IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 February 2023, accepted 28 February 2023, date of publication 31 March 2023, date of current version 5 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3263670

APPLIED RESEARCH

Multimodal Emotion Recognition From EEG Signals and Facial Expressions

SHUAI WANG[1], JINGZI QU[1], YONG ZHANG[2], AND YIDIE ZHANG[1]
[1]School of Computer and Information Technology, Liaoning Normal University, Dalian 116081, China
[2]School of Information Engineering, Huzhou University, Huzhou 313000, China

Corresponding author: Yong Zhang (zhyong@zjhu.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, in part by the Scientific Research Foundation of the Education Department of Liaoning Province under Grant LJKZ0965, and in part by the Huzhou Science and Technology Plan Project under Grant 2022GZ08.

ABSTRACT Emotion recognition has attracted attention in recent years. It is widely used in healthcare, teaching, human-computer interaction, and other fields. Human emotional features are often used to recognize different emotions. Currently, there is more and more research on multimodal emotion recognition based on the fusion of multiple features. This paper proposes a deep learning model for multimodal emotion recognition based on the fusion of electroencephalogram (EEG) signals and facial expressions to achieve an excellent classification effect. First, a pre-trained convolution neural network (CNN) is used to extract the facial features from the facial expressions. Next, the attention mechanism is introduced to extract more critical facial frame features. Then, we apply CNNs to extract spatial features from original EEG signals, which use a local convolution kernel and a global convolution kernel to learn the features of left and right hemispheres channels and all EEG channels. After feature-level fusion, the fusion features of the facial expression features and EEG features are fed into the classifier for emotion recognition. This paper conducted experiments on the DEAP and MAHNOB-HCI datasets to evaluate the performance of the proposed model. The accuracy of valence dimension classification is 96.63%, and arousal dimension classification is 97.15% on the DEAP dataset, while 96.69% and 96.26% on the MAHNOB-HCI dataset. The experimental results show that the proposed model can effectively recognize emotions.

INDEX TERMS Multimodal emotion recognition, EEG, facial expressions, deep learning, attention mechanism.

I. INTRODUCTION

Emotion plays a vital role in human communication. It affects our daily life. Human beings have different responses to different emotional states, and emotion recognition uses the collected emotional response characteristics and signals to evaluate human emotional states. Emotion representations can usually be divided into two models, the discrete model and the dimensional model. The discrete model has six primary emotional states, including anger, disgust, fear, happiness, sadness, and surprise, which are related to different facial expressions [1]. In the dimensional model, the emotions are mapped into the valence, arousal, and dominance dimensions [2]. Currently, the most commonly used dimensional model is the valence-arousal model proposed by Russell [3], which only includes valence and arousal. The arousal is from a happy to an unpleasant state, and the valence is from a calm to an excited state. This paper also uses the valence-arousal model to represent emotional states.

Traditional emotion recognition often uses machine learning methods, such as support vector machine (SVM), decision tree, and k-nearest neighbor (KNN) to classify [4], [5]. The deep learning model can automatically extract features compared to traditional emotion recognition methods [6]. Gao et al. [7] proposed an emotion classification method based on a multilayer convolutional neural network (CNN) and combining differential entropy (DE)

The associate editor coordinating the review of this manuscript and approving it for publication was Seifedine Kadry.

and brain networks, which achieved an average accuracy of 91.45%. Li et al. [8] proposed a multi-domain adaptive graph convolutional network (MD-AGCN) method, which combines the frequency and time domains to utilize the complementary information of electroencephalogram (EEG) signals. The MD-AGCN also considers the topological structure of EEG channels by combining the correlation between channels and the information within channels.

Human beings have different reactions to different emotional states. These reactions can be divided into physiological reactions and non-physiological reactions. Non-physiological reactions include posture, expression, etc., while physiological reactions include skin temperature, respiratory rate, heart rate, etc. Emotion recognition research often uses various sensors and other devices to collect the data signals of these reactions, such as EEG, electrocardiogram (ECG), blood volume pulse, electromyogram (EMG), etc. In the past, many researchers chose to extract EEG features for single-modal emotion recognition [9]. Still, more and more researchers use multimodal models that integrate multiple data and features to carry out emotion recognition, which can extract more abundant features and improve the effect of emotion recognition [10]. For example, Nakisa et al. [11] proposed a time-series multimodal fusion method with a deep learning model to capture the nonlinear emotional correlation between and within EEG and blood volume pulse (BVP) signals and improve emotion classification performance. This method uses a CNN and long short-term memory (LSTM) model to fuse EEG and BVP signals. After learning each modal feature using a single deep network, they can jointly learn and explore the highly correlated emotional expression between different modes. Zhao et al. [12] studied a network based on attention mechanisms and classified three human emotions using EEG, eye movement features, and original eye movement image data. The key idea of this model is to use a new co-attention layer, which increases the weight of crucial feature channels and establishes the correlation between different modes. In this model, the co-attention layers are stacked to form a hierarchical structure, and the effect is enhanced layer by layer to predict the emotional state more accurately. The accuracy rate was 87.63% in the emotion recognition task of the three modes.

