IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 23, 2020, accepted August 27, 2020, date of publication September 7, 2020, date of current version September 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3021994

Expression-EEG Based Collaborative Multimodal Emotion Recognition Using Deep AutoEncoder

HONGLI ZHANG
Department of Educational Technology, Inner Mongolia Normal University, Hohhot 010022, China
e-mail: zhanghl_1974@163.com

This work was supported in part by the Key Technology Project of Inner Mongolia Autonomous Region under Grant 2020GG0170, and in part by the Natural Science Foundation Project of Inner Mongolia Autonomous Region under Grant 2015MS0634.

ABSTRACT Emotion recognition has shown many valuable roles in people's lives under the background of artificial intelligence technology. However, most existing emotion recognition methods have poor recognition performance, which prevents their promotion in practical applications. To alleviate this problem, we proposed an expression-EEG interaction multi-modal emotion recognition method using a deep automatic encoder. Firstly, decision tree is applied as objective feature selection method. Then, based on the facial expression features recognized by sparse representation, the solution vector coefficients are analyzed to determine the facial expression category of the test samples. After that, the bimodal deep automatic encoder is adopted to fuse the EEG signals and facial expression signals. The third layer of BDAE extracts features for training of supervised learning. Finally, LIBSVM classifier is used to complete classification task. We carried out experiments on a constructed video library to verify the proposed emotion recognition method. The results show that the proposed method can effectively extract and integrate high-level emotion-related features in EEG and facial expression signals. The recognition rate of discrete emotion state type and the average emotion recognition rate have been improved relatively, in which the average emotion recognition rate is 85.71%. Overall, the emotion recognition ability has been greatly improved.

INDEX TERMS EEG signals, facial expression features, multi-modal fusion, deep automatic encoder, emotion recognition, decision tree, sparse representation.

I. INTRODUCTION

Emotion is very complex mental state or process of human beings, which can reflect human perceptions and attitudes and play an important role in the communication between people [1]. The research of emotion recognition has very important value in the application of human-computer interaction [2]. The environment in human-computer interaction system is complex and dynamic. In many occasions, it needs coordinate operations with people, therefore system with emotional interaction ability can better adapt to such environment. If the human-computer interaction system can quickly and accurately identify human emotions, the interaction process will be more friendly and natural [3]. EEG signals have the strong ability to characterize changes in human brain state, so emotion recognition based on EEG signals has become a popular trend in research.

With the development of multimedia and human-computer interaction technology, it is of great significance to automatically recognize people's emotional states [4]. For example, in the process of playing a game, the difficulty of the game can be adjusted by identifying the emotional state. In a negative emotional state, the system will provide a simple and cheerful game for relaxing, which is easier to pass through. When being in a positive emotional state, it will increase the game difficulty, which can bring more challenging to people's game experience [5]. In addition, emotions can also be adjusted through the recommendation of music and video. When watching videos or listening to musics, one can alleviate the effects of negative emotions by pushing positive multimedia content. With the development of artificial intelligence, technological research of robots has developed rapidly. Nowadays, it has been applied in many fields, like household appliances, food and other industries. As the robots enter all aspects of daily life, people have put forward higher requirements for them. In terms of humanization, it is hoped

The associate editor coordinating the review of this manuscript and approving it for publication was Juan M. Gorriz.

that the robot can have human brain thinking and emotional colors, which can recognize the emotional state of the human to achieve better human-computer interaction. The guarantee of emotional recognition rate is one of the key technologies that emotional robots can put into practical application, which is also great significance to its realization [6].

In the medical field, emotion recognition is also very important. For patients, the quality of the emotional state will have a great impact on the disease development process and the corresponding treatment management system [7]. Although current medical research has no definite evidence to prove the relationship between emotional states and diseases, the positive emotional states are actually conducive to disease recovery and physical mental health. When patients are in a negative state and does not cooperate with treatment, the cure of disease is usually very slow, and may even lead to the disease exacerbation. Therefore, the detection of emotional state is relatively important for patients. As much as possible to put the patients in a positive emotional state can promote the good development of the disease. Meanwhile, emotion recognition is also helpful for the prevention and treatment of depression and other diseases [8].

In the field of public safety, emotion recognition also has certain practical value. Polygraph is an important tool for public security personnel to interrogate suspects [9]. In the process of criminal suspect's statement, the corresponding emotional state can be judged based on the physiological signal, which will serve as a basis for the truth of the statement. In addition, emotion recognition can be used in teaching management. Through wearing a harmless portable device with emotion recognition function, teathers can detect students' emotional states in real time. When student's mood is relatively poor or even shows extreme behavior, the teacher can timely communicate with the student and their parents to avoid tragedy [10].

Since the theory of sentiment computing was proposed, its related theories and analytical methods have developed rapidly, and the research on sentiment recognition has also been focused by researchers [11]. In this paper, a multi-modal emotion recognition method based on expression-EGG interaction using deep autoencoder is proposed, and experiments on the constructed video library verify the effectiveness of our method.

II. RELATED WORKS

The features used in traditional emotion recognition methods are mainly external features such as facial expressions, body postures and speech [12]. There is unnecessary to wear sensors for obtaining these signals, which has the advantages of easy acquisition and low cost. [13] utilized the people facial expressions in the videos for emotion recognition. The authors applied Naive Bayes algorithm to recognize seven different emotions including happiness, surprise, anger, disgust, fear, sadness, and neutrality. The emotion recognition accuracy rate between facial expressions of different people is 64.3%, while testing the same person achieves an accuracy rate of 93.2%, indicating that facial expressions can be adapted to effectively recognize emotions. [14] combined acoustic features and speech content for emotion recognition based on speech signals. A support vector machine-belief network architecture is used to subdivide six different emotions of anger, disgust, fear, neutrality, sadness and surprise, and the recognition accuracy is up to 93%. These experiment results confirmed the effectiveness of speech signals for emotion recognition. However, these signals are relatively sensitive, which are easily affected by the subjective factors of the testers. The system cannot make a correct judgment when the subject's inner true emotions and external performance are inconsistent. Meanwhile, pure external performance is only a part of emotional performance, which cannot express the rich emotions of human beings. The physiological changes are dominated by the central nervous system of the person, which can more objectively reflect the emotional state of the person. Therefore, the use of human physiological signals for emotion recognition is currently a novel international research trend in emotion computing.

Currently, researchers usually use Electroencephalogram (EEG), Electromyogram (EMG), Galvanic Skin Response (GSR), Electrooculogram (EOG), Electrocardiogram, (ECG), blood pressure, blood volume pulse (BVP), epidermal temperature, eye movement signals and other physiological signals for emotional recognition research [15]. The emotion recognition method based on physiological signals can achieve high accuracy while collected data can objectively reflect the emotional state of the subjects [16]. According to the different signals adopted, existing emotion recognition methods can be roughly divided into EEG signals based, facial video features-based, and multi-modal emotion recognition.

A. EMOTION RECOGNITION METHOD BASED ON EEG SIGNALS

At present, there are many researches on EEG-based emotion recognition, which have proved the effectiveness of EEG signals for emotion recognition. The attention of scholars mainly focus on aspects of feature extraction, feature selection and classification model selection of EEG. EEG-based emotion recognition methods are generally divided into two categories [17], [18]: one is supervised learning, which is trained and tested based on emotion labels, such as KNN, Fisher algorithm and etc. The other is unsupervised learning method, in which the sample data do not contain labels. Professional researchers will automatically divide all samples into different categories according to a certain strategy, and then assign corresponding labels. Unsupervised learning methods usually include K-means, fuzzy C-means (FCM), and self-organizing maps.

