IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 27, 2020, accepted January 2, 2021, date of publication January 6, 2021, date of current version January 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3049516

Multimodal Emotion Recognition Using a Hierarchical Fusion Convolutional Neural Network

YONG ZHANG, CHENG CHENG, AND YIDIE ZHANG
School of Computer and Information Technology, Liaoning Normal University, Dalian 116081, China

Corresponding author: Yong Zhang (zhyong@lnnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772252, in part by the Natural Science Foundation of Liaoning Province of China under Grant 2019-MS-216, and in part by the Program for Liaoning Innovative Talents in University under Grant LR2017044.

ABSTRACT In recent years, deep learning has been increasingly used in the field of multimodal emotion recognition in conjunction with electroencephalogram. Considering the complexity of recording electroencephalogram signals, some researchers have applied deep learning to find new features for emotion recognition. In previous studies, convolutional neural network model was used to automatically extract features and complete emotion recognition, and certain results were obtained. However, the extraction of hierarchical features with convolutional neural network for multimodal emotion recognition remains unexplored. Therefore, this paper proposes a hierarchical fusion convolutional neural network model to mine the potential information in the data by constructing different network hierarchical structures, extracting multiscale features, and using feature-level fusion to fuse the global features formed by combining weights with manually extracted statistical features to form the final feature vector. This paper conducts binary classification experiments on the valence and arousal dimensions of the DEAP and MAHNOB-HCI data sets to evaluate the performance of the proposed model. The results show that the model proposed in this paper can achieve accuracies of 84.71% and 89.00% on the two corresponding data sets, indicating that the model proposed in this paper is superior to other deep learning emotion classification models in feature extraction and fusion.

INDEX TERMS Deep learning, electroencephalogram, hierarchical convolutional neural network, multimodal emotion recognition, multiscale features.

I. INTRODUCTION

In recent years, emotion recognition has received increasing attention in many fields and is an important factor in implementing human-computer interaction systems. Emotions are complex psycho-physiological processes associated with many external and internal activities. The expression of emotions includes not only some physiological responses, such as skin temperature and heart rate, but also some nonphysiological responses, such as facial expressions and body language [1]. Previous studies have explored the brain or surrounding signals separately, but with the deepening of research, the single mode cannot provide complementary information among various modes and cannot fully express emotional states, etc. Therefore, the use of fusion of multimodal signals has gradually developed in emotion recognition [2].

Feature extraction and fusion are the key steps in multimodal emotion recognition [3]. Li *et al.* [4] proposed a hierarchical modular neural network and applied it to multimodal emotion recognition. By connecting the submodules in each module to integrate information from different patterns, the disadvantages of feature-level fusion and decision-level fusion are solved, and the performance of neural network is improved. Domínguez-Jiménez *et al.* [5] proposed a three-emotion recognition model based on physiological signals. The model used a random forest (RF) to recursively eliminate and select features and support vector machines as classifiers to calculate the time domain and frequency domain features of heart rate information and galvanic skin responses, forming the best emotion recognition model. Lu *et al.* [6] proposed a pattern learning framework based on dynamic

The associate editor coordinating the review of this manuscript and approving it for publication was Taous Meriem Laleg-Kirati.

entropy. The framework uses the dynamic entropy of quantitative electroencephalogram (EEG) measurements to extract continuous entropy values that change with time from EEG signals to achieve emotion recognition irrelevant to the subject, and the best average accuracy rate reached up to 85.11%. These methods can achieve good results, but manual feature extraction and early level fusion leads to feature redundancy and loss of key features in multiple modalities due to the large amount of EEG data. Therefore, how to effectively extract features and reduce computations is still the subject of much research [7]. Although researchers have proposed a variety of EEG feature extraction and fusion methods, these methods have problems such as high time complexity and insufficient precision [8]. Therefore, to address the above problems, many deep learning models have been widely proposed in the field of emotion recognition.

Deep learning is a method that advocates end-to-end learning. In this method, deep neural networks process the original data without any preprocessing and decompose the data into multiple levels of abstraction to automatically extract relevant features [9], which can learn high-level representation from the raw input features and effectively achieve the classification goal. Deep learning results show that automatic feature extraction performs better than manual feature extraction; and various deep learning technologies [10], including autoencoders (AEs), convolutional neural networks (CNNs) [11], and recurrent neural networks (RNNs) [12], are widely used in different domains. Among these technologies, CNNs have the ability to find robust spatial features from images, RNNs are suitable for extracting the temporal features of video and speech for classification, and AEs are more suitable for learning unsupervised features [13]. In a CNN, each CNN layer contains some relevant features that represent important information at their respective level of abstraction of the input data. As the convolution process progresses layer by layer, the initial layer extracts local and spatial features while the end layer extracts global features exclusively [14].

Multimodal emotion recognition aims to combine the predictive capabilities of individual behavioral trails and biometric features for accurate classification [15]. The challenges of multimodal emotion recognition are as follows:

- It is more complex than unimodal emotion recognition systems due to combine and model multiple modal data.
- Even within multimodal emotion recognition, there needs a high degree of predictive accuracy, which requires techniques and methods for feature extraction and fusion.