However, most of the existing emotion recognition methods based on facial expressions need to train a new model to extract facial features, which consumes many computing resources. The pre-training technology has been widely used in computer vision and natural language processing tasks, which can reduce many computing resources and time resources required to develop CNN models. Therefore, this paper employs EEG signals and facial expressions for multimodal emotion recognition. First, we extract facial features from the facial expressions by employing the pre-trained CNN and attention mechanism, and extract spatial features from EEG signals using CNNs. Then, we perform feature-level fusion to obtain the fused multimodal features for emotion recognition. The main contributions of this paper are:

- Utilize the pre-trained CNN as an encoder to extract facial features from facial expression images.
- Employ the attention mechanism to assign weight to facial expression features of different frames to obtain key facial frame features.
- Use CNNs with local and global convolution kernels to obtain the spatial features of left and right brain channels and all EEG channels.

The rest of this paper is organized as follows. Section II introduces the facial features, EEG feature extraction, and multimodal emotion recognition methods. The multimodal emotion recognition model based on the fusion of EEG and facial expressions is proposed in Section III. Section IV describes the experiments and the results compared with other methods. Finally, section V concludes this paper.

II. RELATED WORK

Facial expressions play a significant role in daily human communication. People can judge emotions according to various facial expressions. The emotion recognition methods of facial expressions have also made great progress over the years. There are many manually extracted features in traditional facial emotion recognition, such as histogram of gradients (HOG), local binary pattern (LBP), Gabor wavelet transform, etc. Many studies use these traditional features for facial emotion recognition. For example, Kumar et al. [13] proposed a facial expression recognition framework, which can infer the emotional state in real time so that computers can interact with people more intelligently. The directional gradient histogram feature is extracted from the active face block instead of the whole face, which makes the system robust to scale and change. The feature vector is further input into the SVM classifier. The experimental results show that the accuracy rate is 95% in the Cohn Kanade (CK+) dataset for 5-fold cross-validation. Happy et al. [14] proposed a facial expression classification algorithm, which uses a Haar classifier for face detection, takes LBP histograms of different block sizes of face images as feature vectors, and employs principal component analysis (PCA) to classify various facial expressions.

In recent years, deep learning has played an important role in many fields. Many facial emotion recognition studies also use deep learning models to improve the accuracy and efficiency of classification. For example, Zhang et al. [15] proposed a facial expression recognition method based on CNN and image edge detection. Firstly, the facial expression image is normalized, and the edges of each image layer are extracted in the convolution process. Next, the extracted edge information is superimposed on each feature image to retain the edge structure information of the texture image. Then, the maximum pooling is used to reduce the dimension

of the extracted features. Finally, a softmax classifier is employed to classify and recognize the expression of the test sample image. In order to verify the robustness of this method for facial expression recognition, the FER-2013 facial expression database and LFW dataset were mixed for experiments. Experimental results show that the average recognition rate of this method achieves 88.56%. Moghaddam et al. [16] proposed a deep network model for facial emotion recognition using abundant spatial angle information in the light field image. First, the VGG16 CNN is used to extract spatial features. Then, the bi-directional LSTM (BI-LSTM) recurrent neural network is utilized to learn the spatial angle features from the viewpoint feature sequence and explore the forward and reverse angle relationships. The model can selectively focus on the most critical spatial angle features using the attention mechanism to achieve more effective learning results. Finally, the fusion scheme is used to obtain the classification results of emotion recognition.

For emotion recognition of EEG, the traditional methods extract features such as DE, wavelet transform, and power spectral density (PSD). For example, Zhu et al. [17] adopted DE as the feature, used a linear dynamic system (LDS) for feature smoothing, and finally employed SVM for classification. Bhatti et al. [18] extracted the features in three fields: time, frequency, and wavelet, from the recorded EEG signals. Further, they used them to identify human emotions by a classifier. Recently, the research of emotion recognition based on EEG has also begun to use deep learning models to extract features [7], [19]. For example, Li et al. [8] used various deep learning models to extract features from EEG signals for emotion recognition. Zhang et al. [20] proposed an emotion recognition model which uses normalized mutual information to measure the relationships between EEG features and corresponding emotions. Then, the proposed model employed heterogeneous CNNs to extract the convolutional features and fuse them by multimodal factorized bilinear pooling.

Recently, more and more multimodal emotion recognition models combined EEG data with facial video features. Fusing multimodal features can achieve a better emotion recognition effect. For example, Choi et al. [21] proposed a multimodal fusion network integrating video and EEG modalities. To calculate the attention weights of facial video features and corresponding EEG features, they described a bilinear pooled multimodal attention network based on low-rank decomposition. Finally, the efficiency value is calculated using the output of the two modal networks and the attention weight. Huang et al. [22] proposed two multimodal fusion methods between the brain and peripheral signals for emotion recognition, which utilizes a decision-level fusion of EEG and facial expression detections. Tan et al. [23] presented a multimodal emotion recognition method based on facial expressions and EEG to establish a human-robot interaction system with a low sense of disharmony.