In the case of facial expression recognition, [19] proposed the utilization of a deep learning network (DLN) to discover unknown feature correlation between input signals that are crucial for the learning task. The DLN is implemented with a stacked autoencoder (SAE) using hierarchical feature learning approach. Input features of the network are power spectral densities of 32-channel EEG signals from 32 subjects. [20] shown 40 patients with Parkinson's disease (PD) from China and 40 healthy controls, and designed 24 black/white portraits and 24 music excerpts to express happiness, sadness, fear and anger. Four tests were used to evaluate participants' executive functions, including trace production test (TMT), clock drawing test (CDT), semantic spoken fluency test (VFT) and digital span test (DST). The experimental results shown that the recognize ability for anger face in PD group was impaired. It may be related to executive dysfunction, while shown better performance in recognizing musical emotions. To further improve the accuracy of the CNN-based modules, [21] devised a multi-column structured model, whose decision is produced by a weighted sum of the decisions from individual recognizing modules. We apply the model to EEG signals from DEAP dataset for comparison and demonstrate the improved accuracy of our model. [22] proposed a real-time emotion recognition hardware system architecture with EEG, which performing binary and quaternary classification in multiphase convolutional neural network (CNN) algorithm based on a 28-nanometer technology chip and a field programmable gate array (FPGA). Sample entropy, differential asymmetry, short-time Fourier transform and channel reconstruction methods are used for emotional feature extraction. EEG signal features can be divided into time domain features, frequency domain features and time-frequency features. Time domain features are mainly the statistical features of the signal. Frequency domain characteristics mainly include frequency band energy and higher order spectrum characteristics (HOS).

According to the asymmetric effects of emotions, differential asymmetric features and ratio asymmetric features can be extracted. Experimental results show that the performance of multiple feature selection strategies is better than the univariate method. In addition, the emotion recognition accuracy rate of EEG features obtained by more advanced extraction algorithms is higher than band spectrum features obtained by traditional methods.

B. EMOTION RECOGNITION METHOD BASED ON FACIAL VIDEO FEATURES

Emotions are caused by specific scenes, the process is usually very complicated. Emotion recognition needs to consider the specific situation, and cannot be simply judged by the external performance, while ignoring the content attributes and semantics of emotions [23]. Eye-tracking signals can provide various indicators of eye activity. They can guide system observe the subtle subconscious behavior of users, which provide important reference for the user's current activity context [24]. The technique of recording individual's eye movement is called eye tracking. In the basis of this technique, one can determine where the subject is looking at a certain moment, and also get the eye's movement trajectory during a certain period of time. [25] proposed a novel multi-pattern correlation network for emotion recognition, which aims to achieve more powerful and accurate detection by the combination of audio and video channels information. This method first preprocessed the audio and visual signal for feature extraction and then obtain the Mel spectrogram, which will be treated as image to obtain representative frames from visual segments. The Mel spectrogram and representative frames are then fed to a CNN to obtain audio features. [26] studied the reduction of facial expressions that might lead to autism spectrum disorder (ASD) and caused impaired emotion recognition and expression. On this basis, their algorithm evaluated the purpose of reducing facial emotion recognition (FER) deficiencies in the acceptability, feasibility, and initial efficacy of attention-correcting interventions.

C. MULTIMODAL EMOTION RECOGNITION METHOD

At present, multi-modal fusion emotion recognition method is introduced to further improve the accuracy of emotion recognition. Multimodal fusion model can obtain emotion recognition results by fusing different physiological signals. [27] is aimed to analyze the performance of a Convolutional Neural Network which uses AutoEncoder Units for emotion recognition in human faces. The combination of two Deep Learning techniques boosts the performance of the classification system. 8000 facial expressions from the Radboud Faces Database were used during this research for both training and testing. The outcome showed that five of the eight analyzed emotions presented higher accuracy rates, higher than 90%. [28] completed the emotion recognition (ER) task and self-reported depression severity based on 644 outpatients (57.6% were female and the average age was 31.31). The study used 10 items specific to unipolar depression identified through factor analysis. 34.6% of the participants had clinical depression, while all other participants had clinical anxiety or other unspecified emotional disorders. In a large number of diagnosed clinical samples of adults with emotional disorders, we found that the accuracy of ER showed a decline with age increases, especially for negative emotions such as sadness and fear. [29] proposed a method of facial expression recognition during the speech process. The method uses a hybrid deep network architecture to perform multi-modal fusion of EEG signals and facial expressions. Experimental results prove the emotion recognition rate in multi-modal fusion model is higher than the model in each individual modal. [30] analyzed the epoch data from the EEG sensor channel, and performed a variety of machine learning technology tests including Support Vector Machine (SVM), K nearest neighbor, linear discriminant analysis, logistic regression, and decision tree. Whether to use Principal Component Analysis (PCA) for dimension reduction. Grid search is also used to adjust the hyperparameters of each tested machine learning model through the Spark cluster to shorten the execution time. [31] proposed a speech emotion recognition algorithm based on the superposed sparse depth model. The improvement of this algorithm is based on the automatic encoder, denoising automatic encoder and sparse automatic encoder. The first layer structure uses a noise reduction autoencoder to learn hidden features, whose dimension is larger than that of the input features. The second layer uses a sparse autoencoder to learn sparse features. Finally, the wavelet kernel sparse SVM classifier is used to classify the features, but how to further improve the emotion recognition rate of multi-modal fusion is also worth considering.

For the problems of single modality and low accuracy in the above researches, a multi-modal emotion recognition method based on expression-EEG interaction using deep autoencoder is proposed. The innovations are summarized as follows:

1) To accurately obtain facial expression features, the proposed method recognizes facial expression features based on sparse representation, and uses orthogonal matching tracking algorithm to analyze the solution vector coefficients and to determine the facial expression category of the test sample.

2) Different from the single modal recognition problem, our proposed method uses dual-modal depth automatic encoder (BDAE) to fuse EEG signals and facial expression signals, and inputs them to supervised learning LIBSVM classifier framework to get an emotion classification model.

III. OVERALL FRAMEWORK AND FEATURE EXTRACTION

A. FRAMEWORK OF THE PROPOSED METHOD

EEG signals have a strong ability to characterize changes in human brain state, so emotion recognition based on EEG signals has been a common studying method. Besides the EEG signals, another external physiological characterization signal—facial expression signal is added for emotion recognition. Deeply exploring the ability of EEG signals and facial expression signals is to distinguish and characterize different emotions, and combining EEG and facial expression signals through different model fusion strategies (including deep neural networks) is to establish a multi-modal emotion recognition model combining with internal neural models and external subconsciousness behaviors. The framework is shown in FIGURE 1 [32].

