IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON EMERGING DEEP LEARNING THEORIES AND METHODS FOR BIOMEDICAL ENGINEERING

Received June 30, 2020, accepted July 9, 2020, date of publication July 20, 2020, date of current version July 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3010311

Multimodal Fused Emotion Recognition About Expression-EEG Interaction and Collaboration Using Deep Learning

DI WU[1,2], **JIANPEI ZHANG**[1], **AND QINGCHAO ZHAO**[1]
[1]College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
[2]College of Computer and Control Engineering, Qiqihar University, Qiqihar 161006, China

Corresponding author: Jianpei Zhang (zhangjianpei@hrbeu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61672179, Grant 61370083, and Grant 61402126, in part by the Research Fund for the Doctoral Program of Higher Education of China under Grant 20122304110012, and in part by the Youth Science Foundation of Heilongjiang Province of China under Grant QC2016083.

ABSTRACT The proposed emotion recognition model is based on the hierarchical long-short term memory neural network (LSTM) for video-electroencephalogram (Video-EEG) signal interaction. The inputs are facial-video and EEG signals from the subjects when they are watching the emotion-stimulated video. The outputs are the corresponding emotion recognition results. Facial-video features and corresponding EEG features are extracted based on a fully connected neural network (FC) at each time point. These features are fused through hierarchical LSTM to predict the key emotional signal frames at the next time point until the emotion recognition result is calculated at the last time point. Specially, a self-attention mechanism is applied to show the correlation of the stacked LSTM at different hierarchies. In this process, the ''selective focus'' is used to analyze the human-emotional temporal sequences in each model, which improves the utilization of the key spatial EEG signals. Moreover, the process includes the temporal attention mechanism to predict the key signal frame at next time point, which utilizes the key emotion data in temporal domain. The experimental results prove that the classification rate (CR) and F1-score of the proposed emotion recognition model are significantly increased by at least 2% and 0.015, respectively, compared to other methods.

INDEX TERMS Emotion recognition, long-short term memory neural network, attention mechanism, multimodal signal fusion, electroencephalogram, time sequence.

I. INTRODUCTION

Emotions are very complex mental states or processes of human beings. They reflect the attitudes and cognition of people. Emotion communication is an important part of social life [1]. Therefore, emotionalization for human-computer interaction is not only a hot topic in the life science and information research, but also a focus with many unsolved problems in the cognitive field. In 1990, Salovey and Mayer firstly proposed the concept of emotional intelligence. It involves the ability to perceive, express, and regulate the emotion [2]. Most existed human-computer systems have certain logical intelligence. However, their emotional intelligence is not enough to interact and communicate with users smartly. With the rapid development of human-computer interaction, more humanized computers are expected to understand people better, and to assist people with many duties. This means the computers need to recognize different emotions. During the human-computer interaction, if the computer can identify the emotions of users quickly and accurately, it will adjust its working methods and content intelligently. A more friendly and natural human-computer interaction is created to provide a better user experience [3], [4]. Consequently, studies on how to obtain sentiment state information efficiently and how to model, recognize, and adapt the human emotion state, have been an important part in human-computer interaction and artificial intelligence.

Emotions includes three parts: subjective experience (personal feelings), physiological arousal (physiological signal changes in body), and outer performance (quantified response to actions for each body part). During daily communication, people usually identify the emotional state of others

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

by external performances, such as facial expressions, gestures, and intonations, which is also widely used in traditional emotion recognition. But besides above external performances, emotion changes are related to physiological indexes in the central nerves and autonomic nerves system. All these indexes provide the quantified base to the emotion changes [4], [5]. Actually, both external performance and physiological arousal reflect emotion changes in different respects and angles. Therefore, it is difficult to fully represent the exact emotion state just through actions or body physiological signals. That means single-model emotion recognition is unable to meet the requirement of human-computer interaction [6]. The future human-computer emotional interaction research tends to: use the complementarity for the multimodal signals to emotional state, build a more robust emotional recognition model, and establish a multimodal emotional interaction algorithm combined outer performance with body signals [7], [8].

Emotion recognition in this paper is defined as a process to analyze the time series signals from each emotional modal in a ''selective focus'' manner. This process was inspired by attention mechanism of human visual system. When human observe a specific scene, instead of accepting and understanding the whole situation at one time, they dynamically ''focus'' on partial or local information slices in the visual space. Then they integrate the obtained information to understand the current scene. Similarly, when the proposed emotion recognition model receives the modal signals of emotion, it learns the signal at each time point and predicts the key signal frame that is needed ''focus'' at the next time point. Repeated analysis and prediction will continue until sufficient information is obtained to give finial emotion recognition results.

II. RELATED RESEARCH

External features, such as the facial expression, body gesture and voice, are widely used in traditional emotion recognition methods [9]. The acquisition of these signals does not need wearable sensors. It is simple, convenient and low-cost. However, these signals are easily affected by the subjective factors from each participant. In the case of inconsistent between the real emotions and external performances, the system is unable to make a correct judgement. Moreover, external performance is not all of the emotions, it is not enough to express rich emotions. Physiological changes are controlled by central nervous system, which reflects the emotional state more objectively. In consequence, using physiological signals for emotion recognition is now an international hotspot in the emotional computing field [10].

Physiological signals, for example electroencephalogram, electromyogram (EEG), galvanic skin response (GSR), electrooculogram (EOG), electrocardiogram, blood pressure (ECG), blood volume pulse (BVP), epidermal temperature, and eye movement signal, are used to study emotion recognition by many scholars worldwide. Picard *et al.* [11] used four signals, including electromyogram, blood pressure, galvanic skin response and heart rate, to identify eight emotions (grief, pleasure, anger, annoyance, neutrality, and reverence), resulting in 81% classification accuracy. As a dominant emotion, physiological activities in brain can not only represent different emotions, but also objectively reflect the real one. In 2011, Nie *et al.* [12] extracted energy values from five frequency bands after the short-time Fourier transform of the EEG signals and achieved 87.53% recognition accuracy.