TABLE 1. The structure of DeepVANet.

Layer	Details
Convolution	32@3×3 ReLu activator
Max pooling	3×3
Convolution	64@3×3 ReLu activator
Max pooling	3×3
Convolution	128@3×3 ReLu activator
Max pooling	3×3

This paper integrates the EEG and facial expression features for multimodal emotion recognition to enrich the features. We extract spatial features of left and right brain channels and all EEG channels from the original EEG signals using CNNs with local and global convolution kernels. Meanwhile, we employ the pre-trained CNN and attention mechanism to extract facial expression features from the face image sequences. The attention mechanism is introduced into the CNN feature extraction model to obtain more critical facial frame features and improve emotion recognition's effect. Finally, the features of the two types are input to the classifier for classification after feature-level fusion.

III. PROPOSED METHODOLOGY

The multimodal emotion recognition model proposed in this paper is shown in Figure 1. First, EEG signal features and facial expression features are extracted from the original data. Next, the extracted features are fused at the feature level. Finally, the multimodal fusion features are classified by the classifier. The model is described in detail below.

A. FACIAL FEATURE EXTRACTION

For the facial features, this paper uses CNN for transfer learning and combines the attention mechanism to extract the facial spatial features of important frames. We use the DeepVANet [24] pre-trained on the AFEW-VA database as an encoder to extract facial features. Table 1 shows its structure. The AFEW-VA database [25] contains 600 videos selected from movies showing various facial expressions. The database labels 30000 frames of expressions, of which 19858 frames were chosen in the pre-training. Each frame is marked with values ranging from -10 to 10 according to the valence and arousal dimensions.

The input to the pre-trained CNN is a sequence of face images. The shape of each sample $F_i = \{f_1, f_2, \ldots, f_T$ is $T \times 3 \times 64 \times 64$, where T is the sequence length, 3 is the number of RGB channels, and 64×64 is the size of each frame of images. Since five frames per second are extracted in data processing, $T = 5$. Using the pre-trained CNN, we can extract the spatial features of facial expressions $f'_1, f'_2, \ldots, f'_T$. Then, the attention mechanism is used to allocate attention weights to different frames of face images. The attention

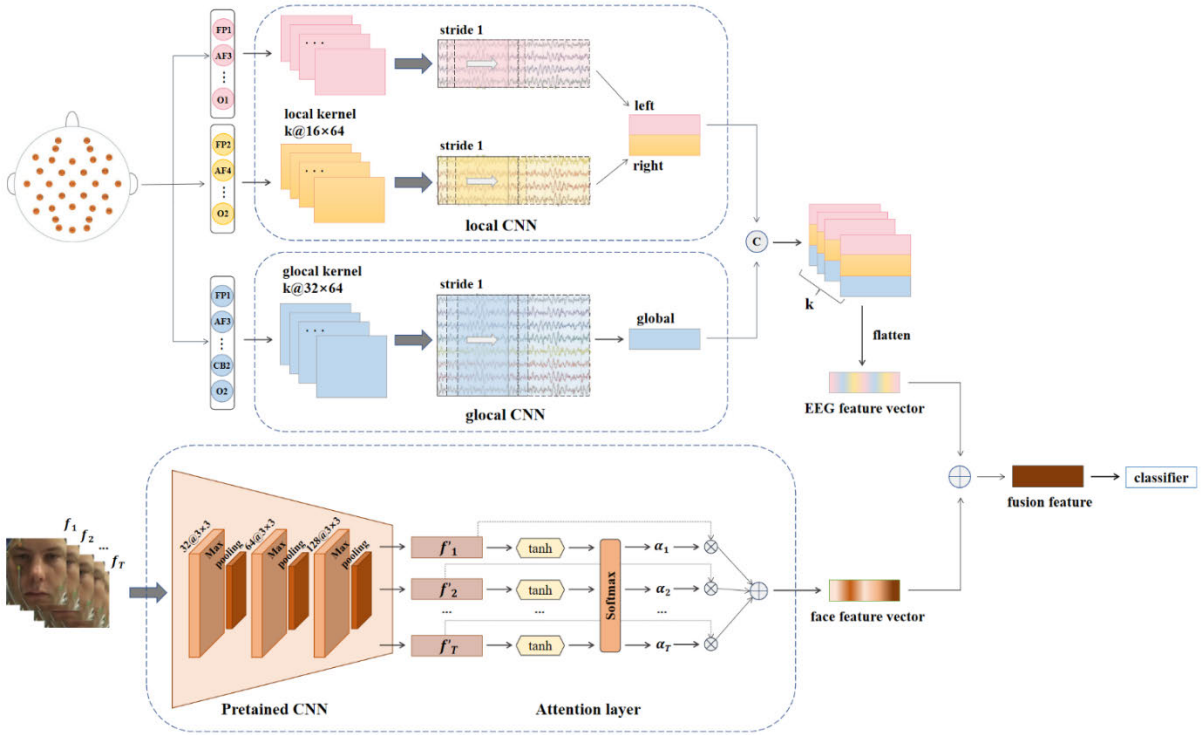

FIGURE 1. The proposed multimodal emotion recognition model.