With the aim, in this study, we propose a novel hierarchical fusion convolutional neural network (HFCNN) model to construct a layered incremental architecture by setting different convolution kernel sizes and numbers on the CNN convolutional layers. We use the information hidden in the HFCNN to construct the feature representation of the multimodal signals and combine the statistical features in the time and frequency domain to improve the classification ability of the model.

The contributions of this research are as follows:

- A hierarchical fusion convolutional neural network model is proposed based on multimodal feature extraction.
- Based on the hierarchical network structure in the CNN, the weights are used to fuse hierarchical convolutional features to form a global feature vector.
- The physiological signals and emotional scores of multiple video clips are used to enhance the emotional recognition system.

The remaining sections of this paper are arranged as follows. Section II introduces the previous studies of multimodal feature extraction and classification methods using machine learning and deep learning. Section III introduces the model proposed in this paper. Section IV discusses the main content and results of the experiments in this paper, and Section V gives a summary of the experiments in this paper.

II. RELATED WORK

Existing research in the field of emotion recognition shows that there are two methods of emotion recognition: one method is to detect physiological signals and the other method is to detect emotional behaviors. Since human beings can rely on their active consciousness to cover their behaviors and existing technologies cannot elicit the true emotional state of human beings, behavior-based emotion recognition has certain limitations. Therefore, research on the use of physiological signals to recognize emotions has received increasingly more attention [16]. In view of the different information carried by the signals of different modes, in order to make an emotion recognition system more accurate, the fusion of the signals from multiple modalities to recognize emotions has attracted the interest of an increasing number of researchers.

In the traditional field of machine learning, Bao *et al.* [17] used EEG and eye movement signals to analyze the impact of gender differences on classification and used two neural network classifiers to assess the gender differences of five emotions. The results showed that the overall accuracy of same-sex strategies was relatively high. Chen *et al.* [8] proposed a multichannel EEG feature extraction method that combines differential entropy and linear discriminant analysis, and the method achieved an accuracy rate of 82.5%. Huang *et al.* [1] proposed a multimodal emotion recognition framework that combines facial expressions and EEG signals. The framework uses different support vector machine classifiers to detect valence and arousal learning targets and finally combines them using fusion technology. The results show that the effect after fusion is significantly better than that of a single form, and the best result reaches 70%. Although traditional machine learning methods can be applied well in emotion recognition and achieve good results, considering that the training data of some signals are very large, the feature vectors belonging to different signals are directly connected in series, resulting in the formation of high-dimensional feature vectors. In doing so, the differences between different modalities are ignored, which leads to problems such as feature

redundancy, insufficient expression of key features, or unnecessary computational cost. However, with the increase in processing speed and computational power of computers, the design and implementation of deep learning networks has become possible, which will have great impact on signals and signal processing [18].

Deep learning models are also gradually being used in the field of multimodal emotion classification and recognition based on EEG [19]. Xing *et al.* [20] proposed a multichannel EEG-based emotion recognition framework that is composed of a mixed linear EEG model and an emotion timing model. By decomposing the EEG source signals from the collected EEG signals and improving the classification accuracy by using the contextual correlations of the EEG feature sequences, the average accuracy rate on the DEAP data set is 81.10%. Granados *et al.* [21] used a CNN to automatically extract features from a variety of physiological signals and perform sentiment prediction through fully connected network layers. The experiments show that this method has achieved better precision in the classification of emotional states. To investigate the application potential of eye-tracking glasses in multimodal emotion recognition, Guo *et al.* [22] used eye images, eye movements and EEG signals to classify five emotions. The results showed that the three modes can improve the ability to recognize the five complementary emotions. Moreover, the results showed that the classifier using the fused features of human eye images and eye movements can achieve a classification accuracy of 71.99%.

Currently, many studies have used feature fusion and multiple CNN models to extract intermediate features and fuse them into models with different architectures, achieving some success. However, when only one particular layer of the CNN or the final output feature of the whole model is extracted, the characteristic features at multiple scales are ignored, and the phenomenon of missing important features occurs. The information of each mode cannot be fully expressed. To solve this problem, this paper proposes a HFCNN model based on the traditional CNN. The HFCNN uses different parameter settings in the convolutional layer to construct different layered network structures to extract the convolutional features of each layer and form a global feature vector by weight fusion. By combining these features, we aim to increase the multiscale representation of multimodal signals and uncover domain-specific knowledge and class-discriminating features that have extracted at different levels. Manually extracted statistical features are added to classify human emotions, which improve the accuracy of multimodal emotion recognition.

III. PROPOSED METHODOLOGY

The multimodal emotion recognition model based on the HFCNN is shown in Figure 1. Considering that a single signal cannot be used to accurately distinguish categories of emotions, multimodal observations of EEG signals and peripheral physiological signals (PPS) are used to track real human emotions. The HFCNN is used to extract the features of the multimodal signals, and the multiscale expression of the multimodal signals is enhanced by weighted fusion, which improves the accuracy of emotion recognition. The emotion recognition model proposed in this paper will be discussed in detail below.

