R-CNN, in Proceedings of International Conference on Computer Vision, Santiago, 2015, pp. 1440–1448.
[12] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016, pp. 779–788.
[13] M.W. Bhatti, Y. Wang, L. Guan, A neural network approach for human emotion recognition in speech, in Proceedings of IEEE International Symposium on Circuits and Systems, Vancouver, 2004, pp. 181–184.
[14] S. Koelstra, C. Muhl, M. Soleymani, J.S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, Deap: a database for emotion analysis using physiological signals, IEEE Trans. Affect. Comput. 3(1) (2012), 18–31.
[15] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature. 323 (1986), 533–536.
[16] J. Masci, U. Meier, D. Ciresan, J. Schmidhuber, Stacked convolutional auto-encoders for hierarchical feature extraction, in Proceedings of International Conference on Artificial Neural Networks, Espoo, 2011, pp. 52–59.
[17] H. Huang, X.T. Hu, Y. Zhao, M. Makkie, Q. Dong, S. Zhao, L. Guo, T. Liu, Modeling task fMRI data via deep convolutional autoencoder, IEEE Trans. Med. Imaging. 37(7) (2018), 1551–1561.
[18] Y.Q. Wang, Z.G. Xie, K. Xu, Y. Dou, An efficient and effective convolutional auto-encoder extreme learning machine network for 3D feature learning, Neurocomputing. 174 (2016), 988–998.
[19] K.G. Dizaji, A. Herandi, C. Deng, W. Cai, H. Huang, Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization, in Proceedings of International Conference on Computer Vision, Venice, 2017, pp. 5747–5756.
[20] B.S. Hua, M.K. Tran, S.K. Yeung, Pointwise convolutional neural networks, in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018, pp. 984–993.
[21] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018, pp. 7132–7141.
[22] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in Proceedings of International Conference on Learning Representations, San Diego, 2015, pp. 1–14.
[23] B. Shi, X. Bai, C. Yao, An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, IEEE Trans. Pattern Anal. Mach. Intell. 39(11) (2016), 2298–2304.
[24] L. Bottou, Large-scale machine learning with stochastic gradient descent, in Proceedings of International Conference on Computational Statistics, Paris, 2010, pp. 177–186.
[25] T. Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, A.Y. Wu, An efficient k-Means clustering algorithm: analysis and implementation, IEEE Trans. Pattern Anal. Mach. Intell. 24(7) (2002), 881–892.
[26] C. Darken, J. Chang, J. Moody, Learning rate schedules for faster stochastic gradient search, Neural Netw. Signal Process. 44 (2002), 3–12.
[27] C.L. Cheng, X.W. Wei, J. Zhou, Emotion recognition algorithm based on convolution neural network, in Proceedings of International Conference on Intelligent Systems and Knowledge Engineering, Nanjing, 2017, pp. 1–5.