weight is calculated as follows:

$$\alpha_j = Softmax\left(\tanh\left(W_j f_j'\right), u_j\right) = \frac{exp(\tanh\left(W_j f_j'\right)^T u_j)}{\sum_{j=1}^{T} exp(\tanh\left(W_j f_j'\right)^T u_j)} \quad (1)$$

where W_j is the weight matrix and u_j is a random initialization parameter. The expression feature $f_j'(j = 1,2,\ldots, T)$ of each frame extracted by the pre-trained CNN is subjected to nonlinear transformation, and then input into the softmax function in equation (1) to calculate the attention weight α_j. It is worth noting that all the calculated weights α_j add up to 1, and can be adjusted adaptively in training. The more important the feature, the higher its weight. The attention mechanism helps extract more important frame features and improve model classification accuracy.

After the attention weight is obtained, the feature vector extracted by the CNN of each facial expression image and the attention weight are multiplied and summed, respectively, to get the features of the final facial video modality V_i:

$$V_i = \sum_{j=1}^{T} \alpha_j f_j' \quad (2)$$

B. EEG FEATURE EXTRACTION

The proposed model also uses CNNs to extract spatial features from original EEG signals, which employs two convolution kernels: the local convolution kernel K_{local} and the global convolution kernel K_{global}. The global convolution kernel covers all EEG channels for convolution, and its size is (c, h), where c is the number of EEG channels. The size of the local convolution kernel is $(c/2, h)$, and the convolution kernel can convolute the EEG signals of the channels in the left and right brain regions. In DEAP and MAHNOB-HCI datasets, the number of channels and EEG electrode distribution is the same. Therefore, in this subsection, c is 32, h is 64, and the number of convolution kernel K is 20. The distribution position of EEG electrodes is shown in Figure 2, and the arrangement sequence of EEG electrode channels is shown in Table 2.

Let the ith EEG signal $P_i \in R^{c\times l}$, where l is the length of each EEG sample. The local convolution kernel can convolute the left brain electrode channels (1-16 channels) and the right brain electrode channels (17-32 channels) to extract their respective spatial features. The local features are shown as follows:

$$X_{local} = Maxpooling(Conv(P_i, K_{local})) \quad (3)$$

where $Conv(\bullet)$ represents convolution operation and $Maxpooling(\bullet)$ represents maximum pooling, respectively.

Global channel feature X_{global} can be obtained by a convolution kernel K_{global}:

$$X_{global} = Maxpooling(Conv(P_i, K_{global})) \quad (4)$$

The features obtained by the convolution of two convolution kernels are connected in series to obtain the EEG spatial features $X_i = \left[X_{local}, X_{global}\right]$.

TABLE 2. Arrangement of EEG electrode channels.

Channel no.	Channel name	Channel no.	Channel name
1	Fp1	17	Fp2
2	AF3	18	AF4
3	F3	19	Fz
4	F7	20	F4
5	FC5	21	F8
6	FC1	22	FC6
7	C3	23	FC2
8	T7	24	Cz
9	CP5	25	C4
10	CP1	26	T8
11	P3	27	CP6
12	P7	28	CP2
13	PO3	29	P4
14	O1	30	P8
15	Oz	31	PO4
16	Pz	32	O2

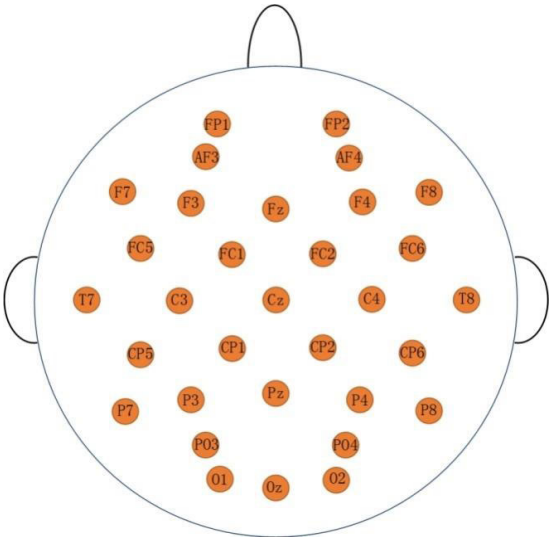

FIGURE 2. Location map of EEG electrodes.