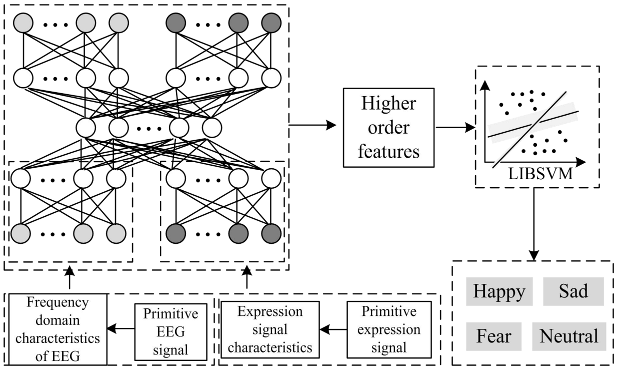

Higher order features
LIBSVM
Happy
Sad
Fear
Neutral
Frequency domain characteristics of EEG
Primitive EEG signal
Expression signal characteristics
Primitive expression signal

FIGURE 1. Multimodal emotion recognition framework.

Meanwhile, to research the stability of express emotions ability for EEG and facial expression signals over time, BDAE is used for model fusion. The accuracy of emotion recognition model has been effectively improved after fusion of EEG signals and facial expression signals. After training process, the middle layer (i.e., the third layer of BDAE, a total of five layers) is used as the extracted features, and then send them to the LIBSVM classifier for supervised learning training to get the emotion classification model. In this way, the emotion recognition ability of the model will be significantly improved, which is benefit to the high-order features related to emotions in the two signals extracted by the deep neural network.

B. EEG SIGNAL FEATURE SELECTION

To alleviate the shortcomings of the traditional EEG signal feature selection methods, a decision tree-based EEG signal feature selection method is proposed to achieve both objectivity of feature selection and high classification accuracy.

1) PRINCIPLE OF DECISION TREE

Decision tree is one of the most classic and commonly used algorithms in data mining. Compared with other data mining algorithms, the decision tree has three advantages [33]: (1) The decision tree is a very easy-to-understand algorithm; (2) In process of training the decision tree, there is no need for researchers to understand the relevant background knowledge of the training data; (3) The classification accuracy of the decision tree is relatively high. Considering the three advantages of decision trees, researchers usually use decision tree algorithms to conduct data classification studies.

The decision tree uses a "divide and conquer" greedy algorithm, outputting a tree-like structure. The structure of the decision tree is shown in FIGURE 2. Each non-leaf node stores a split attribute, branches are divided according to different attribute values of the split attribute, and each leaf stores a category label. Completing the decision tree algorithm requires two steps: constructing the decision tree; pruning the complete decision tree to form a simplified decision tree.

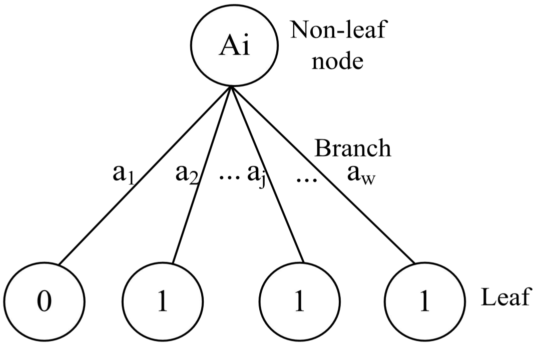


FIGURE 2. Schematic diagram of decision tree structure.

2) EEG SIGNAL FEATURE SELECTION BASED ON DECISION TREE

EEG signal is a kind of non-stationary, non-linear, high-dimensional weak physiological signal, and the decision tree has three advantages: easy to understand, no relevant background knowledge during training, and high classification accuracy. Therefore, using decision trees can not only objectively select features from high-dimensional and complex EEG signals, but also improve classification accuracy [34]. The proposed framework of EEG signal feature selection method based on decision tree is shown in FIGURE 3.

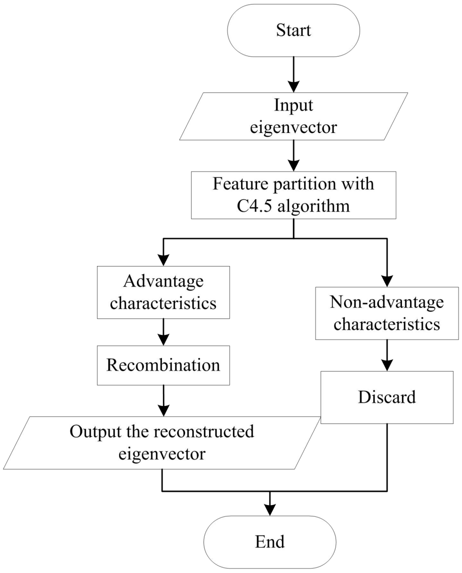

Start
Input eigenvector
Feature partition with C4.5 algorithm
Advantage characteristics
Non-advantage characteristics
Recombination
Discard
Output the reconstructed eigenvector
End

FIGURE 3. Flow chart of EEG feature selection method based on decision tree.

The method is divided into the following 5 steps:

Step 1. Input the EEG signal feature vector after performing feature extraction operation;

Step 2. Apply decision tree C4.5 algorithm to divide the input EEG signal feature vector into dominant features and non-dominant features;

Step 3. Discard non-dominant features from the input EEG signal feature vector;

Step 4. Recombine the dominant features from the input EEG signal feature vectors to obtain new feature vectors;

Step 5. Output the restructured feature vector after feature selection operation.

Step 2 will be discussed in detail. For EEG signal feature vectors after feature extraction X, X contains N samples, denoted as $X = \{x_1, x_2, \cdots, x_N\}$. Each sample x_i contains d dimensional features, denoted as $x_i = \{v_{i1}, v_{i2}, \cdots, v_{id}\}\,(i = 1, 2, \cdots, N)$. Regarding d dimensional features as d attributes, the candidate attribute set A can be denotes as $A = \{A_1, A_2, \cdots, A_d\}$. Each attribute contains M $(M \leq N)$ different values, that is $A_k = \{v_{1k}, v_{2k}, \cdots, v_{Mk}\}\,(k = 1, 2, \cdots, d)$. Building the feature sample set based on the method flow.

Initially, feature vector sample set X is treated as training set, calculating the information entropy of X:

$$IE\,(X) = -\sum\nolimits_{j=1}^{l} p_j \log_2 \left(p_j\right) \tag{1}$$

In formula, p_j is the probability that training set X belongs to category j, l is the number of all categories.

For the given attribute A_k, the information entropy of training set X is:

$$IG\,(X, A_k) = IE\,(X) - \sum\nolimits_{v \in Values(A_k)} \frac{|X_v|}{|X|} IE\,(X_v) \tag{2}$$

In the formula, $Values\,(A_k)$ is a collection of different values contained in attribute A_k, X_v is the subset of X, all samples of X_v on attribute A_k are value v, and in the set of $X - X_k$, all samples of attribute A_k are not v.

For given attribute A_k, split information of training set X is:

$$IS\,(X, A_k) = -\sum\nolimits_{j=1}^{c} \frac{|X_j|}{|X|} Ib \left(\frac{|X_j|}{|X|}\right) \tag{3}$$

In the formula, X_j is a subset of X, all samples in X_j are belong to j, and in the set of $X - X_k$, all samples are not belong to j.

For given attribute A_k, the information entropy of training set X is:

$$IGR\,(X, A_k) = \frac{IG\,(X, A_k)}{IS\,(X, A_k)} \tag{4}$$

In the process of constructing a complete tree, the attribute that maximizes the information entropy gain rate of the training set is selected as the split attribute, and then the branches are divided according to different values, each branch is operated recursively. When pruning a complete tree, the post-pruning method is selected to form a simplified decision tree.