At present, as is proved in many EEG-based emotion recognition studies, the use of EEG signals for emotion recognition is effective. These studies mainly focus on different aspects like feature extraction, feature selection, and classification models of EEG. In 2012, Duan *et al.* [13] firstly applied Short-time Fourier transform to original EEG signals, extracted energy, rational asymmetry (RASM) and differential asymmetry (DASM) features from five frequency bands (Theta, Delta, Alpha, Beta, Gamma). They succeed to identify two types of emotions (calm and excited) with these three kinds of EEG features, achieving more than 80.03% classification rate. Their result shows that the brain discharge has asymmetric characteristics when emotions are generated. Zheng *et al.* [14] put forward a novel deep belief network (DBN) based method to recognize emotions. They trained DBN to recognize three emotions (positive, neutral, and negative) and the obtained average recognition accuracy is 86.08%. Furthermore, they explored critical regions and frequency bands by examining the weight distribution learned by DBN. As a result, the Bets band and Gamma band at the temporal lobes of the head were found to be mostly relevant to emotion.

The basic theory proves that when emotions arise, a variety reaction of human physiological and external behavior are activated [15]. For example, some changes may occur in face, voice, gesture and other physiological states (such as heart rate increasing) when people get angry. All these contribute to application of multimodal signals in emotion recognition. Even the difficulties inherent in single-modal emotion recognition will bring more challenges to multi-modal occasions, the advantages of the latter still attract scholars to further studies.

For example, Wagner *et al.* [16] extracted voice, face and gesture features and performed multimodal fusion in both feature and decision level. At the same time, a problem was solved that researchers were able to use other modal data for emotion recognition when the data of one modal was missing. Experiment results shown that the accuracy rate of emotion recognition using just one signal is between 42% and 51%. But after three-modal fusion, the accuracy rate reaches to 55%. In 2016, Ranganathan *et al.* [17] applied DBN and convolutional deep belief network (CDBN) to fuse facial expressions, body gestures, voice and physiological signals together. And then four emotions were recognized through unsupervised learning method. As a result, they found that the deep belief network can extract multi-modal features with high robustness, and the accuracy of emotion recognition is also higher than that of the existing emotion recognition technologies. In addition, the convolutional deep belief network

designed by them is able to learn significant facial features. At that time, their model is better than the other methods in recognizing low-intensity emotions or subtle expression.

Facial expressions, speech, voice, body gestures and other external behaviors are often used as inputs in above studies to emotion recognition, which are also the main signals used in multimodal conditions. Moreover, some central nervous system signals (such as EEG signals) and other physiological signals are also common in multimodal emotion recognition. For instance, Verma and Tiwary [18] recognized thirteen emotions using EEG, GSR, blood pressure, breath model, skin temperature, EMG and EOG signals.

Many studies have proved that EEG signals can identify different emotions effectively. However, scholars cannot highly improve the recognition accuracy after combining physiological signals with EEG signals. It may because that effective information which is related to emotions are not precisely extracted from these signals. A lot of redundancy and noise existed in it. Therefore, this method does not work well in emotion recognition. In 2015, Oleymani *et al.* [19] proposed an emotion recognition algorithm based on EEG and eye movement signals. The algorithm was not depending on individuals. From the Arousal level, the three states of calmness, moderate excitement and arousal are identified, and finally the accuracy rate of 68.5% is achieved. From the valence level, three states of the pleasant, moderate pleasant and unpleasant are recognized, and finally the accuracy rate of 76.4% is achieved.

In recent two years, research has gradually focused on the heterogeneous multi-modal fusion about facial images and EEG signals. For example, Huang *et al.* [20] studied emotion recognition technology through two inputs involving face image and EEG. The input signal is collected by emotional stimulus from the movie clips corresponding to different emotions. Facial expression detection recognized 4 basic emotion states. EEG detection recognized four basic emotion states and three emotion intensity levels, respectively. Emotion recognition and classification is based on two decision-level fusion methods. Li *et al.* [21] proposed a decision-level fusion framework of both EEG and facial expression in continuous emotions recognition. Three types of movie clips (positive, negative, and neutral) were used to elicit specific emotions of subjects, simultaneously, the EEG and facial expression signals were recorded. The power spectrum density (PSD) features of EEG were extracted by time-frequency analysis, and then EEG features were selected for regression. For the facial expression, facial landmark localization was applied to calculate the facial geometric features. The decision-level fusion was completed, and temporal dynamics of emotions were captured by Long short-term memory networks (LSTM). Islam *et al.* [22] revealed the emotional essence through two cases, and then judged whether emotion fluctuated by comparing two emotions. In addition, the data in two cases could obtain emotion state effectively. One process was facial expression recognition, the other was EGG extraction. Combining the evaluation result and analysis data of these two processes, we can get the fluctuation of emotional state.

However, two critical problems are still needed solving in this field. One is how to integrate the heterogeneous multi-modal emotion signals in an interactive and collaborative manner, to obtain more accurate recognition results. The second is how to quickly locate key emotional information from multi-modal signals which contain a lot of redundancy, to improve the efficiency and accuracy of the model. Take a two-minute face video as an example, it recorded the expression of one subject when he was watching comedy images. In this video, the subject only laughed for about 10 seconds, and the remaining video frames were all redundant for emotion recognition.

In response to the above problems, a LSTM-based emotion recognition model of video-EEG signal interaction and collaboration is proposed in this paper. The main contributions are as follows:

1)For better expressing the correlation between different levels of stacked LSTM, a self-attention mechanism is used in this research. The traditional attention model introduces external information, which is more similar to an alignment mechanism, to measure the matching degree between the corresponding positions of the output and the input. While the self-attention mechanism can update the learning parameters only through its own information instead of using external information.

2)The proposed model analyzes the time series from each emotion modal in a ''selective focus'' manner. With the help of the time domain attention mechanism, it can use the emotion date efficiently and predict the time information in the key signal frame at the next time point.

III. SYSTEM MODEL AND PROBLEM MODELING

A. EMOTION RECOGNITION MODEL OF VIDEO-EEG SIGNAL INTERACTION AND COLLABORATION

The target face videos are facial expression signals collected by ordinary cameras, depth cameras, etc., which helps to analyze intuitive and external emotions. The EEG signal is an amplified biological signal at the scalp generated by brain activity. It reflects the activity of billions of neurons in the cerebral cortex, which can help to analyze the deep and internal sentiments. Theoretically, the emotion recognition model which used both face video and EEG signals can integrate the external and internal emotion information, thus giving more accurate recognition results.

FIGURE.1 shows the structure of the proposed video-EEG model, it involves two important processes: feature extraction and interaction coordination. The two processes are coupled and related with each other. Then time series in each emotional model is further analyzed in a ''selective focus'' manner to give final recognition results.