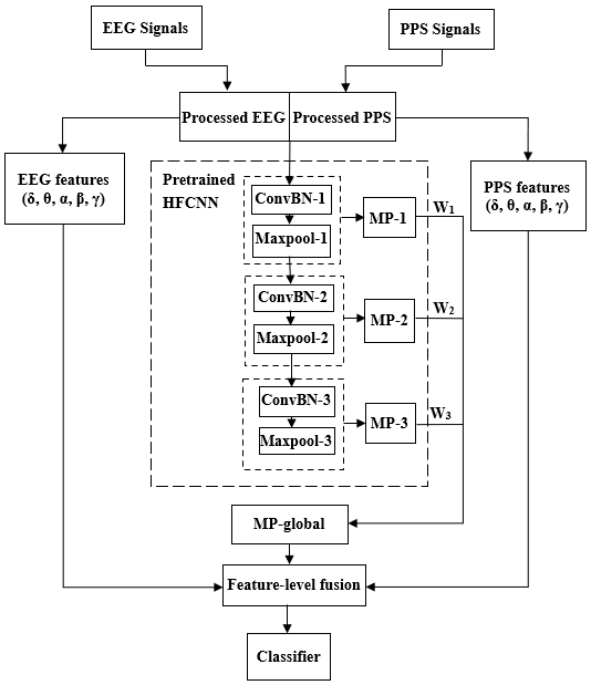

FIGURE 1. Multimodal emotion recognition model using the hierarchical fusion convolutional neural network.

A. DATA PREPROCESSING AND OBSERVATION-LEVEL FUSION OF MULTIMODAL SIGNALS

This paper uses video stimulation to induce human emotions. The subjects were asked to watch a group of selected video clips of different emotions and simultaneously, their physiological signals were collected through sensors. We select six forehead channels (FP1, FP2, AF3, AF4, F3, and F4) and four PPS signals, including galvanic skin response (GSR), respiration belt (RESP), skin temperature (TEMP), and plethysmograph (PLET), as input to the emotion recording model and use the middle 30 s of each video as experimental data. To eliminate noise interference, a band-pass filter and a low-pass filter are applied to EEG signals and PPS signals to eliminate noise and downsample the signals. In the experiment, according to the valence and arousal dimension scores of the videos, a median of 5 is used as the threshold to divide the levels into two categories, namely, high valence (HV)/low valence (LV) and high arousal (HA)/low arousal (LA). Due to the problem of imbalanced number of samples in the categories in the experiment, the sample points are expanded to increase the number of sample points in the smaller categories to balance the data set (the specific expansion method can be found in Section IV), which ensures that the number of samples of the two categories is approximately equal to

TABLE 1. The number of expanded sample points.

DEAP dataset				MAHNOB-HCI dataset			
Label	Data Quantity	Label	Data Quantity	Label	Data Quantity	Label	Data Quantity
HA	33930	HV	32580	HA	11475	HV	12015
LA	33670	LV	32020	LA	11025	LV	11985
Total	67600	Total	64600	Total	22500	Total	24000

improve the normalization ability of the classification model. The final sample size is shown in Table 1.

After preprocessing the sample data in the data set, the observation-level feature fusion method is used to merge the observation-level vectors of EEG signals and PPS signals into a single vector as input to the HFCNN and feature function computation. Then the hierarchical convolutional feature extraction is performed.

B. THE HIERARCHICAL FUSION CONVOLUTIONAL NEURAL NETWORK

The traditional CNN consists of stacks of multiple convolutional layers and pooling layers. It can adaptively adjust the convolution kernels in each layer to obtain some desired features [23]. The CNN computes the local features of different scales which are the output of the previous layer by adjusting the sizes of different convolution kernels in the convolutional layer, and these features are combined to obtain the output of the next layer by the activation function. The pooling layer, which can reduce the parameter settings of the next layer due to the retention of the main features to prevent overfitting, further optimizes the convolutional layer [24]. The training of CNN model is mainly divided into two stages. The first stage is the forward propagation. The data enters the convolutional layer and the pooling layer, and then the final output is obtained by the activation function. The second stage is the backward propagation and updating the weights. According to the error between the predicted value and the true value obtained by forward propagation, the error function of each network layer is obtained by backward propagation and the weights are updated [25].

This paper uses a layered incremental architecture within the CNN and constructs different network structures by setting different convolution kernel sizes and numbers in the convolutional layer. Then, hierarchical local convolutional features are extracted from the multimodal input signals to solve the problem of missing key features in traditional neural networks, and the multiscale representation of the multimodal signals is added. This greatly improves the performance of the model. The specific process is shown in Figure 2.