The split attribute of all nodes included in the simplified decision tree is defined as the dominant attribute. After removing dominant feature from the initial feature, the remaining features are non-dominant features. For the initial feature vector X, reorganize them according to the dominant feature to form a new feature vector. This recombined feature vector is selected after applying the feature selection method based on the decision tree.

C. FACIAL EXPRESSION FEATURE EXTRACTION

Facial expression is a reflector of visual information, which is a key part of conveying emotions. Human face and facial emotions can reflect the different emotional states of peoples. For example, when people are in a happy state, their lips are often open, the corners of their mouths are raised, and their eyes become smaller. When people are in an angry state, they often open their eyes and frown, lock eyebrow, twitch zygomatic muscles. Recognizing these states by computers is facial expression recognition.

1) BASIC PRINCIPLE OF SPARSE REPRESENTATION

The model of facial expression recognition based on the sparse representation first composes the training sample set as a sparse dictionary, and then solves the most sparse solution vector for the test sample [35]. Finally, the analysis of solution vector coefficients determines the facial expression category of the test sample. The specific implementation process is as follows:

Step 1. Pretreatment phase. Input four types of facial expression samples: angry, depressed, happy, and normal. Form the u_g training samples corresponding to the g category into a matrix B_g. Assume the size of each sample $b_{g,h}$ is $m \times n$ (Pixel) grayscale image, where $b_{g,h} \in R^{m\times n}$.

Step 2. Stacking each column of sample $b_{g,h}$ to form column vector $u_{g,h}$, $u_{g,h} \in R^{m\times n}$, new matrix is denoted as $B_g = \left[u_{g,1}, u_{g,2}, \cdots, u_{g,n_i}\right] \in R^{m\times n}$. Each column represents the facial expression training sample of the g th object, A total of n training samples of type k form a total training sample set matrix B:

$$\begin{aligned} B &= \left[B_1, B_2, \cdots, B_g, \cdots, B_k\right] \\ &= \left[u_{g,1}, \cdots, u_{g,n_i}, \cdots, u_{k,n_k}\right] \end{aligned} \tag{5}$$

Step 3. After obtaining matrix B_g, for test samples, first convert it to a column vector form, and any g-th object can be approximated by a linear combination of training samples of this category:

$$y = a_{g,1}u_{g,1}a_{g,2}u_{g,2}, \cdots, a_{g,h}u_{g,h} \tag{6}$$

In the formula, $a_{g,h} \in R$, $h = 1, 2, \cdots, n_g$, $y \in R^m$. The test sample y is shown as:

$$y{=}Ax, x{=}\left[0, \cdots, 0, a_{g,1}, a_{g,2}, \cdots, a_{g,n_h}, 0, \cdots, 0\right] \tag{7}$$

In the formula, x is the expansion coefficient vector of test sample y relative to the total training sample set B:

$$\begin{Bmatrix} y_1 \\ y_2 \\ \vdots \\ y_g \end{Bmatrix} = \begin{Bmatrix} u_{11} & u_{12} & \cdots & u_{1h} \\ u_{21} & u_{22} & \cdots & u_{2h} \\ \vdots & \vdots & \vdots & \vdots \\ u_{g,1} & u_{g,2} & \cdots & u_{g,h} \end{Bmatrix} \cdot \begin{Bmatrix} a_{g1} \\ a_{g2} \\ \vdots \\ a_{gh} \end{Bmatrix} \tag{8}$$

Step 4. Finding the approximate sparse solution $\hat{x}$ of $y = Ax$ is to get the emotional state of different facial expressions.

The variables number in the system of equations is greater than the number of equations, therefore, the solution of y is not unique. Since the sparsity is defined by the 0 norm, it can be solved using the L_0 norm minimization method, as shown below:

$$\hat{x} = \arg\min \|x\|_0 \quad \textit{subjec to } y = Ax \tag{9}$$

In the formula, $\| \|_0$ is the L_0 norm of vector, representing the number of non-zero elements in vector A x. According to recent studies on compressed sensing, when the coefficient vector x is sufficiently sparse, it can be solved using the L_1-norm approximation method, as follows

$$\hat{x} = \arg\min \|y - V_t x\|_1 \quad \textit{subject to } y = Ax \tag{10}$$

In the formula, $\| \|_1$ is the L_1 norm of vector, representing the sum of the absolute values of the elements in vector x.

2) ORTHOGONAL MATCHING PURSUIT ALGORITHM

There are multiple solutions to equation (9), including matching pursuit (MP) algorithms, orthogonal matching pursuit algorithms (OMP) and etc. The OMP algorithm is used to solve the sparse coefficients, through repeated iterations to select the column vector in the training matrix that has a high correlation with the residual signal. The sparse solution of the test sample can effectively recognize different facial expressions. The pseudo code of the OMP algorithm is shown in Algorithm 1.

Algorithm 1 Pseudo Code of OMP Algorithm

```
Begin
1. Approximate sparse solution for $y = Ax$ is $\hat{x}$. First selecting the element with the highest correlation with the residual $r_0 = y$ in the matrix A: $n_t = \arg\max(r_{t-1}, v_i)$, $i = 1, 2, \cdots, N$, $t$ is the iteration numbers, $t = 1$, Index set $V = \phi$.
2. Update selected column space: $V_t = \lfloor V_{t-1}, v_{n_i} \rfloor$, and update the sparse system values of the selected columns $\hat{x}$.
$\hat{x} = \arg\min \|y - V_t x\|_1$
3. Update margin: $r_t = y - V_t\hat{x}$, Set the maximum residual value $\theta$
4. $t = t+1$, judging $r_t < \theta$, when satisfied, stop and output $\hat{x}$, otherwise go back to the first step.
5. The sparse coefficient combination of the sample to be tested can be obtained based on the above steps. Analyzing them can determine the emotional category number of the sample corresponding to the item with the smallest residual in the training set, that is, the emotional state of the test sample.
End
```

It can be seen from the specific algorithm that the sparse coefficient after solving not only excludes the interference of most samples of different types, but also can find the most similar sample category among a small number of samples, which greatly reduces the false judgment rate, and can effectively solve the image occlusion, lighting and other issues. Benefiting from the multiple advantages of the sparse representation classifier, it has been used in many modal recognition fields.

IV. MULTIMODAL EMOTION RECOGNITION BASED ON DEEP NEURAL NETWORK

In recent years, deep neural networks have been used in extensive researches, and shown superiority in the applications of artificial intelligence such as speech recognition, natural language processing, and image recognition. However, there is a lack of research on the use of deep neural networks for emotion recognition. Recently, Liu *et al.* used deep neural networks to combine EEG signals and eye tracking signals for emotion recognition. It has found that using deep neural networks to fuse the two signals can effectively improve the accuracy of emotion recognition, and has better performance than traditional model fusion methods. This observation inspired us to use deep neural networks to fuse EEG signals and facial expression signals, so as to further improve the accuracy of emotion recognition [36]. The following parts will introduce the details of deep neural network model: Bimodal Deep Auto-Encoder.

A. RESTRICTED BOLTZMANN MACHINE

Restricted Boltzmann Machine (RBM) was originally proposed by Smolensky and *et al.* It is a kind of undirected graph model that contains a visible layer and a hidden layer, which is shown in the FIGURE 4.

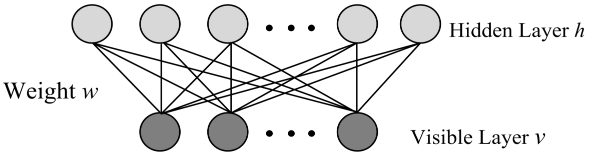


FIGURE 4. RBM model diagram.