On the one hand, in the feature extraction stage, the original EEG signals are first visualized as the image sequence of α, β and θ waves, in order to save more time-spatial

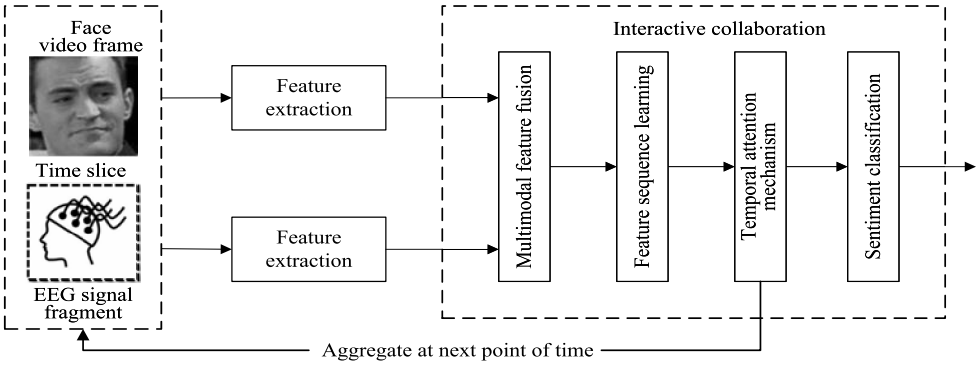

FIGURE 1. Emotion recognition model of video-EEG signal interaction and collaboration.

information from the signals, then the two models can collaborated efficiently. Corresponding EEG image features are finally extracted from the FC-based face-video frame. On the other hand, in the interactive collaboration stage, LSTM was trained to fuse the features from the two modalities and learned from them. Then it predicts the time information from the ''focused'' key signal frame at the next time point. The network will feed the prediction back to the feature extraction stage. That forms a loop, the above loop will repeat until the end of the sequence. The emotion recognition results for the entire signal is calculated at last. During the whole process, through the attention mechanism of spatial frequency band, the importance of images from three EEG waves (α β and θ) is computed, thus the key spatial information in EEG signals can be used to an effective degree. Through the time domain attention mechanism, the time information from the key signal frame at next time point is predicted, thus the significant information in emotion data will also be used at most efficiency.

In this model, a closed loop is formed between the inputs and the model reactions. It was also an emotion recognition process by selectively ''repeat focusing'' the multi-modal signals of human beings.

B. FC-BASED FEATURE EXTRACTION

The input signals in this paper are the face video and EEG signals collected from each subject when they are watching the emotion-stimulus video. Among them, the facial video, as a visual signal, is collected by camera from the expression activity of each subject. EEG, as a physiological signal, refers to the electric potential generated from spontaneous and rhythmic movements of cortical neurons according to the chronological order. Subjects wear EEG caps while watching emotion-stimulus video. Then 32 different channels of EEG signals from the cerebral cortex is collected. Since these two signals have different structure that are difficult to be fused directly, extracting features with strong expression and generalization ability is recommended to interact themselves effectively in two models. For the face video, compared with other feature extraction algorithms, the FC-based expression feature performs better at further mining distributed expression features of data. Meanwhile, the EEG signal in this paper is firstly converted into images of three frequency band. This visualization process preserves the time-spatial features of EEG signals while unifying the two modal signals into images. Then the EEG image features are extracted by attention mechanism based on FC and spatial frequency band.

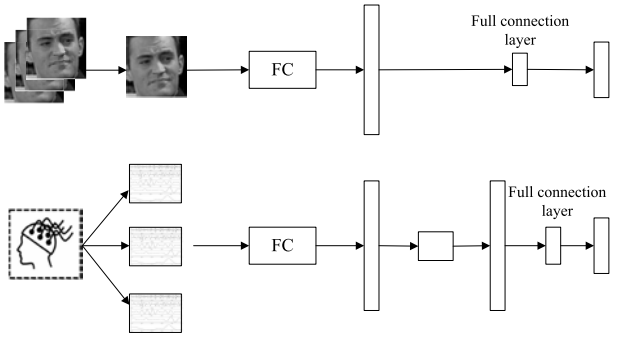

FIGURE 2. Feature extraction flowchart for facial video and EEG.

FIGURE.2 presents the flowchart for face video feature extraction. The face detection was firstly done by Faster-RCNN, then the facial features was extracted with FC algorithm, feature vectors are eventually obtained through fully connection layer $\boldsymbol{x}_{v,n}$.

Especially, the feature extraction was the most complex part. Firstly, artifacts in original EEG signal are removed by wavelet soft threshold algorithm, to achieve pure signals. Next, referring the data processing proposed by Bashivan *et al.* [24], the EEG signals are divided into different segments with a duration of T (1/T is the frame rate of facial video), then spectral energy are extracted from three types of waves (α, β and θ) at t^{th} data segment. The EEG images corresponding to 32 channels was achieved by visualizing above spectral energy. It can be seen that as the rising of emotion activation, β wave on the forehead will significantly increase. At last, $\boldsymbol{e}_{\alpha,n}$, $\boldsymbol{e}_{\beta,n}$ and $\boldsymbol{e}_{\theta,n}$ features from three frequency band are extracted and fused by FC algorithm, as is shown in equation (1) and(2).

Importance indexes $\boldsymbol{e'}_n$ from three sets of features are calculated by attention mechanism of spatial frequency band, and inputted into the fully connection layer to obtain feature vector $\boldsymbol{x}_{e,n}$.

$$\boldsymbol{e}_n = \boldsymbol{e}_{\alpha,n}\theta_{en,1} + \boldsymbol{e}_{\beta,n}\theta_{en,2} + \boldsymbol{e}_{\theta,n}\theta_{en,3} \tag{1}$$

$\theta_{en,1}, \theta_{en,2}, \theta_{en,3}$ represent the importance assigned to $\boldsymbol{e}_{\alpha,n}$, $\boldsymbol{e}_{\beta,n}$ and $\boldsymbol{e}_{\theta,n}$ respectively:

$$\theta_{en,i} = \frac{\exp\left(\boldsymbol{W}_{h,i}\boldsymbol{h}_{n-1} + b_{n,i}\right)}{\sum_{j=1}^{3}\exp\left(\boldsymbol{W}_{h,j}\boldsymbol{h}_{n-1} + b_{n,j}\right)} \quad i = 1, 2, 3 \tag{2}$$

$\boldsymbol{W}_{h,i}$ indicates the weight matrix waited for training, $b_{n,i}$ is the deviation. $\boldsymbol{h}_{n-1}$ shows the hidden state at last time point $(n-1)$ in LSTM network.