First, the representations of the multimodal signals at the observation level are combined into a unified representation vector to enter the input layer of the HFCNN. Then, according to the characteristics of emotion data, three incremental convolutional processes are used as the structure of the HFCNN model to extract hierarchical convolutional features. There are 8 filters of size 7*7 in the first convolutional layer, 12 filters of size 5*5 in the second and 16 filters of size 3*3 in the third convolutional layer. In the proposed architecture, a maximum pooling is used in the pooling layers. The maximum pooling is obtained in 2*2 with a stride of 2. The observation-level representations of the multimodal signals enter three incremental convolutional layers and pass through the maximum pooling layer to successively output features MPs-1, MPs-2, and MPs-3, as shown in Figure 2. The output features of the next layer are obtained by using the Relu activation function to combine the results of the convolutional features of the previous layer. We extract these features after maximum pooling layers to obtain compressed features without losing meaningful and important information. We assign different importance to the convolutional features, each of which has undergone three convolutional layer operations. Finally, the MPs-1, MPs-2 and MPs-3 features extracted by the three incremental processes are combined with weights to form MPs-global features. Before performing feature extraction, the CNN model is pre-trained. Then the model parameters and weights are frozen after training, and the fusion model is trained. In this paper, maximum pooling is used to optimize the convolutional features, reduce the parameter settings of the next layer and prevent overfitting. We use the strategy provided by recent experimental advances in CNNs in deep learning and machine learning, and the convolution output uses standardization techniques. To prevent further overfitting, a dropout layer is added before the convolutional features of the incremental process are fused. In addition, a supervised learning strategy is used where each input sample of the model is mapped to an output class. The softmax layer is the target class and produces predicted values. According to the error between the predicted value and the true value, a stochastic gradient descent with a small batch size is performed, and the back-propagation algorithm optimizes the network parameters.

C. STATISTICAL FEATURES AND HIERARCHICAL LOCAL CONVOLUTIONAL FEATURE EXTRACTION

The EEG signals are divided into the time domain, frequency domain, and time-frequency domain. We manually extract the Hjorth parameters in the time domain signal. For the frequency domain signal, the mean value of power spectral density and the mean value of difference entropy are calculated after Fourier transform. For the preprocessed PPS signals, the mean, variance, maximum, minimum, and difference between the maximum and minimum of the signal, and the mean, variance, maximum, minimum, and difference between the maximum and minimum of the first-order

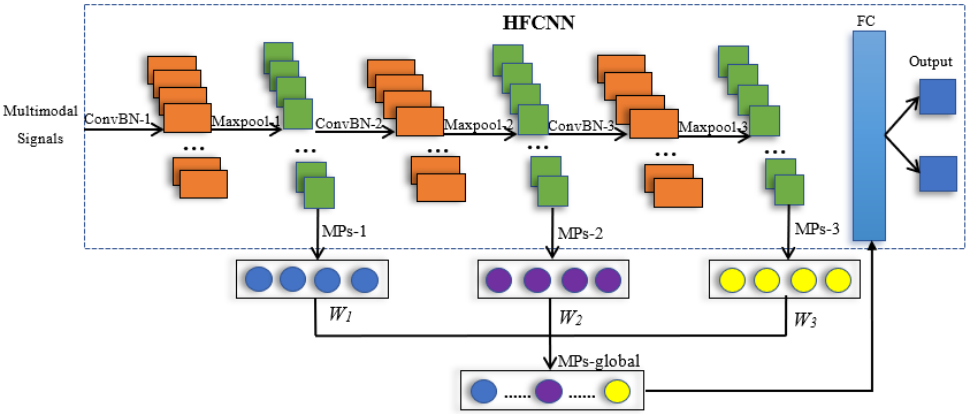

FIGURE 2. The specific process of the HFCNN proposed in this paper.

and second-order difference are calculated as characteristic parameters.

Since EEG data is dynamic and has a low signal-to-noise ratio, manually extracted features cannot achieve highly accurate results. Therefore, we add hierarchical local convolutional features to improve the effect of the emotion recognition model, apply the trained HFCNN model to the representation-level features of the multimodal signals of each sample, and extract the local features after maximum pooling in the model sequentially for MPs-1, MPs-2, and MPs-3. The magnitude of the contribution value of the features to the output is taken as the weight. The features are intercepted after convolution of each layer and finally the MPs-global is formed. The goal of adopting hierarchical architecture is to reveal the domain-specific knowledge extracted by the CNN at different levels and the class-distinguishing features, while avoiding the problem of missing key features during the extraction process [14]. The weighted feature fusion method can compress features without losing meaningful and important information.

D. WEIGHT-BASED FEATURE FUSION

In this paper, the weights are calculated based on feature-level fusion, and the observation-level fusion of the two modalities is used as the input of the model. After the different sized convolution kernels, the maximum pooling layer is used to extract the local convolutional features, and the features are weighted according to the best accuracy of the output. The specific formula is as follows [26]:

$$H = W_1 * A(F_1) \oplus W_2 * A(F_2) \oplus W_3 * A(F_3) \qquad (1)$$

where H represents the convolutional features formed after the CNN, A represents a convolution operation in the CNN, W_i $(i = 1,2,3)$ represents the weight, F_i $(i = 1,2,3)$ represents the feature signal of different levels, and $\oplus$ represents the connection operation applied to the output of different scale features.