Then, the features of the two modalities are fused at the feature level to obtain the multimodal feature vector $M_i = V_i \oplus X_i$. The obtained multimodal feature vectors are fed into the fully connected neural network classifier for classification. The classifier generates the predicted value S_i. Accordingly, the predicted label can be obtained as follows.

$$\hat{y} = \begin{cases} high, & S_i > 0.5 \\ low, & S_i \leq 0.5 \end{cases} \tag{5}$$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

This paper uses two standard datasets: DEAP [26] and MAHNOB-HCI [27] datasets to verify the model performance. Both datasets use the data of all subjects, and the model performance is discussed by the 5-fold cross-validation method.

TABLE 3. The emotional dataset description.

Attribute	DEAP	MAHNOB-HCI
Subjects	32	25
Trail of each subject	40	20
Available channels	40	38
Length of each trail	30 s	30 s
Items for rating emotion	Valence, Arousal	Valence, Arousal

A. EXPERIMENTAL DATASETS AND DATA PREPROCESSING

DEAP is a multimodal dataset used to study human emotional states. In this database, 32 subjects were examined, and 40 videos with a duration of 63s were selected as trigger stimuli. In addition, the central nervous system, peripheral physiological systems, and the facial expressions of the first 22 subjects were also recorded. After watching each video, the participants were asked to conduct self-assessments of valence, arousal, dominance, and liking. EEG signals on the DEAP dataset were down-sampled to 128 Hz and filtered from 4.0 Hz to 45 Hz; also, eye artifacts were eliminated by the blind source separation technique. The assessments are described in Table 3.

MAHNOB-HCI is a multimodal emotion database that records signals from 30 subjects while they watch 20 videos, including central nerve signals, peripheral physiological signals, and eye movement signals. After watching each video, participants mark their emotional scores on arousal, valence, control, and predictability. Since the video duration in the experiment is not uniform, we select the middle 30 seconds as the experimental data. In the investigation, the experimental recording files of three individuals were corrupted due to problems with the experimental equipment and recordings, and the experimental data of two individuals were incomplete. Therefore, we extracted the records from the remaining 25 participants for the experiment. The assessments are also described in Table 3.

This paper mainly concerns a model with four emotional states in an arousal and valence space, namely high arousal (HA) / low arousal (LA) and high valence (HV) / low valence (LV). Values less than five are considered low, while values between 5 and 9 are high.

Data preprocessing refers to the methods proposed by Zhang et al. [24] and Yang et al. [28] to process the data. The data is divided into segments according to the length of 1 second, and the number of samples of facial expression data and EEG data is consistent. According to this processing method, the number of samples per subject in the DEAP dataset is 2400, and the number of samples per subject in the MAHNOB-HCI dataset is 1611.

For facial expression data, we first intercept video pictures at five frames per second, then use the tool [29] to detect the face of each picture, align with 68 facial landmarks, and finally crop the face image. For EEG data, the sampling rate

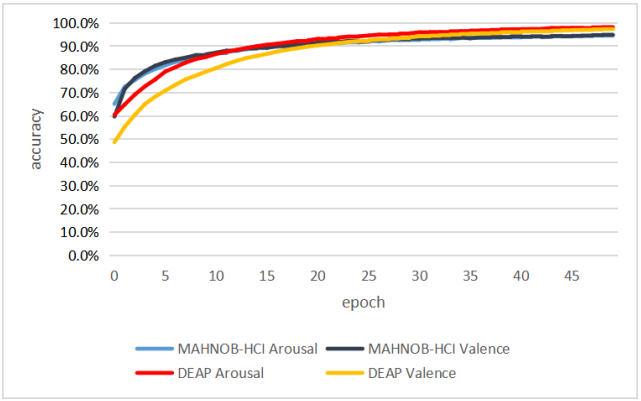

FIGURE 3. Accuracy-epoch line chart.

is reduced to 128Hz, and the band-pass filter in the toolbox of EEGLAB [30] is used to filter and remove artifacts. Then, the baseline data of the first 3 seconds is divided into three 1-second segments, and the average value of the baseline data of the first 3 seconds is used to represent the baseline signal data. Finally, the baseline is removed by subtracting the average value of the baseline data in the first three seconds from each EEG signal sample, and the formula is as follows:

$$final\_EEG_i = exper\_EEG_i - \frac{\sum_{n=1}^{3} base\_EEG_i^n}{3} \quad (6)$$

where $exper\_EEG_i$ represents the ith EEG data segment, $base\_EEG_i^n$ represents the baseline signal data, and $final\_EEG_i$ represents the EEG data with the baseline removed.

B. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed model in this paper is implemented by the *PyTorch* framework. We set the training batch size as 128 and the learning rate as 0.00001. We use the Adam algorithm as the optimizer during training and the binary cross entropy as the loss function.

The changing trend of the accuracy rate on DEAP and MAHNOB-HCI datasets with the training process is shown in Figure 3. It can be seen that the training accuracy rate gradually rises and finally tends to be stable, indicating that the model's training process is stable and can converge within a certain number of iterations.