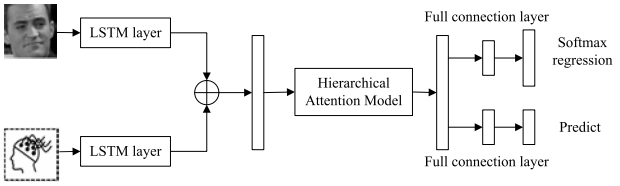

FIGURE 3. Interactive collaboration flowchart of hierarchical LSTM-based attention mechanism.

IV. INTERACTIVE COLLABORATION OF HIERARCHICAL LSTM-BASED ATTENTION MECHANISM

A. INTERACTIVE COLLABORATION

FIGURE.3 explains the interactive collaboration process in this paper, a two-layer LSTM (the first layer includes two LSTMs with shared parameters) is applied to fuse and learn the feature sequences from each model. LSTM is good at handling time series, avoiding the long-distance dependence of traditional recurrent neural networks. Time-domain attention mechanism is also introduced, in a reinforcement learning manner, to learn and predict the signal frame that needs to be ''focused'' at the next time point. Finally, emotion recognition is completed by the Softmax classifier.

Based on hard attention, time selective attention mechanism is proposed in this paper, which involves four parts: glimpse, core, action and reward. For the specific face video and EEG signals in a certain segment of T, the largest length of action sequence was pre-set to $N_{\max}$, so at the time point n:

1) Glimpse part: the focused position f_n and two feature vectors $\boldsymbol{x}_{v,n}, \boldsymbol{x}_{e,n}$ at this target frequency are accepted, the first LSTM layer handles the features from two models and the state from last time point as two sets of hidden states. Then the two sets are spliced into one feature vector, finishing multimodal signal fusion.

2) Core part: this part was made up by the second LSTM layer. The inputs for this layer involve two elements: partial emotion features produced from glimpse period, and the hidden state $\boldsymbol{h}_{n-1}$ at the last time point. The outputs are the latest hidden state $\boldsymbol{h}_n$ at current time point. The whole process integrates emotional information in history.

3) Action part: this part is used to predict the time position f_{n+1} of the key signal frame at the next time point, and finally to output the emotion recognition result at the last time point by Softmax classifier. It gives as: $p = (p_1, \ldots, p_C)^T$, $p_k = p\left(C_k \left| \boldsymbol{h}_N \right.\right), k = 1, \ldots, C$ indicates the probability that the emotion belongs to class C_k. Especially, the ending condition of the prediction is: the time position of the key signal frame at the next time point is the last frame of the given emotional signal, it is given by $f_{n+1} = N_{\max}$, or when the length of the behavior sequence reaches the maximum value, namely for $N = N_{\max}$.

4) Reward part: after each time of sampling, a reward message is fed back. With the time attention mechanism, the emotion recognition model is working in a reinforcement learning manner. But The model cannot observe the whole environment at one time, instead, the partial or local information from the two modal signals is obtained from each sampling. In this case, the model can autonomously learn strategies $\boldsymbol{\pi}\left(\boldsymbol{\alpha}_n \left| \boldsymbol{s}_{1:n}; \theta \right.\right)$. $\boldsymbol{\alpha}_n$ represents the behavior of the emotion recognition model under the strategy at the time point n, that means, the time information f_{n+1} of the ''focused'' signal at the next moment needed calculating. $\boldsymbol{s}_{1:n}$ refers to the past state (including the state at current time point), which is the input and output time series of the time attention mechanism. Therefore, the strategy $\boldsymbol{\pi}$ with the parameter θ(namely for $\boldsymbol{\pi}_\theta$) is the analysis results combined the current inputs with historical glimpse, which is significant to calculate the key signal frame to be ''focused'' at next time point. As a result, our goal is to find a strategy to maximize the cumulative sum of reward time. But there is a delay in reward accumulation, as is given by $R_N = \sum_{n=1}^{N} r_n$, R_N refers to the total reward obtained from one-time emotion recognition during total N time points, r_n is the reward that is obtained by each ''focus'' analysis behavior in one recognition. It is consistent with the reward obtained at the end of the entire behavior sequence.

B. HIERARCHICAL ATTENTION MECHANISM

As the LSTM is stacked on the time axis, new information is added during the iteration of each time step. The network will update by passing on this information. The essence for this process is to provide infinite memory depth for LSTM, but the network update process does not deal with any layered information. The hierarchical architecture allows greater model complexity. At lower level, it focuses on the basic concepts, and at the higher level, it concerned more about abstract features. The stacked LSTM layer is used in this paper to express the hierarchical time series. A vector sequence, obtained from each LSTM layer, is the input of the next LSTM layer. This creates a more effective feature representation for the input data and enhances the modal expression.

In order to express the correlation between different layers of stacked LSTM, a self-attention mechanism is proposed in this paper. Traditional attention model involves many external

information. It is similar to an alignment mechanism to measure the matching degree between the current output position and the specified input position. But the self-attention mechanism does not require any external information, but only update the learning parameters through its own information.

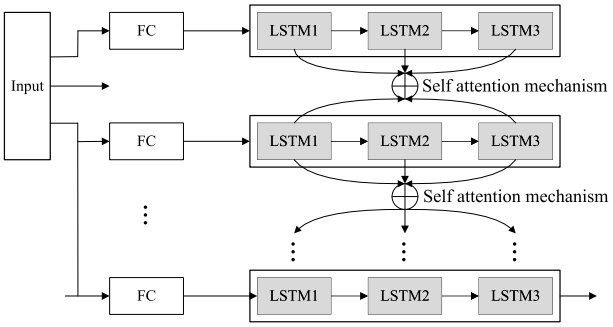

FIGURE 4. Flowchart for hierarchical attention model.