In this weighted feature fusion, different importance is given to the convolutional features at different levels. Finally, the manual features and the convolutional features are combined into a feature vector using feature level fusion. We define the features from two different operations as $H_1^1 = \{H_1^1, H_2^1, \ldots, H_m^1\}$ and $H_1^2 = \{H_1^2, H_2^2, \ldots, H_m^2\}$, respectively. The combined features can be obtained via feature-level fusion:

$$H^{1\oplus 2} = \left\{H_1^1, \ldots, H_m^1, H_1^2, \ldots, H_m^2\right\} \qquad (2)$$

IV. EXPERIMENTS AND RESULT ANALYSIS

This paper focuses on the feature extraction and fusion method for multimodal physiological signals. To evaluate the effectiveness of our proposed method, we conducted experiments on the DEAP and MAHNOB-HCI data sets. The hierarchical local convolutional features extracted by the HFCNN are used to form a global feature vector with weights, and finally feature fusion with manual features is performed. RF is used as a classifier to train and test the two dimensions of arousal and valence using ten-fold cross-validation method.

A. DATA SET SETTINGS

This paper evaluates the proposed model on two data sets containing multiple physiological signals and emotional assessments.

In the DEAP data acquisition experiment, 32 subjects watched 40 videos with different emotions, where each video was approximately one minute in length. Various physiological signals were detected and recorded, and then the different dimensions were scored from 1-9 [27]. Taking into account that the brain regions related to the frontal lobe have high recognition accuracy [28], the 6-channel EEG signals of the forehead and the PPS signals of other remaining channels are used as experimental data in the experiment. The data is downsampled to 128 Hz, and five bands including the delta (4-8 Hz), theta (8-13 Hz), alpha (13-30 Hz), beta (30-43 Hz), and gamma bands (4-43 Hz) are filtered out. Due to the error in the first 3 s of the video in the experiment, the first 3 s of the video are removed, and the middle 30 s of the remaining duration of the video are used as experimental

data. The dimension of the data of each subject is 128 (time points) * 6 (channels) * 30 (video times). To overcome the problem of insufficient number of sample points in deep learning, a 6-s window with 50% overlap is applied to the multimodal source signals to increase the number of experimental sample points, and then applied to each band. Taking the DEAP data set as an example, after increasing the overlap window with a length of 3 s, the total number of signal segments for each subject in the 40 experiments is $40*9 = 360$, and then after applying the window to each band, the number of sample points for each subject becomes $360* 5 = 1800$. The final dimension of the data for each subject is 4608 (128 (time points) * 6 (channels) * 6 (video times)) * 1800 (epochs). Each video is divided into two categories on the dimensions of valence and arousal, with a median score of 5 as the threshold. If the score of a video on valence or arousal dimension is greater than or equal to 5, it belongs to HV/HA; otherwise, it belongs to LV/LA. Therefore, the dimension of the label corresponding to each subject is 1*1800.

The MAHNOB-HCI data set [29] was generated by 30 subjects (17 females and 13 males) watching 20 videos with different emotions, including EEG signals, peripheral physiological signals and eye movement signals. These 20 videos are from famous commercial movies which can derive emotions from subjects. Each video clip was about 34 to 117 seconds [30]. During the experiment, participants' behavior was recorded by cameras, microphones and gaze trackers. Every respondent scores his/her emotion on arousal, valence, control and predictability after watching the videos. Due to technical problems and data collection failures in the experiment, there were 3 subjects with missing data records and 2 subjects with incomplete data records; therefore, we used the data of the remaining 25 subjects in the experiment. During processing, the data was downsampled to 256 Hz. EEG signals of the same channel as DEAP data set and the remaining peripheral physiological signals in the MAHNOB-HCI data set are selected. After removing the neutral video length, we selected the middle 30 s of the remaining video length as experimental data. The other processing methods were the same as those of the DEAP data set for comparison. The dimension of each subject is 9216 (256 (time points) * 6 (channels) * 6 (video times)) * 1800 (epochs). In manual feature extraction, the two processed data sets are extracted every second; therefore, a total of 30 features in the time-domain and frequency-domain of EEG signals and 90 PPS features are obtained.

B. MULTIMODAL EMOTION RECOGNITION RESULT

To eliminate the influence of the classifier parameters on the experimental results, we input the fusion features of the two data sets into the RF classification model to investigate the influence of the changes of the parameters in the RF on the experimental accuracy. Figure 3 shows the change curve of accuracy with increasing number of decision trees in RF on the dimensions of arousal and valence of DEAP and MAHNOB-HCI data sets, respectively. From the figure, it can be seen that the accuracy of the multimodal features after processing gradually increases as the number of decision trees in RF increases, and the accuracy eventually reaches a certain value that is close to stable. Therefore, we fixed the parameters of the two data sets with different dimensions to certain values. The decision trees for arousal and valence in DEAP are 450 and 200, respectively, and the decision trees in MAHNOB-HCI are 600 and 500, respectively. The corresponding average accuracy rates are shown in Table 2.