The observation data corresponds to the visible layer, and the extracted features correspond to the hidden layer, which can be regarded as a feature detector. The model has no edge connections between nodes on the same layer. Assuming that visible layer variable is $\boldsymbol{v} \in \{0, 1\}^M$ and hidden layer variable is $\boldsymbol{h} = \{0, 1\}^N$, the energy possessed by an RBM system can be defined as:

$$E(\boldsymbol{v}, \boldsymbol{h}; \theta) = -\sum_{i=1}^{M}\sum_{j=1}^{N} W_{ij} v_i h_j - \sum_{i=1}^{M} b_i v_i - \sum_{j=1}^{N} a_j h_j \tag{11}$$

In the formula, parameter $\theta = \{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{W}\}$, W_{ij} is the symmetric weight between the visible layer node i and the hidden layer node j, b_i and a_j are the bias of the visible layer node i and the hidden layer node j.

With the energy equation, the joint probability distribution of the visible layer node and the hidden layer node can be obtained:

$$\begin{aligned} P(\boldsymbol{v}, \boldsymbol{h}; \theta) &= \frac{1}{Z(\theta)} e^{-E(\boldsymbol{v}, \boldsymbol{h}; \theta)} \\ Z(\theta) &= \sum_{\boldsymbol{v}, \boldsymbol{h}} e^{-E(\boldsymbol{v}, \boldsymbol{h}; \theta)} \end{aligned} \tag{12}$$

In practical applications, we are concerned with the observation data, that is, the edge probability distribution $P(\boldsymbol{v}\,|\theta)$ of the visible layer $\boldsymbol{v}$. This distribution is also called the Likelihood Function.

$$P(\boldsymbol{v}\,|\theta) = \frac{1}{Z(\theta)} \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v}, \boldsymbol{h}; \theta)} \tag{13}$$

In the formula, calculating $Z(\theta)$ needs to traverse the values of all visible layer nodes i and hidden layer nodes j, total 2^{m+n} times, which is intolerable in practical applications. However, because RBM has a special structure, it has edge connections between layers, but there is no connection between nodes in the layer. Therefore, when the state of a certain layer (such as the visible layer) is determined, the state of each node in the other layer (that is, the hidden layer) is conditionally independent. Therefore, the conditional probability of hidden layer node j activation is:

$$P\left(h_j = 1\,|\boldsymbol{v}; \theta\right) = \sigma\left(a_j + \sum_{i=1}^{M} v_i W_{ij}\right) \tag{14}$$

In the formula, $\sigma(x) = \frac{1}{1+\exp(-x)}$ is called sigmoid activation function.

Since the structure of RBM is symmetrical, the conditional probability of activation of node i in the visible layer can also be obtained by giving hidden layer $\boldsymbol{h}$:

$$P\left(v_i = 1\,|\boldsymbol{h}; \theta\right) = \sigma\left(b_i + \sum_{j=1}^{N} h_j W_{ij}\right) \tag{15}$$

The goal of training RBM is to get the value of parameter θ, and θ can be obtained by maximizing the log-likelihood of RBM on the training data:

$$\theta^* = \arg\max_{\theta} \Gamma(\theta) = \arg\max_{\theta} \sum_{t=1}^{T} \log P\left(\boldsymbol{v}^{(t)}\,|\theta\right) \tag{16}$$

In the formula, T represents the number of training samples. Based on partial derivation of θ with respect to parameter $\Gamma(\theta)$, the gradient can be obtained as:

$$\frac{\partial \Gamma}{\partial \theta} = \sum_{t=1}^{T} \left(\left\langle \frac{\partial\left(-E\left(\boldsymbol{v}^{(t)}, \boldsymbol{h}; \theta\right)\right)}{\partial \theta} \right\rangle_{P(\boldsymbol{h}|\boldsymbol{v}^{(t)}, \theta)} - \left\langle \frac{\partial\left(-E(\boldsymbol{v}, \boldsymbol{h}; \theta)\right)}{\partial \theta} \right\rangle_{P(\boldsymbol{v}|\boldsymbol{h}, \theta)} \right) \tag{17}$$

In the formula, $\langle\cdot\rangle P$ is the mathematical expectation about distribution P. $P\left(\boldsymbol{h}\,|\boldsymbol{v}^{(t)}, \theta\right)$ is the probability distribution of each node of the hidden layer under the given training samples (that is, the value of each node of the visible layer is determined), which is easy to calculate. However, $P(\boldsymbol{v}\,|\boldsymbol{h}\,, \theta)$ represents the joint probability distribution of the visible layer and the hidden layer, which involves the normalization factor $Z(\theta)$ and is difficult to obtain. Therefore, the approximate value can only be obtained by sampling. The commonly used sampling method is Gibbs sampling [37].

Hinton proposed an RBM fast learning method based on Contrastive Divergence (CD) algorithm in 2002. Hinton suggested that Gibbs sampling can be used in k steps (usually k = 1) to get a good enough approximation. First, set the visible layer as the value of a training sample, and use the probability obtained by Equation 14 to calculate the value of the hidden layer node. Then use the probability obtained by Equation 15 to determine the value of the visible layer node, so as to form a reconstruction of the training sample. In this way, the update rule based on the parameters of a sample becomes:

$$\begin{aligned} \Delta W_{ij} &= \gamma\left(\langle v_i h_i\rangle_{dt} - \langle v_i h_i\rangle_r\right) \\ \Delta b_{ij} &= \gamma\left(\langle v_i\rangle_{dt} - \langle v_i\rangle_r\right) \\ \Delta a_{ij} &= \gamma\left(\langle h_j\rangle_{dt} - \langle h_j\rangle_r\right) \end{aligned} \tag{18}$$

In the formula, γ is learning rate, $\langle\cdot\rangle r$ represents the distribution of the model after one-step reconstruction. Then, traverse all sample points and constantly update the model parameters to train the RBM.

The nodes used in RBM obey Bernoulli Distribution, that is, the value of the node is 0 or 1. Moreover, the contrast divergence algorithm (CD) is applied to train the RBM. In addition, when updating the weights of RBM, only learning one sample at a time will greatly increase the amount of calculation [38]. If dividing the samples into mini-batches with tens or hundreds of samples for training, the efficiency will be improved, which is mainly benefited from using the graphics processor GPU (Grapic Processing Unit) or the efficient matrix multiplication operation in MATLAB. In this paper, the batch size of small batch data is set to 100.

Let the number of visible layer units is n, the hidden layer units number is m, w represents the connection weight between the visible layer and the hidden layer ($m \times n$ dimension), vector α (m-dimensional column vector)represents the offset vector of the hidden layer, vector b (n-dimensional column vector) represents the offset vector of the visible layer. The CD-based RBM fast learning method is shown in FIGURE 5.