FIGURE.4 shows the hierarchical attention model, which consists of two parts: stacked LSTM and self-attention mechanism. The self-attention mechanism functions between multi-levels from the same time step of the stacked LSTM. The inputs are the hidden and unit state of the stacked LSTM, the outputs are weight vectors.

$$\boldsymbol{u}^t = \boldsymbol{v}^T \tanh\left(\boldsymbol{W}_s \boldsymbol{X}_t + \boldsymbol{b}\right) \tag{3}$$

$$\boldsymbol{a}^t = \text{Softmax}\left(\boldsymbol{u}^t\right) \tag{4}$$

The dimension of the vector $\boldsymbol{X}_t$ is $n \times r$,the dimension of the vector $\boldsymbol{W}_s$ is $r \times d_a$; $\boldsymbol{b}$ and $\boldsymbol{v}^T$ are all the vectors with the dimension of d_a (n is the number of samples for batch processing, r is the number of neurons in LSTM hidden layers, d_a is a hyper parameter that can be set randomly). $\boldsymbol{W}_s$, $\boldsymbol{b}$ and $\boldsymbol{v}^T$ are the parameters learned from the model. $\boldsymbol{X}_t$ is the input, it means the hidden state $\boldsymbol{H}_t$ or the unit state $\boldsymbol{C}_t$ of LSTM, as is given by:

$$\boldsymbol{H}_t = \left(h_t^{(1)}, h_t^{(2)}, \ldots, h_t^{(l)}\right) \tag{5}$$

$h_t^{(l)}$ is the corresponding hidden state in the l th LSTM layer at the tth time step.

$$\boldsymbol{C}_t = \left(c_t^{(1)}, c_t^{(2)}, \ldots, c_t^{(l)}\right) \tag{6}$$

$c_t^{(l)}$ is the corresponding unit state in the lth LSTM layer at the tth time step.

Then it gives the dot product between weight vector $\boldsymbol{a}^t$ and the state value of LSTM, as:

$$\boldsymbol{Y}_t = \boldsymbol{a}^t \boldsymbol{X}_t \tag{7}$$

$\boldsymbol{Y}_t$ respectively indicates the weighted hidden state $\boldsymbol{H'}_t$ and the weighted unit state $\boldsymbol{C'}_t$ of the update LSTM. The attention mechanism is used to calculate the weights, and then the hierarchical state is weighted according to the importance. The weighted state $\boldsymbol{Y}_t$ can dynamically capture the relationship between the hierarchical levels and improve the performance of the hierarchical characteristics.

C. PENALTY FACTORS

In order to prevent the self-attention mechanism from always providing similar sum weights for LSTM unit states between different time steps, a penalty index is required to encourage the diversity of weight vectors from different layers of the LSTM unit between multiple time steps. The hierarchical attention mechanism adaptively extracts the relationship between high-layer features and low-layer features, so the penalty factor can be considered as a kind of attention enhancement. That means, by optimizing the weight coefficients between features, the penalty index further strengthens the expression features at the interest level, weakens redundant features, and improves the quality during feature extraction.

The weight vector indicates a specific feature layer of the focused sequence frame. Multiple attention transfers are required at different time steps. As a result, an attention matrix is formed by all the attention vectors at different levels in each time step. In order to differentiate the hierarchical relationship, statistical variance is chosen to optimize the network weights, as is shown in below formula:

$$P = \frac{1}{T}\sum_{t}^{T}\sum_{i}^{L}\left(\left(\boldsymbol{\alpha}_{ti} - \mu\right)^2 + \left(\boldsymbol{\beta}_{ti} - \eta\right)^2\right) \tag{8}$$

$$\mu = \frac{1}{L}\sum_{i}^{L}\boldsymbol{\alpha}_{ti}, \quad \eta = \frac{1}{L}\sum_{i}^{L}\boldsymbol{\beta}_{ti} \tag{9}$$

$\boldsymbol{\alpha}_{ti}$ is the attention weight of the hidden sate $\boldsymbol{h}$ at layer i, time t. $\boldsymbol{\beta}_{ti}$ is the attention weight of the unit sate $\boldsymbol{c}$ at layer i, time t. L is the number of LSTM layers, μ and η represent the average value of the attention matrix in column t of hidden state $\boldsymbol{h}$ and unit state $\boldsymbol{c}$, respectively.

Variance indicates the dispersion degree of the data. A larger variance means more obvious difference between various weights. Therefore, variance is taken as a penalty factor to achieve the differentiation of weights between levels. Specifically, the penalty index P is multiplied by a coefficient λ ($\lambda > 0$), then it is given as a negative value and minimized together with the original loss function

$$L_d = -\log\left(p\left(y\,|a\right)\right) - \lambda P \tag{10}$$

$-\log\left(p\left(y\,|a\right)\right)$ is the cross entropy loss function, expresses the difference between the true sample label and the predicted probability;a is the actual output of the model, y is the sample label, λ is the penalty factor. $\lambda = 1$ is the hyperparameter which is determined by a random search.

D. REWARD FUNCTION

f_{n+1} is non-differentiable, thus a reinforcement learning is used for training based on policy gradient. For the specific sequence space $\boldsymbol{A}$, $p_\theta\left(\boldsymbol{\tau}\right)$ represents the distribution on $\boldsymbol{A}$ with the parameter $\boldsymbol{\theta}$, $\boldsymbol{\tau} \in \boldsymbol{A}$ is a set of state sequences. The reinforcement learning function is given

$$W\left(\boldsymbol{\theta}\right) = \sum_{\boldsymbol{\tau}\in\boldsymbol{A}} p_\theta\left(\boldsymbol{\tau}\right) r\left(\boldsymbol{\tau}\right) \tag{11}$$

$r(\boldsymbol{\tau})$ shows the reward that is brought by each sequence; $W(\boldsymbol{\theta})$ represents the expected reward under the possible sequence distribution. The network parameter $\boldsymbol{\theta}$ is learned to maximize the expected reward of the sequence f_{n+1}.

The gradient of the objective function is given as:

$$\nabla W(\boldsymbol{\theta}) = \sum_{\boldsymbol{\tau} \in \boldsymbol{A}} p_\theta(\boldsymbol{\tau}) \nabla \log p_\theta(\boldsymbol{\tau}) r(\boldsymbol{\tau}) \tag{12}$$

In general, in model-free reinforcement learning tasks, the policy gradient is estimated by sampling. Especially, the Monte Carlo strategy gradient algorithm is used in this paper. The basic idea of this method is continuous exploration, that means, the model explores the scene and generates a state-action sequence from the start to the end of the state according to the current strategy.