The classification accuracy of the model proposed in this paper is shown in Table 4. In the binary classification task of the DEAP dataset, the accuracy of the valence and arousal dimensions is 96.63% and 97.15%, respectively. In the binary classification task of the MAHNOB-HCI dataset, the accuracy of the valence and arousal dimensions is 96.69% and 96.26%, respectively.

In order to verify the classification effect of the proposed model, it is compared with the multimodal emotion recognition methods using EEG and facial expressions, and the results are shown in Table 4. For example, Huang et al. [31] proposed a multimodal emotion recognition framework combining EEG and facial expressions. For facial expressions,

TABLE 4. The classification results of the proposed model and existing approaches.

Dataset	Authors	accuracy
DEAP	Huang *et al.* [31]	Valence:80.30% Arousal:74.23%
	Zhu *et al.* [32]	Valence:70.01% Arousal:77.08%
	Li *et al.* [33]	Valence:71.00% Arousal:58.75%
	Ours	Valence:96.63% Arousal:97.15%
MAHNOB-HCI	Koelstra *et al.* [34]	Valence:74.00% Arousal:70.00%
	Huang *et al.* [31]	Valence:75.21% Arousal: 75.63%
	Li *et al.* [33]	Valence:70.04% Arousal:72.14%
	Ours	Valence:96.69% Arousal:96.26%

the framework applies a pre-trained multi-task CNN model to extract facial features automatically, and uses a single modal framework to detect valence and arousal values. When extracting the PSD features of EEG, wavelet transform is used to capture the time-domain features of EEG, and then two different SVM models are used to identify the valence and arousal values. Finally, the decision-level fusion parameters are obtained based on the fusion method's training data. Zhu et al. [32] studied the multimodal decision-level fusion of EEG signals, peripheral physiological signals, and facial expressions. Their experiment uses the open multimodal DEAP database. It uses a one-dimensional convolution kernel neural network for EEG signals to extract the features of 32 EEG channels. For peripheral physiological signals, the neural network is used to detect and extract the relevant features. For facial expression video, three-dimensional CNN is used to extract features. Then the single-mode results are fused at the decision level to obtain multimodal accuracy. Li et al. [33] developed an open-source software toolkit named MindLink-Eumpy to identify emotions by integrating EEG and facial expression information. The toolbox first uses a series of tools to obtain the physiological data of the subjects automatically, then analyzes the facial expression data and EEG data obtained, and finally fuses two different signals at the decision level. MindLink-Eumpy uses a multi-task CNN in facial expression detection based on transfer learning. In EEG detection, MindLink-Eumpy provides two algorithms, an SVM model and an LSTM network model. In decision-level fusion, weight assignment and the AdaBoost algorithm are used to fuse multimodal features. Koelstra and Patras [34] used facial expressions and EEG signals analysis methods to study the possibility of multimodal fusion in emotion recognition and implicit tags. The PSD of five bands and the lateralization features of 14 pairs of left and right sides

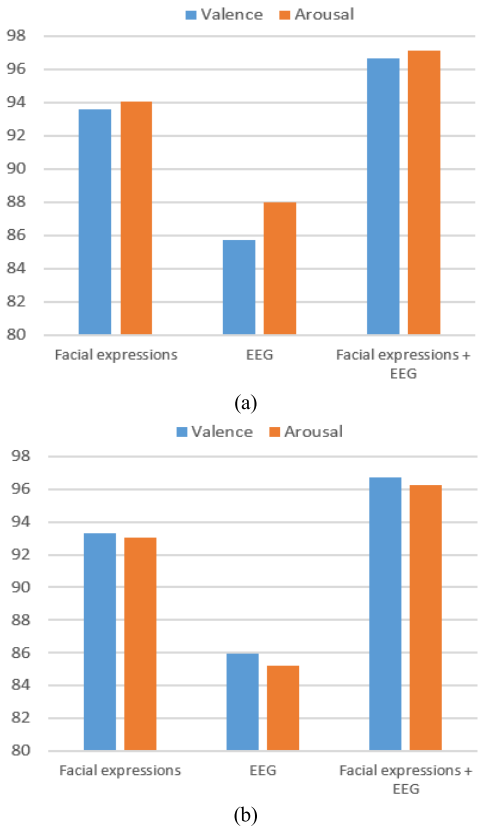

FIGURE 4. Ablation study results (recognition accuracy %). (a) DEAP dataset; (b) MAHNOB-HCI dataset.

of the brain were extracted for EEG signals. The facial action unit detection method is used for facial expression analysis. Finally, feature-level and decision-level fusions are used to classify.

In the above methods, Koelstra and Patras [34] used traditional feature extraction methods on facial expressions and EEG signals, Zhu et al. [32] used CNNs to extract features on facial expressions and EEG signals, while Huang et al. [31] and Li et al. [33] used pre-training models to extract features in facial expression analysis. Our proposed method uses the deep learning model in EEG feature extraction, which can directly extract effective, relevant features compared with traditional methods. In facial expressions, the pre-training CNN model is used to extract facial expression features automatically, and attention mechanism weights are also used to extract more important expression frame features. Through the above optimization methods, the model proposed in this paper outperforms other similar studies mentioned above in the classification tasks of DEAP and MAHNOB-HCI datasets.