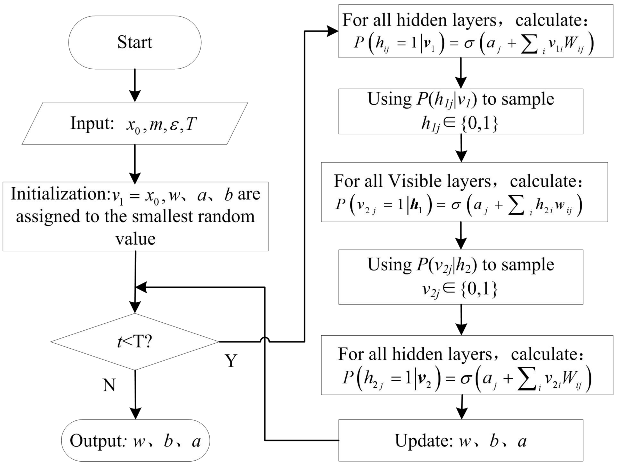

Start
Input: x_0, m, ε, T
Initialization:$v_1 = x_0$, w、a、b are assigned to the smallest random value
t<T?
Y
N
Output: w、b、a
For all hidden layers，calculate: $P(h_{ij}=1|\boldsymbol{v}_1)=\sigma(a_j+\sum_i v_{1i}W_{ij})$
Using $P(h_{1j}|v_1)$ to sample $h_{1j}\in\{0,1\}$
For all Visible layers，calculate: $P(v_{2j}=1|\boldsymbol{h}_1)=\sigma(a_j+\sum_i h_{2i}w_{ij})$
Using $P(v_{2j}|h_2)$ to sample $v_{2j}\in\{0,1\}$
For all hidden layers，calculate: $P(h_{2j}=1|\boldsymbol{v}_2)=\sigma(a_j+\sum_i v_{2i}W_{ij})$
Update: w、b、a

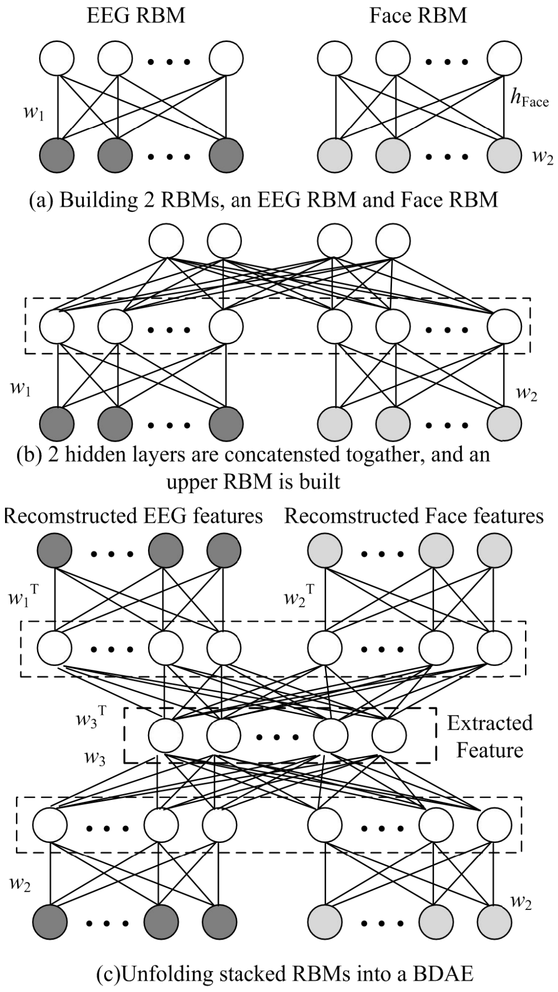
FIGURE 5. RBM fast learning method based on CD.

B. CONSTRUCTION OF MULTIMODAL EMOTION RECOGNITION MODEL

A BDAE is used to construct a deep neural network, which includes two parts: encoding and decoding. In the encoding part, two RBM models are trained using the characteristics of EEG signals and eye movement signals, as shown in FIGURE 6(a).

The hidden layer h_{EEG}, h_{Face} and weights $w_1 w_2$ of the two RBMs can be obtained after training. h_{EEG} and h_{Face} are merged together into another new RBM visible layer, and then train the RBM to obtain the corresponding weights, which is shown in FIGURE 6(b).

In the decoding part, two layers of RBM are developed to reconstruct the input features, so as to form a deep automatic encoder. The weights between the network connection layers are w_1, w_2, w_3, w_3^T, w_1^T, w_2^T, shown in FIGURE 6(c).



(a) Building 2 RBMs, an EEG RBM and Face RBM

(b) 2 hidden layers are concatensted togather, and an upper RBM is built

(c)Unfolding stacked RBMs into a BDAE

FIGURE 6. Schematic diagram of dual-mode depth automatic encoder.

Then, the back-propagation algorithm (BP) method of unsupervised learning is used to adjust the parameters of the network. After training BDAE, let the middle layer (the third layer of BDAE, a total of five layers) be the extracted features. Then send them to the LIBSVM classifier for supervised learning training, and get the final emotion classification model.

The proposed method uses LIBSVM as a classifier to identify emotions. LIBSVM is the improvement and supplement to some parameters of the original SVM. LIBSVM is usually used to solve binary classification problems. Based on the established hyperplane, it can distinguish positive examples and negative examples as much as possible.

In the process of emotion classification, each channel involved in emotion recognition will eventually get a result of emotion recognition [39]. At this time, each channel can be regarded as a set of separate EEG signals, which can form a separate LIBSVM classifier. Perform decision layer fusion for the classification results generated by each LIBSVM classifier. The apply of multi-classifier fusion based on fuzzy integration can not only fuse the results obtained by each LIBSVM classifier, but also reflect the importance of each LIBSVM classifier in the fusion process [40]. Fuzzy integral uses the fuzzy measure as the weight of each LIBSVM classifier, and fully considers the relationship between each LIBSVM.

V. EXPERIMENTAL VERIFICATION AND RESULT ANALYSIS

In order to demonstrate the performance of the proposed multi-modal deep learning emotion recognition method, a video library is first established for video emotion-evoked EEG experiments. The video library contains 90 video clips. These video clips are collected from different movies and TV shows, and are unified into wmv format. The 90 video clips in the video library contain three types of emotions: violent, neutral and pornographic. The violent video clips and pornographic video clips were taken from two different types of famous movies: action films and drama films. In the video library contained 90 video clips, violent, neutral and pornographic videos each contain 30 video clips. Each video clip contains only one emotion type, and the duration of each video clip is about 6s. Each video clip in the video library is objectively evaluated by 6 researchers (3 men and 3 women) before being included in the video library. When selecting video clips, only these six researchers believe that a certain video clip belongs to a certain emotional type clip, then this video clip can be selected into the video library.

13 healthy subjects participated in the video emotion-evoked electroencephalogram experiment, including 7 males and 6 females. The subjects were 24-28 years old, and the naked vision or corrected vision reached 1.0. In order to induce EEG signals containing different video emotions, the above video library contained 90 video clips was used as a stimulus. The subjects wore a 64-lead Quik-Cap electrode cap and watched the video clips continuously playing on the computer to induce EEG signals containing different video emotions.

The electrodes of the electrode cap are arranged according to the 10-20 system electrode coordination method, as shown in Fig. 7. Neuroscan system is used to collect and preprocess the EEG signals induced by experiments. The E-Prime software developed by PST company is used in the design of video emotion evoked EEG experiment.

The flow chart of the designed experiment is shown in Figure 8. At the beginning of the experiment, the computer screen in front of the subjects will display the instructions and cautions. After understanding the experimental process and the general content of the experiment, the subjects started the experiment by pressing the space bar.