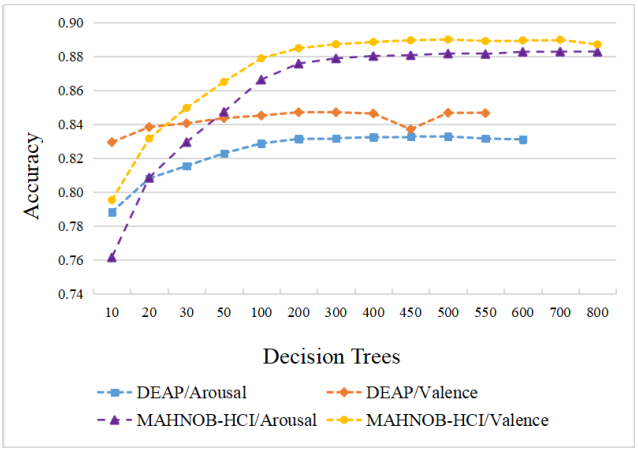

FIGURE 3. The change curve of accuracy with the parameter setting in the classifier.

It can be seen from Table 2 that in the DEAP data set, the average accuracy of arousal is 83.28% and valence is 84.71%. In the MAHNOB-HCI data set, the accuracy of arousal is 88.28% and valence is 89.00%. It can be concluded that using the same data set in the parameter optimization case, the performance of valence is better than that of arousal, indicating that valence is more active in the model proposed in this paper. When the classifier parameters are optimal, the accuracy of the model proposed in this paper when applied to the MAHNOB-HCI data set is higher than the accuracy when the model is applied to the DEAP data set. It can be seen from Table 2 that the experimental results obtained by using the HFCNN for feature extraction are significantly better than the results obtained by using the CNN alone for feature extraction. The reason is that although the CNN can automatically learn and extract multilayer feature representations from the original data, the neural network does not focus on one feature for each neuron, but a group of neurons focus on one feature, and there may be the problem of missing important features in feature extraction. Therefore, we extract convolutional features by setting up an incremental network structure with different parameters to achieve the multiscale fusion of multimodal signals and add manual features to improve the accuracy of the HFCNN model.

In order to better compare the superiority of the method proposed in this paper, we compare the proposed method with state-of-the-art methods using the same data set, and guarantee the same number of emotional states. The detailed

TABLE 2. The experimental results under the optimal parameter settings.

Dimension	Arousal		Valence	
Dataset	DEAP	HCI	DEAP	HCI
CNN	75.98%(RF450)	70.17%(RF600)	79.52%(RF200)	71.38%(RF500)
HFCNN	83.28%(RF450)	88.28%(RF600)	84.71%(RF200)	89.00%(RF500)

TABLE 3. Compare the results with other methods.

Author	Method	Dataset	Average Accuracy	Number of emotion status
Xing et al. [20]	SAE+LSTM	DEAP	81.10%	2
Zhu et al. [28]	Multi-hypergraph neural networks	DEAP	82.95%	2
Chao et al. [31]	Caps Net	DEAP	68.28%	2
Chen et al. [32]	Hierarchical Bidirectional GRU	DEAP	67.90%	2
Ours	HFCNN	DEAP	84.71%	2
		MAHNOB-HCI	89.00%	2

information of the comparison method as shown in Table 3. Xing *et al.* [20] built and solved a linear mixing model of multichannel EEG signals based on stack autoencoder (SAE), and established an emotion timing model based on the long short-term memory RNN (LSTM-RNN) to decompose the EEG source signals. Zhu *et al.* [28] used the multi-hypergraph neural network to identify the emotional status from physiological signals. The correlation between different subjects was represented by multi-hypergraphs, and each physiological signal constituted a hypergraph to explore the potential correlation between multiple physiological signals and the relationship between different subjects. Chao *et al.* [31] extracted multiband features from multichannel EEG signals to constitute the multiband feature matrix (MFM), and entered the capsule network to mine three related features. Chen *et al.* [32] introduced the attention mechanism combined with the hierarchical bidirectional Gated Recurrent Unit (GRU) network to mirror the hierarchical structure of the EEG signals. By paying different levels of attention to content with different importance, the model can learn more significant feature representation of EEG sequence which highlights the contribution of important samples and epochs to its emotional categories.

Through the analysis of the comparative research methods, most relevant studies made the same choice. To ensure that the comparative results are fair, our experiment focuses on two dimensions of arousal and valence. It can be seen from Table 3 that the accuracy of our proposed HFCNN model is significantly higher than other models in multimodal emotion recognition. The HFCNN adopts a hierarchical incremental approach, which can mine various potential information in multimodal emotion recognition, increase the multiscale feature representation of multimodal signals, avoid the interference caused by the manual adjustment of parameters, and improve the multimodal emotion recognition performance.

In order to improve the generalization ability of the model, the model proposed in this paper performs emotion recognition using the data of all subjects' EEG signals and PPS signals without considering the individual differences between subjects. The results show that the model has achieved good results in multimodal emotion recognition tasks for all subjects. Furthermore, each layer of features extracted by the proposed HFCNN was tested separately, and the results proved that only the weighted feature fusion method that fuses the features of the three incremental network structures can achieve good experimental results. Although this paper uses a hierarchical architecture, there is no additional cost in the runtime of the experiment compared to other deep learning network models. In any case, the HFCNN model has achieved satisfactory results in terms of multimodality.