The Monte Carlo algorithm is used in sampling and approximate estimation, a sequence is obtained by random sampling according to the current strategy:

$$\nabla W(\boldsymbol{\theta}) \approx \frac{1}{M} \sum_{m=1}^{M} \nabla \log p_\theta(\boldsymbol{\tau}) r(\boldsymbol{\tau}) \tag{13}$$

The sequence is supposed as $\boldsymbol{\tau}^m = \left\{\boldsymbol{s}_1^m, \boldsymbol{\alpha}_n^m, \ldots, \boldsymbol{s}_N^m, \boldsymbol{\alpha}_N^m\right\}$, its likelihood probability is given by:

$$p_\theta\left(\boldsymbol{\tau}^m\right) = \prod_{n=1}^{N} P\left(\boldsymbol{s}_{n+1}^m \left| \boldsymbol{s}_n^m, \boldsymbol{\alpha}_n^m \right.\right) \boldsymbol{\pi}_\theta\left(\boldsymbol{\alpha}_n^m \left| \boldsymbol{s}_n^m \right.\right) \tag{14}$$

P represents the state transition probability; $\boldsymbol{\pi}_\theta$ represents the behavior strategy, the Gaussian strategy used in the training process. At the time point n, under the behavior sequence m, $\boldsymbol{s}_{n+1}^m$ indicates the policy state at the next time point. $\boldsymbol{\alpha}_n^m$ indicates the current behavior of the policy (ie. f_{n+1}), $\boldsymbol{s}_n^m$ indicates the current policy state. Consequently, the gradient expression of Monte Carlo strategy is as follows:

$$\nabla W(\boldsymbol{\theta}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{N} \nabla \log \boldsymbol{\pi}_\theta\left(\boldsymbol{\alpha}_n^m \left| \boldsymbol{s}_n^m \right.\right) R^m \tag{15}$$

R^m represents the reward that is obtained from the sequence m. The reward is obtained after the end of the entire behavior sequences, as shown in equation (16):

$$R' = \begin{cases} \lambda_{tp}, & \text{If correctly tested} \\ \lambda_{fp}, & \text{If error detection} \end{cases} \tag{16}$$

$\lambda_{tp}\,(>0)$, $\lambda_{fp}\,(<0)$ are reward values respectively for correct detection and false detection at each time point. Particularly, false detection will be highly punished in this paper. Meanwhile, considering the sparsity of effective emotional information, the sparsity constraint term $\lambda_{sparse} N < 0$ is also added in equation (17), to ensure fewer signals is observed while the highest accuracy is obtained.

$$R = \lambda_r R' + \lambda_{sparse} N \tag{17}$$

Here, λ_r is a reward factor, its value is larger than zero. λ_{sparse} is a sparsity factor, its value is less than zero. N is the length of behavior sequence. The basic idea of the strategy iteration is given as:

$$\boldsymbol{\theta} = \theta + \zeta \nabla W(\boldsymbol{\theta}) \tag{18}$$

Here, ζ is the step factor, refers to the learning rate of this algorithm.

V. EXPERIMENT AND RESULTS

To verify the effectiveness of the method in this paper, the DEAP dataset is used in our experiments. Emotion activity and effectiveness value are mainly identified in this section. Classification rate and F1-score are used as two evaluation indexes for recognition effects.

A. DATE COLLECTION

DEAP is a dataset for multimodal emotion recognition to analyze human sentiment states. It has multiple modal data, including EEG and peripheral physiological signals. This database is often used not only in comparison experiments about physiological signals, but also in research on EEG emotions. The DEAP dataset has 32 subjects. EEG and peripheral physiological signals are obtained through them. Each subject usually watches a one-minute long music video clip every time, 40 sets of music videos in total. As a result, 40 sets of physiological signals is collected at the same time. Each experimental signal includes 32 channels of EEG signals and 8 channels of peripheral physiological signals, details are: 2 channels of EOG signals, 1 channel of GSR signal, 2 channels of EMG signals, 1 channel of respiration recording, 1 channel of plethysmography, 1 channel of temperature recording. Each channel can be divided into a 3-second baseline and a 60-second emotion segment. There are wo sampling rates: 512Hz for original data and128Hz for preprocessed data. Dimension is used as emotional label, including: Valence, Arousal, Dominance, and Like/Dislike, which respectively represent value, arousal degree, advantage, like/dislike, and the scoring interval is [1], [9]. Valence means the satisfaction or pleasure, the higher the score, the more satisfied. The Arousal level, the higher the score means the subject is more likely to be awake, the lower the score means the subject is closer to sleep. The Dominance changes from low to high, indicating the subject turned from the passive to the active. Like/Dislike changes from low to high, indicating that the attitude to the video is turned from hate to like, but this label was rarely used by scholars. Instead, Valence and Arousal are the most commonly used. Therefore, verification experiments with and without baseline strategy is carried out based on these two scores.

B. EXPERIMENT SETUP AND EVALUATION INDEX

The DEAP data is divided into training set A, verification set A0 and test set B at a ratio of 5:1:1. During data preprocessing, the face video from the dataset is down sampled to 8fps. Meanwhile, the face image in the video is detected and reduced with the re-scaled image size of 227×227. In training process, the Adam algorithm was applied to update

the parameters. 12 samples (mini batch) from training set A is extracted through the experience playback mechanism, which is used as the sample set used for each update. To prevent overfitting, the dropout value is set to 0.6, and the value of the max time step $N_{\max}$ is 40.

Recognition accuracy and F1-score in this paper are used to evaluate the recognition results. Recognition accuracy rate (Classification rate, CR) refers to the percentage of the quantity of correct-classified samples and that of the total samples in the test set, as is shown in equation (12). F1-score is an index to measure the accuracy of multi-classification models in statistics. It can be regarded as a weighted average of model precision and recall, which can take the balance between the former and the latter.

$$CR = \frac{N_{TP}}{N_{data}} \tag{19}$$

$$CR = \frac{2N_{TP}}{2N_{TP} + N_{FP} + N_{FN}} \tag{20}$$

N_{data} is the sample quantity for emotion data in test set. N_{TP}, N_{FP} and N_{FN} are the quantity for correct-detections, fail-detections and miss-detections in test samples, respectively.

C. RESEARCH ON STACKED LSTM LAYERS

The investigation was conducted to study the quantitative effect of LSTM stack layers on the experimental results in a multi-layer architecture. FC is trained by Faster-RCNN and then used to extract image features. Then these features are put into the single-layer LSTM network without adding the layered attention mechanism. The output at the last time step is sent to the Softmax classifier for discrimination.