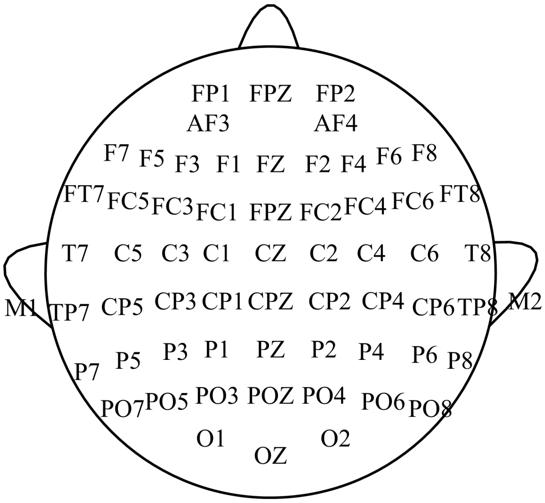


FIGURE 7. 64 guide electrode position distribution.

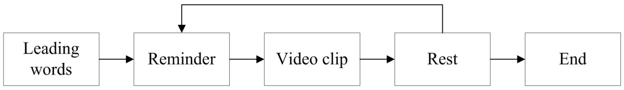


FIGURE 8. Flow chart of video emotion evoked EEG experiment.

For each subject, 30 video clips were randomly selected from the video library contained 90 videos, and each emotion type contained 10 video clips. In order to prevent subjects from forming inertial memories, the playback of these 30 video clips is random. Before playing each video clip, a cross-shaped prompt will be displayed on the computer screen to attract the attention of the subjects. After each video clip is played, there will be a period of rest to allow the subjects to calm down. After the 30 selected video clips are all played, the experiment ends. Throughout the experiment, the subjects' EEG signals were collected, and the sampling rate of the EEG signals was 1000 Hz. The above experiment was repeated for each subject, and could obtain final 13 subjects' EEG signals.

After preprocessing the collected original EEG data, a relatively "pure" EEG signal is obtained. Then, we extract the features of "pure" EEG signals to obtain the initial EEG features. In this method, wavelet packet decomposition (WPD) is used to extract the features of the preprocessed EEG signals. In the stage of EEG feature extraction, DB6 is used as wavelet base. Through the experiment, the wavelet packet decomposition level within 10 does not have a great impact on the classification results, so this paper sets the decomposition level to 3. The WPD features with dimension 8 are extracted from the EEG signals of each window length. For a video clip, 48 dimensional WPD features are extracted from a subject's EEG signal collected by an electrode. When 64 electrodes are used in the experiment, 3072 dimensional WPD features are extracted from a subject's EEG signal for a video clip. The feature selection method based on decision tree is used to select the initial EEG features. In this paper, each dimension of EEG feature is regarded as an attribute. WPD feature vector after feature extraction is input into decision tree C4.5 for tree building and pruning, and a simplified tree is formed. All the attributes included in the simplified tree are the selected EEG features. Experiments show that the simplified tree contains 14 attributes. Therefore, 14 dimensional EEG features are selected from 3072 dimensional WPD features for later classification.

In addition, the nodes number in all RBM hidden layers of the deep neural network is same, and cross-validation is used in {700, 500, 450, 400, 350, 300, 250, 200, 150, 100, 50, 20, 10} to select the optimal node number parameter. Moreover, for BDAE, set the number of epochs for adjusting the network weights of BPAE, and the optimal parameters are cross-validated in selection of {1000, 700, 500, 300, 250, 200, 150, 100, 50, 10}.

A. EXTRACTION COMPLEMENTARY ANALYSIS OF EEG SIGNAL AND EXPRESSION SIGNAL

In order to study the difference in the ability of EEG signals and facial expression signals to recognize different emotional states, EEG signals, expression features, feature layer fusion (FLF) method and BDAE method were used to identify the four emotion-like confusion matrices, the results are shown in Table 1. Each row of the confusion matrix represents the true category of the data, and each column represents the predicted category. The i-th row and the j-th column indicate the number of samples that the model distinguishes the data that truly belongs to the i-th category as the j-th category.

TABLE 1. (a) EEG. (b) Face. (c) FLF. (d) BDAE.

(A)

	Sad	Fear	Happy	Neutral
Sad	2578	228	311	935
Fear	136	1721	298	498
Happy	58	135	1688	235
Neutral	123	278	372	2678

(B)

	Sad	Fear	Happy	Neutral
Sad	2358	942	471	278
Fear	347	1776	368	163
Happy	314	116	1423	265
Neutral	188	234	265	2766

(C)

	Sad	Fear	Happy	Neutral
Sad	2771	641	338	300
Fear	366	2100	299	498
Happy	59	138	1866	233
Neutral	122	278	228	3101

(D)

	Sad	Fear	Happy	Neutral
Sad	3218	248	192	175
Fear	149	2566	113	100
Happy	161	99	1562	299
Neutral	49	36	198	3180

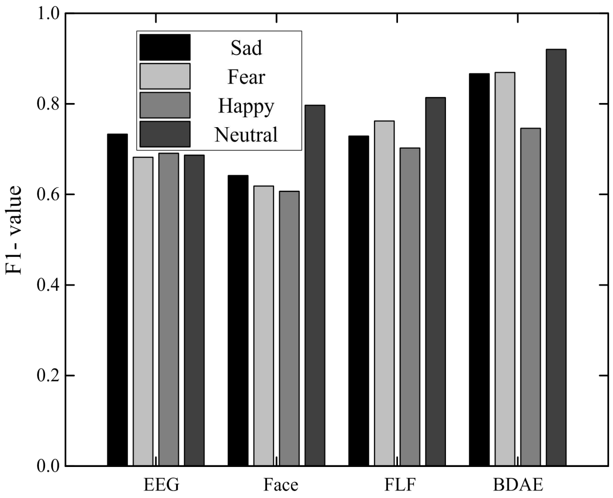


FIGURE 9. F1-score of different modalities for different classes.

Furthermore, the confusion matrix can be used to calculate Precision, Recall, and F1-score. Precision represents the percentage of data identified as a certain category that really belongs to this category. Recall indicates the percentage of data that the model judges correctly in the data that really belongs to a certain category. The F1-score combines the two measure characterizes, which is a weighted average with value between 0 ∼ 1. It is also a method of measuring accuracy. FIGURE9 can be obtained by calculating the accuracy, recall, and F1-values of the four confusion matrices, which focuses on comparing the F1-values of EEG signals, eye movement signals, FLF, and BDAE methods on different emotion types.

Observing the F1-values of EEG signals in four types of emotional states from the above figure, it can be seen that the F1-values of sad and happy states are higher, indicating that the model trained with EEG signals is better at distinguishing sad and happy state. From the F1-values of facial expression characteristics under four types of emotional states, it can be seen that the F1-values under neutral emotional states are significantly higher than those under the other three emotional states, indicating that models trained with facial expression characteristics are better at distinguishing neutral emotions. On the basis of above results, models trained with EEG signals and facial expression characteristics have great differences in the ability to distinguish four different emotions, indicating that the two models have a certain complementarity in the ability to characterize different emotions, which can effectively improve the accuracy of emotion recognition.