To discuss the effect of multimodal emotion recognition and whether it has advantages over single-modal emotion recognition, this section uses EEG signals and facial expressions alone for emotion recognition on two datasets. The experimental results are listed in Table 5. To explore the emotion recognition performance of different modalities

TABLE 5. The classification results of single-modal and multimodal emotion recognition on two datasets (%).

Dataset	Dimension	Facial expressions	EEG	Facial expressions + EEG
DEAP	Valence	93.59	85.72	96.63
	Arousal	94.03	87.97	97.15
MAHNOB-HCI	Valence	93.34	85.98	96.69
	Arousal	93.06	85.23	96.26

more intuitively, we give the ablation experimental results as shown in Figure 4. It can be seen that the accuracy of multimodal emotion recognition based on the fusion of EEG signals and facial expressions is higher than that of single-modal emotion recognition in both arousal and valence dimensions.

V. CONCLUSION

This paper proposes a multimodal emotion recognition model based on the fusion of EEG signals and facial expressions. The end-to-end model can directly extract the features of EEG signals and facial expressions. For the facial expressions, the pre-trained CNN is used to extract facial features, and the attention mechanism is introduced to extract the features of crucial expression frames. Meanwhile, CNNs are employed to extract spatial features from EEG signals, which use a local convolution kernel and a global convolution kernel to learn the features of left and right brain channels and all EEG channels. After feature level fusion, two types of features are fed into a classifier for emotion recognition, and the predicted valance and arousal labels are output. The experimental results show that the proposed model can effectively carry out emotion recognition, and the multimodal emotion recognition effect of EEG and facial expression fusion is better than that of using EEG or facial expressions alone. Next, we will explore a more reliable pre-training model to extract facial expression features and reduce the resources and time required for model operation. We will also try to introduce more modalities, such as non-physiological signals, to enrich the multimodal emotion recognition model.

REFERENCES

[1] P. Ekman, "Universals and cultural differences in the judgments of facial expressions of emotion," *J. Personality Social Psychol.*, vol. 53, no. 4, pp. 712–717, 1987.
[2] P. Ekman, "An argument for basic emotions," *Cognit. Emotion*, vol. 6, nos. 3–4, pp. 169–200, 1992.
[3] J. A. Russell, "Core affect and the psychological construction of emotion," *Psychol. Rev.*, vol. 110, no. 1, pp. 145–150, 2003.
[4] J. Atkinson and D. Campos, "Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers," *Expert Syst. Appl.*, vol. 47, pp. 35–41, Apr. 2016.
[5] P. Ackermann, C. Kohlschein, J. A. Bitsch, K. Wehrle, and S. Jeschke, "EEG-based automatic emotion recognition: Feature extraction, selection and classification methods," in *Proc. IEEE 18th Int. Conf. e-Health Netw., Appl. Services (Healthcom)*, Sep. 2016, pp. 1–6.
[6] Y. Zhang, C. Cheng, and Y. Zhang, "Multimodal emotion recognition based on manifold learning and convolution neural network," *Multimedia Tools Appl.*, vol. 81, no. 23, pp. 33253–33268, Apr. 2022.