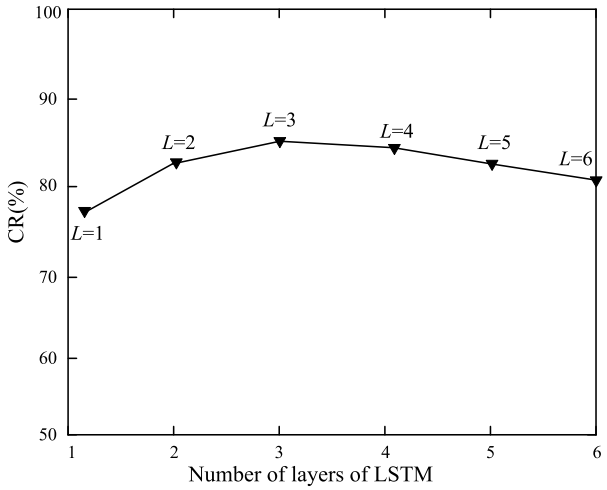

FIGURE 5. Relation between LSTM layer number and CR.

FIGURE. 5 shows the recognition results of the LSTM layer number under different values on the DEAP data set. It can be analyzed that: compared with the single-layer LSTM, multi-layer LSTM combines low-level features in a more descriptive manner, so that the higher LSTM layers can better extract abstract features from the sequence and have higher recognition performance. Particularly, when $L = 3$, the recognition is the best. But when the stacked LSTM has more than 3 layers, the performance gradually decreases because the saturation activation function is used. Too many layers cause the gradient disappearance, so that the weight of the shallow LSTM cannot be updated. Based on below recognition results on the DEAP data set, three-layer LSTM is selected in this paper.

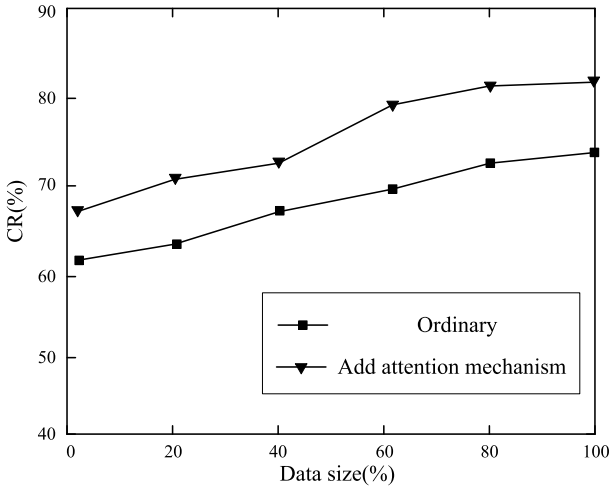

FIGURE 6. Relation between LSTM layer number and CR.

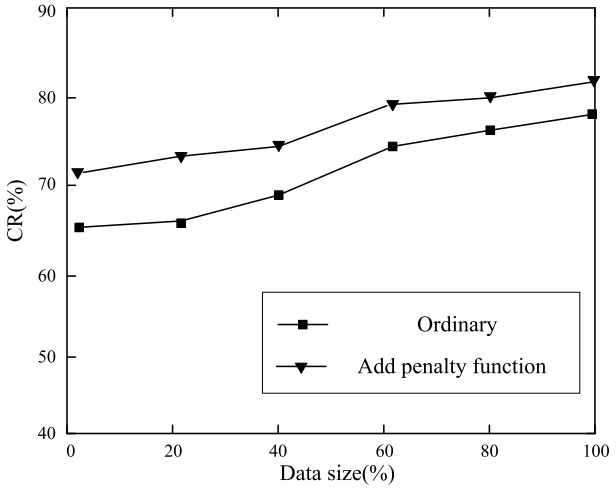

FIGURE 7. Classification rate comparison for penalty experiment.

D. HIERARCHICAL ATTENTION MECHANISM EXPERIMENT

The attention mechanism is used to selectively focus on different levels at various time steps. Since the attention mechanism between the levels has an effect on the experimental results, a comparative experiment is designed to do further analysis. The results in FIGURE. 6 shows: with the common stacked LSTM structure, the recognition rate of 85.9% is achieved on the DEAP dataset. This indicates that the hierarchical attention mechanism performs well in extracting the effective information from the video frame.

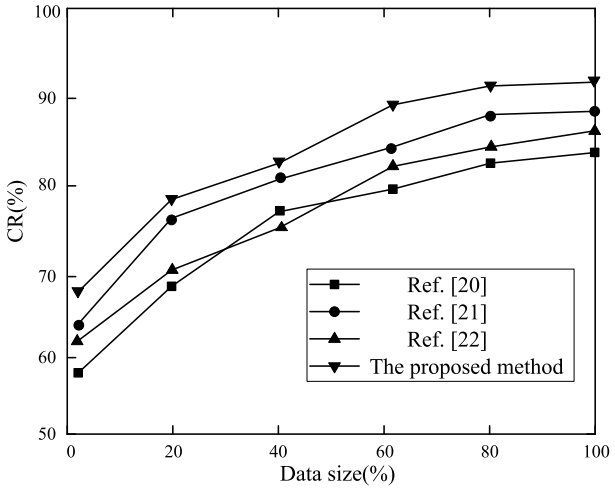

FIGURE 8. Comparison of classification rate about different emotion recognition methods.

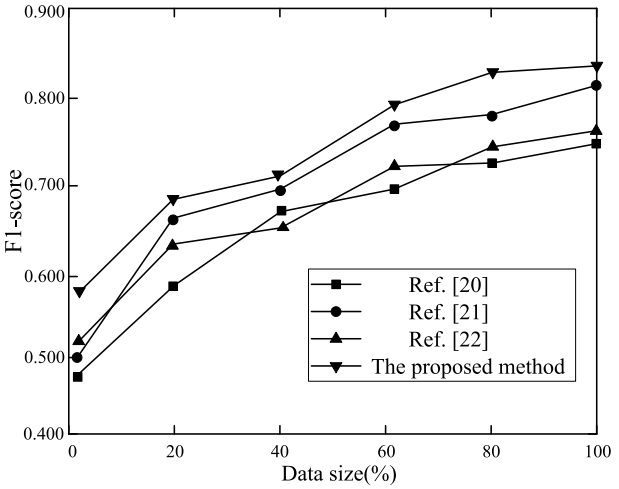

FIGURE 9. Comparison of F1-score about different emotion recognition methods.

The layered architecture of LSTM creates a hierarchical feature representation of input data. The attention mechanism adjusts the information proportion between different levels, which helps to select information that is more valuable to the current task goal. Then more attention is put on the more critical level. The results prove that the classification rate with the attention mechanism has been fully improved.