V. CONCLUSION

This paper proposes a novel hierarchical multimodal learning model based on early fusion in the context of emotion recognition. The proposed hierarchical fusion multimodal deep learning model is based on a CNN network with an end-to-end method. We evaluated the performance of the proposed model on DEAP and MAHNOB-HCI data sets containing multiple physiological signals and emotional ratings. First, we extracted the preprocessed and denoised middle 30-s 6-channel EEG data and retained the peripheral channel data from 32 subjects watching 40 videos in DEAP data set and 25 subjects watching 20 videos in MAHNOB-HCI data set. Then, the 30-s multimodal source signals were divided into 6-s fixed-size windows with a 50% overlap to increase the number of experimental sample points and applied to each band. Then, the dataset was balanced to ensure that the number of the two classes was equal to improve the normalization ability of the classification model. We obtained the global fusion features MPs-global of MPs-1, MPs-2, and

MPs-3 in the manner described above. Finally, a 10-fold cross-validation data set was created for each subject, and the HFCNN model was used to perform binary emotion classification experiments in the valence and arousal dimensions.

The experimental results showed that the fused signals with the features extracted by the HFCNN gave better and more stable performance, which overcame the lack of key features in the traditional CNN and improved the multiscale representation of the multimodal signals. Through analysis, we found that combining HFCNN features with statistical features can achieve considerable accuracy in multiple modalities. We presented a subject-independent emotion recognition system that is suitable for real-time operation and has the ability to recognize the actual emotional state, as it not only avoids the large engineering of manual feature extraction and feature selection before traditional machine learning classification, but also effectively improves the accuracy and stability of multimodal emotion recognition. It also provided a valuable method for developing a powerful brain-computer interface for multimodal emotion recognition and regulation.

Although good experimental results were obtained in the present study, further research is needed on how to extract more discriminative multimodal features to perform cross-subject emotion classification, how to select, construct, and optimize deep learning models with higher accuracy, robustness, and generalization for multimodal emotion recognition, and how to incorporate some emotion-related brain neurogenic analysis into the analysis of experimental results. In addition, significant efforts should be made to reduce the sensory-induced interference for our multimodal signals. All these are the main contents of our next research work.

REFERENCES

[1] Y. Huang, J. Yang, S. Liu, and J. Pan, "Combining facial expressions and electroencephalography to enhance emotion recognition," *Future Internet*, vol. 11, no. 5, pp. 105–121, May 2019.

[2] E. Ghaleb, M. Popa, and S. Asteriadis, "Multimodal and temporal perception of audio-visual cues for emotion recognition," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2019, pp. 552–558, doi: 10.1109/ACII.2019.8925444.

[3] D. Wu, J. Zhang, and Q. Zhao, "Multimodal fused emotion recognition about expression-EEG interaction and collaboration using deep learning," *IEEE Access*, vol. 8, pp. 133180–133189, 2020.

[4] W. Li, M. Chu, and J. Qiao, "Design of a hierarchy modular neural network and its application in multimodal emotion recognition," *Soft Comput.*, vol. 23, no. 22, pp. 11817–11828, Jan. 2019.

[5] J. A. Domínguez-Jiménez, K. C. Campo-Landines, J. C. Martínez-Santos, E. J. Delahoz, and S. H. Contreras-Ortiz, "A machine learning model for emotion recognition from physiological signals," *Biomed. Signal Process. Control*, vol. 55, Jan. 2020, Art. no. 101646.

[6] Y. Lu, M. Wang, W. Wu, Y. Han, Q. Zhang, and S. Chen, "Dynamic entropy-based pattern learning to identify emotions from EEG signals across individuals," *Measurement*, vol. 150, Jan. 2020, Art. no. 107003.

[7] Z. Wang, X. Zhou, W. Wang, and C. Liang, "Emotion recognition using multimodal deep learning in multiple psychophysiological signals and video," *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 4, pp. 923–934, Jan. 2020, doi: 10.1007/s13042-019-01056-8.

[8] D.-W. Chen, R. Miao, W.-Q. Yang, Y. Liang, H.-H. Chen, L. Huang, C.-J. Deng, and N. Han, "A feature extraction method based on differential entropy and linear discriminant analysis for emotion recognition," *Sensors*, vol. 19, no. 7, pp. 1631–1647, Apr. 2019.

[9] M. Riyad, M. Khalil, and A. Adib, "Cross-subject EEG signal classification with deep neural networks applied to motor imagery," in *Proc. Int. Conf. Mobile, Secure, Program. Netw. (MSPN)*, 2019 pp. 124–139.

[10] Z. Jia, Y. Lin, X. Cai, H. Chen, H. Gou, and J. Wang, "SST-EmotionNet: Spatial-spectral-temporal based attention 3D dense network for EEG emotion recognition," in *Proc. 28th ACM Int. Conf. Multimedia (MM)*, Oct. 2020, pp. 2909–2917.