[7] Z. Gao, R. Li, C. Ma, L. Rui, and X. Sun, "Core-brain-network-based multilayer convolutional neural network for emotion recognition," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.
[8] R. Li, Y. Wang, and B.-L. Lu, "A multi-domain adaptive graph convolutional network for EEG-based emotion recognition," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 5565–5573.
[9] L. Feng, C. Cheng, M. Zhao, H. Deng, and Y. Zhang, "EEG-based emotion recognition using spatial–temporal graph convolutional LSTM with attention mechanism," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 11, pp. 5406–5417, Nov. 2022.
[10] C. Cheng, Y. Zhang, L. Liu, W. Liu, and L. Feng, "Multi-domain encoding of spatiotemporal dynamics in EEG for emotion recognition," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 3, pp. 1342–1353, Mar. 2023.
[11] B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire, and V. Chandran, "Automatic emotion recognition using temporal multimodal deep learning," *IEEE Access*, vol. 8, pp. 225463–225474, 2020.
[12] Z.-W. Zhao, W. Liu, and B.-L. Lu, "Multimodal emotion recognition using a modified dense co-attention symmetric network," in *Proc. 10th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, May 2021, pp. 73–76.
[13] P. Kumar, S. L. Happy, and A. Routray, "A real-time robust facial expression recognition system using HOG features," in *Proc. Int. Conf. Comput., Analytics Secur. Trends (CAST)*, Dec. 2016, pp. 289–293.
[14] S. L. Happy, A. George, and A. Routray, "A real time facial expression classification system using local binary patterns," in *Proc. 4th Int. Conf. Intell. Human Comput. Interact. (IHCI)*, Dec. 2012, pp. 1–5, doi: 10.1109/IHCI.2012.6481802.
[15] H. Zhang, A. Jolfaei, and M. Alazab, "A face emotion recognition method using convolutional neural network and image edge computing," *IEEE Access*, vol. 7, pp. 159081–159089, 2019.
[16] A. Sepas-Moghaddam, A. Etemad, F. Pereira, and P. L. Correia, "Facial emotion recognition using light field images with deep attention-based bidirectional LSTM," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3367–3371.
[17] J. Y. Zhu, W. L. Zheng, and B. L. Lu, "Cross-subject and cross-gender emotion classification from EEG," in *World Congress on Medical Physics and Biomedical Engineering, Toronto, Canada*. Cham, Switzerland: Springer, Jun. 2015, pp. 1188–1191.
[18] A. M. Bhatti, M. Majid, S. M. Anwar, and B. Khan, "Human emotion recognition and analysis in response to audio music using brain signals," *Comput. Hum. Behav.*, vol. 65, pp. 267–275, Dec. 2016.
[19] Y. Zhang, Y. Zhang, and S. Wang, "An attention-based hybrid deep learning model for EEG emotion recognition," *Signal, Image Video Process.*, Dec. 2022, doi: 10.1007/s11760-022-02447-1.
[20] Y. Zhang, C. Cheng, S. Wang, and T. Xia, "Emotion recognition using heterogeneous convolutional neural networks combined with multimodal factorized bilinear pooling," *Biomed. Signal Process. Control*, vol. 77, Aug. 2022, Art. no. 103877.
[21] 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.
[22] Y. Huang, J. Yang, P. Liao, and J. Pan, "Fusion of facial expressions and EEG for multimodal emotion recognition," *Comput. Intell. Neurosci.*, vol. 2017, pp. 1–8, Jan. 2017.
[23] Y. Tan, Z. Sun, F. Duan, J. Solé-Casals, and C. F. Caiafa, "A multimodal emotion recognition method based on facial expressions and electroencephalography," *Biomed. Signal Process. Control*, vol. 70, Sep. 2021, Art. no. 103029.
[24] Y. Zhang, M. Z. Hossain, and S. Rahman, "DeepVANet: A deep end-to-end network for multi-modal emotion recognition," in *Proc. 18th Int. Conf. Hum.-Comput. Interact.*, Bari, Italy, Aug. 2021, pp. 227–237.
[25] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "AFEW-VA database for valence and arousal estimation in-the-wild," *Image Vis. Comput.*, vol. 65, pp. 23–36, Sep. 2017.
[26] S. Koelstra, C. Muhl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis using physiological signals," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012.
[27] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012.
[28] Y. L. Yang, Q. F. Wu, Y. Z. Fu, and X. W. Chen, "Continuous convolutional neural network with 3D input for EEG-based emotion recognition," in *Proc. 25th Int. Conf. Neural Inform. Process.*, Dec. 2018, pp. 433–443.
[29] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1021–1030.
[30] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.
[31] Y. R. Huang, J. H. Yang, S. Y. Liu, and J. H. Pan, "Combining facial expressions and electroencephalography to enhance emotion recognition," *Future Internet*, vol. 11, no. 5, pp. 1–17, 2019.
[32] Q. Zhu, G. Lu, and J. Yan, "Valence-arousal model based emotion recognition using EEG, peripheral physiological signals and facial expression," in *Proc. 4th Int. Conf. Mach. Learn. Soft Comput.*, Jan. 2020, pp. 81–85.
[33] R. Li, Y. Liang, X. Liu, B. Wang, W. Huang, Z. Cai, Y. Ye, L. Qiu, and J. Pan, "MindLink–Eumpy: An open-source Python toolbox for multimodal emotion recognition," *Frontiers Hum. Neurosci.*, vol. 15, Feb. 2021, Art. no. 621493.
[34] S. Koelstra and I. Patras, "Fusion of facial expressions and EEG for implicit affective tagging," *Image Vis. Comput.*, vol. 31, no. 2, pp. 164–174, Feb. 2013.

SHUAI WANG received the bachelor's degree from Zhongshan College, Dalian Medical University, China, in 2019. She is currently pursuing the M.S. degree with Liaoning Normal University, China. Her research interests include machine learning and data mining.

JINGZI QU received the bachelor's degree from Liaoning Normal University, China, in 2021, where he is currently pursuing the M.S. degree in computer science. His research interests include affective computing and data mining.

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 Information Engineering, Huzhou University, Huzhou, China. His research interests include machine learning, intelligence computing, and affective computing.

YIDIE ZHANG received the bachelor's and M.S. degrees in computer science from Liaoning Normal University, China, in 2018 and 2022, respectively. Her research interests include machine learning and data mining.