E. PENALTY EXPERIMENT

In the attention extraction, attention weights represent the relationship between various levels. By introducing a penalty factor, the combination of weight vectors can be adjusted, then there are more various ways to combining attentions between frame-sequence levels. Especially, variance is used to measure the difference between attention weights. It is maximized through back propagation. FIGURE.7 compares the recognition rate of the model before and after adding penalty factors. As is shown in FIGURE.7, the penalty factor improves the overall performance of the network and improves the recognition rate to a certain extent.

F. COMPARISON WITH OTHER ALGORITHMS ON EXPRESSION-EEG MULTIMODAL EMOTION RECOGNITION

The proposed model is respectively compared with the model in Ref [20] Ref [21] and Ref [22]. The results are shown in FIGURE.8 and FIGURE.9.

It can be seen from FIGURE.8 and FIGURE.9 that the classification accuracy rate (CR) and F1-score of the proposed model are significantly higher than other methods. Particularly, CR and F1-score increased by at least 2% and 0.015 respectively. Recognition result is improved because the remaining methods directly analyze the multi-modal emotional signals containing a lot of redundant information. But the proposed model in this paper introduces an information attention mechanism to compress the redundant information and improve the accuracy. In addition, the proposed model analyzes time series of human emotions from each modal in a "selective focus" way, and predicts the time information of the key signal frame at the next time point through the time-domain attention mechanism, which effectively uses the time-domain information from emotion data.

VI. CONCLUSION

An emotion recognition model of video-EEG signal interaction based on hierarchical LSTM. Taking the advantages of the collaboration and complementarity for the two models, Human emotions can be accurately identified by proposed method. In order to fully use the key spatial information in EEG signals, EEG signals are firstly converted into image sequences. Then the importance, corresponding to three frequency bands(α, β and θ) from EEG signals, is calculated by the spatial frequency band attention mechanism. To use the key time information from the data effectively, the time-domain attention mechanism is introduced to automatically locate critical signal frames. The experimental results on the two datasets show that the proposed model achieves a higher classification rate and performs better in recognition.

However, the human emotional state in natural scenes is different from a specific dataset and will change with time. Under the premise of guaranteeing recognition effect, how to identify different emotional states in one emotion segment is still a focus need to be studied in the future.

REFERENCES

[1] 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)*, Cambridge, U.K., Sep. 2019, pp. 552–558.
[2] P. Salovey and J. Mayer, "Emotional intelligence," *Imag., Cogn. Pers.*, vol. 9, no. 3, pp. 185–211, 1990.
[3] B. Liebold and P. Ohler, "Multimodal emotion expressions of virtual agents, mimic and vocal emotion expressions and their effects on emotion recognition," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Geneva, Switzerland, Sep. 2013, pp. 405–410.
[4] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Transfer of multimodal emotion features in deep belief networks," in *Proc. 50th Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2016, pp. 449–453.

[5] A. S. Patwardhan, "Multimodal mixed emotion detection," in *Proc. 2nd Int. Conf. Commun. Electron. Syst. (ICCES)*, Coimbatore, India, Oct. 2017, pp. 139–143.
[6] S. Thushara and S. Veni, "A multimodal emotion recognition system from video," in *Proc. Int. Conf. Circuit, Power Comput. Technol. (ICCPCT)*, Nagercoil, India, Mar. 2016, pp. 1–5.
[7] N. Vryzas, L. Vrysis, R. Kotsakis, and C. Dimoulas, "Speech emotion recognition adapted to multimodal semantic repositories," in *Proc. 13th Int. Workshop Semantic Social Media Adaptation Personalization (SMAP)*, Zaragoza, Spain, Sep. 2018, pp. 31–35.
[8] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019.
[9] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.
[10] 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./Mar. 2012.
[11] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1175–1191, Oct. 2001.
[12] D. Nie, X.-W. Wang, L.-C. Shi, and B.-L. Lu, "EEG-based emotion recognition during watching movies," in *Proc. 5th Int. IEEE/EMBS Conf. Neural Eng.* Cancún, Mexico: IEEE, Apr. 2011, pp. 667–670.
[13] R. N. Duan, X. W. Wang, and B. L. Lu, "EEG-based emotion recognition in listening music by using sup-port vector machine and linear dynamic system," in *Proc. Int. Conf. Neural Inf. Process.* Doha, Qatar: Springer, 2012, pp. 468–475.
[14] W.-L. Zheng, H.-T. Guo, and B.-L. Lu, "Revealing critical channels and frequency bands for emotion recognition from EEG with deep belief network," in *Proc. 7th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Apr. 2015, pp. 154–157, doi: 10.1109/NER.2015.7146583.
[15] P. Ekman, "An argument for basic emotions," *Cognition Emotion*, vol. 6, nos. 3–4, pp. 169–200, May 1992.
[16] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data," *IEEE Trans. Affect. Comput.*, vol. 2, no. 4, pp. 206–218, Oct. 2011.
[17] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*. New York, NY, USA: IEEE, Mar. 2016, pp. 1–9.
[18] G. K. Verma and U. S. Tiwary, "Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals," *NeuroImage*, vol. 102, pp. 162–172, Nov. 2014, doi: 10.1016/j.neuroimage.2013.11.007.
[19] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos (Extended abstract)," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*. Xi'an, China: IEEE, Sep. 2015, pp. 491–497.
[20] Y. R. Huang J. Yang, P. Liao, and J. Pan, "Fusion of facial expressions and EEG for emotion recognition," *Comput. Syst. Appl.*, vol. 27, no. 2, pp. 9–15, 2018.
[21] D. Li, Z. Wang, C. Wang, S. Liu, W. Chi, E. Dong, X. Song, Q. Gao, and Y. Song, "The fusion of electroencephalography and facial expression for continuous emotion recognition," *IEEE Access*, vol. 7, pp. 155724–155736, 2019, doi: 10.1109/ACCESS.2019.2949707.
[22] M. A. Islam, A. Hamza, M. H. Rahaman, J. Bhattacharjee, and M. M. Rahman, "Mind reader: A facial expression and EEG based emotion recognizer," in *Proc. 2nd Int. Conf. Comput. Methodologies Commun. (ICCMC)*, Erode, India, Feb. 2018, pp. 101–107, doi: 10.1109/ICCMC.2018.8487898.
[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[24] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," in *Proc. Int. Conf. Learn. Represent.* San Juan, Puerto Rico: ICLR, 2016, pp. 1–15.
[25] L. Chang and L. Qin-Rang, "Using reinforce learning to train multi attention model," *Acta Automatica Sinica*, vol. 43, no. 9, pp. 1563–1570, 2017.