[11] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Bencherif, and M. S. Hossain, "Multilevel weighted feature fusion using convolutional neural networks for EEG motor imagery classification," *IEEE Access*, vol. 7, pp. 18940–18950, 2019.

[12] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal residual LSTM network," in *Proc. 27th ACM Int. Conf. Multimedia*, Nice, France, Oct. 2019, pp. 176–183.

[13] H. Zhang, "Expression-EEG based collaborative multimodal emotion recognition using deep autoencoder," *IEEE Access*, vol. 8, pp. 164130–164143, 2020.

[14] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Mekhtiche, and M. S. Hossain, "Deep learning for EEG motor imagery classification based on multi-layer CNNs feature fusion," *Future Gener. Comput. Syst.*, vol. 101, pp. 542–554, Dec. 2019.

[15] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168865–168878, 2020.

[16] J. Li, S. Qiu, C. Du, Y. Wang, and H. He, "Domain adaptation for EEG emotion recognition based on latent representation similarity," *IEEE Trans. Cogn. Develop. Syst.*, vol. 12, no. 2, pp. 344–353, Jun. 2020.

[17] L.-Q. Bao, J.-L. Qiu, H. Tang, W.-L. Zheng, and B.-L. Lu, "Investigating sex differences in classification of five emotions from EEG and eye movement signals," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Berlin, Germany, Jul. 2019, pp. 6746–6749.

[18] C. Wei, L.-L. Chen, Z.-Z. Song, X.-G. Lou, and D.-D. Li, "EEG-based emotion recognition using simple recurrent units network and ensemble learning," *Biomed. Signal Process. Control*, vol. 58, Apr. 2020, Art. no. 101756.

[19] P. Zhong, D. Wang, and C. Miao, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Trans. Affect. Comput.*, early access, May 11, 2020, doi: 10.1109/TAFFC.2020.2994159.

[20] X. Xing, Z. Li, T. Xu, L. Shu, B. Hu, and X. Xu, "SAE+LSTM: A new framework for emotion recognition from multi-channel EEG," *Frontiers Neurorobot.*, vol. 13, p. 37, Jun. 2019.

[21] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-González, E. Abdulhay, and N. Arunkumar, "Using deep convolutional neural network for emotion detection on a physiological signals dataset (AMIGOS)," *IEEE Access*, vol. 7, pp. 2169–3536, 2019.

[22] J.-J. Guo, R. Zhou, L.-M. Zhao, and B.-L. Lu, "Multimodal emotion recognition from eye image, eye movement and EEG using deep neural networks," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Berlin, Germany, Jul. 2019, pp. 3071–3074.

[23] S. Xie and H. Hu, "Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 211–220, Jan. 2019.

[24] M. Wu, W. Su, L. Chen, Z. Liu, W. Cao, and K. Hirota, "Weight-adapted convolution neural network for facial expression recognition in human-robot interaction," *IEEE Trans. Syst., Man, Cybern., Syst.*, early access, Mar. 8, 2019, doi: 10.1109/TSMC.2019.2897330.

[25] H. Huang, Z. Hu, W. Wang, and M. Wu, "Multimodal emotion recognition based on ensemble convolutional neural network," *IEEE Access*, vol. 8, pp. 3265–3271, 2020.

[26] Z. Li, L. Huang, and J. He, "A multiscale deep middle-level feature fusion network for hyperspectral classification," *Remote Sens.*, vol. 11, no. 6, p. 695, Mar. 2019.

[27] 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. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012.

[28] J. Zhu, X. Zhao, H. Hu, and Y. Gao, "Emotion recognition from physiological signals using multi-hypergraph neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 610–615, doi: 10.1109/ICME.2019.00111.

[29] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012.

[30] D. Y. Choi, D.-H. Kim, and B. C. Song, "Multimodal attention network for continuous-time emotion recognition using video and EEG signals," *IEEE Access*, vol. 8, pp. 203814–203826, 2020.
[31] H. Chao, L. Dong, Y. Liu, and B. Lu, "Emotion recognition from multiband EEG signals using CapsNet," *Sensors*, vol. 19, no. 9, pp. 2212–2227, May 2019.
[32] J. X. Chen, D. M. Jiang, and Y. N. Zhang, "A hierarchical bidirectional GRU model with attention for EEG-based emotion classification," *IEEE Access*, vol. 7, pp. 118530–118540, 2019.

YONG ZHANG received the M.S. degree in computer science from the University of Shanghai for Science and Technology, in 2002, and the Ph.D. degree in computer science from the Dalian University of Technology, in 2008. He is currently a Professor with the School of Computer and Information Technology, Liaoning Normal University, China. His current research interests include machine learning, intelligence computing, and affective computing.

CHENG CHENG received the bachelor's degree from Anshan Normal University, China, in 2018. She is currently pursuing the M.S. degree from Liaoning Normal University, China. Her research interests include machine learning and data mining.

YIDIE ZHANG received the bachelor's degree from Liaoning Normal University, China, in 2019, where she is currently pursuing the M.S. degree. Her research interests include machine learning and data mining.