The F1-value of the FLF method in four types of emotional states shows that the F1-value in the fear and neutral states is relatively high, indicating that the FLF method is better at distinguishing fear and neutral emotions. In addition, except that the F1-value in the sad state (0.73) is slightly lower than the F1-value in the sad state of the EEG signal (0.74), the F1-values of the other three emotions are higher than those corresponding to the EEG signals and expression characteristics F1-value. The F1-values obtained by the proposed method using BDAE in four types of emotional states have been significantly improved, and the F1-values of sadness, fear and neutral are relatively high, indicating that the BDAE method distinguish ability between sadness, fear and neutral has been significantly improved. This also shows that it uses the advantages of EEG signals to distinguish sad emotions and facial expressions, to distinguish between neutral emotions and to improve their recognize emotions ability. Combining with the confusion matrix table 1(d), for the sad emotions, the model originally trained by the EEG signal is most likely to misclassify the data that is truly sad into neutral, and there are fewer cases of misclassification into fear, but the expression characteristics are just the opposite. BDAE, combined the characteristics of the two models, has effectively improved the situation, in which the data that really belongs to sadness is wrongly divided into neutral and fear. Similarly, the other three types of emotions have similar results.

The comparison results of F1-values of EEG signals, facial expression characteristics, FLF and BDAE methods for each emotional state are shown in FIGURE 10.

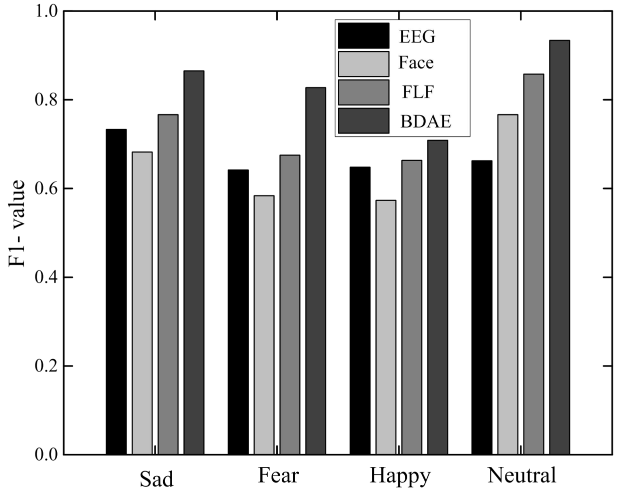


FIGURE 10. F1-score of each class from different modalities.

For sad emotions in the above figure, the F1-value of the FLF method has not been significantly improved, and the ability of the proposed method to recognize sad emotions using BDAE has been significantly improved. For fear and happiness, two F1-values of the single modal are relatively close, but the FLF and BDAE methods improve the ability to recognize fear emotions to a greater extent than that to recognize happy. For neutral emotions, the model trained by EEG signals is not very good at distinguishing this emotion, but the model trained by facial expression characteristics has better performance. Finally, by combining the complementary information between two methods, the obtained model can effectively improve the ability to recognize neutral emotions.

The results show that the EEG signals and facial expression characteristics have differences in the ability to characterize different emotions, and combining them can effectively use the complementarity information to improve the ability of recognize four types of emotions.

B. THE INFLUENCE OF LEARNING RATE ON THE RECOGNITION ACCURACY RATE OF FOUR KINDS OF EMOTION

In order to analyze how the learning rate γ affects the accuracy of the proposed method of emotion recognition, classification accuracy results of the four emotions, including happiness, sadness, fear and neutrality, are shown in FIGURE 11.

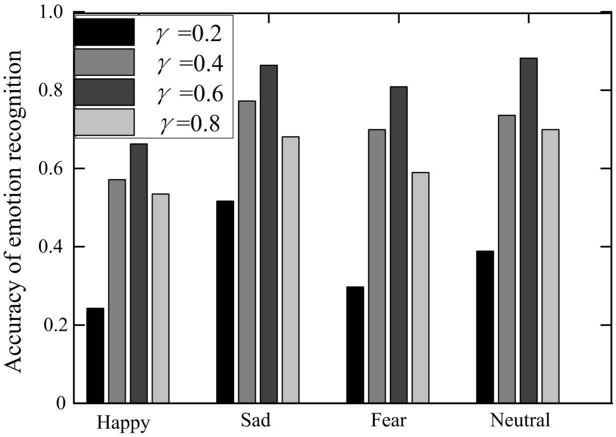


FIGURE 11. The effect of learning rate on the accuracy of four kinds of emotion recognition.

From the FIGURE 11, with the increase of learning rate γ, the accuracy rate of emotion recognition has fluctuated to a certain extent. When the learning rate is too small, there will be a very slow convergence rate for parameters with large gradients; while the learning rate is too large, the parameters that have been optimized similarly may be unstable. Therefore, when γ is 0.4 and 0.6, the accuracy is higher. In addition, the proposed method combines the EEG signals and facial video features, and uses complementary advantages of the EEG signals good at distinguishing sad emotions and expression features good at distinguishing neutral emotions to improve the model ability of recognizing emotions. Therefore, the model gets the highest recognition accuracy of neutral emotion, while the lower recognition accuracy rate of happy emotions.

C. MODEL STABILITY OVER TIME

In practical applications, it is not realistic to train a model each time and then test in order to recognize the user's emotions. It is necessary to train a more stable emotion recognition model that can be used for the task of emotion recognition in a long time. In order to study the stability of the emotion recognition model over time, each participant was asked to do three experiments, each time a few days apart. By using the same subject's data from different experiments as the training set and test set to verify that the change of the proposed model over time is relatively stable. Table 2 shows the stability of the model obtained by the fusion of 6-lead and 62-lead EEG signals and facial expression features after training by the BDAE method. The i-th row and j-th column of each table represent the i-th day use the data as a training set to train the model, and then use the data from the j-th day as the test set to test the recognition accuracy.

From the above table, the model obtained by combining 6-lead EEG signals or 62-lead EEG signals and expression features for training has certain stability over time, and the accuracy rate on the diagonal is higher than other ones.

TABLE 2. Average accuracy of different groups in experimental training tests using the BDAE method.

Derivative	Train	Test		
		1st	2nd	3rd
6	1st	83.62±13.5%	70.52±6.21%	66.89±9.25%
	2nd	63.52±9.56%	86.57±13.61%	62.33±8.69%
	3rd	63.21±8.58%	70.63±9.12%	83.56±8.74%
62	1st	82.65±9.68%	70.65±8.99%	68.25±7.56%
	2nd	67.56±7.29%	88.87±8.49%	64.59±10.43%
	3rd	66.34±7.85%	70.35±6.89%	83.62±8.39%

The model trained on the same day can achieve a higher accuracy of predicting results in the emotional state experiment. When used to test the emotional state of other different groups in experiments, although the accuracy is not as high as that of the day, it still achieves satisfactory accuracy compared with four classifications. The results of training and testing in the same group of experiments are better than those between different groups of experiments. The environmental variables of each experiment will slightly change, for example, the impedance of the EEG cap is different, and the interference from the outside will also change. When training or test data comes from different experiments, the noise will affect the distribution of samples, which will reduce the accuracy. In addition, the emotional induction of subjects under different experiments will also be different, which will also lead to relatively low accuracy of training or tests conducted by different groups of experiments.

In addition, with the passage of time, the longer the interval of experimental time, the lower the accuracy of emotion recognition obtained by training and testing, indicating that the stability of the model will decrease as time goes by. And the difference between the recognition accuracy of 6 leads and 62 leads is not very obvious. The model trained by the combination of 6 EEG signals and facial expression features also has stability equivalent to lead 62 over time, which also proves that 6 EEG signal has a strong emotion representation capability, and can only use small number of electrode EEG signals for emotion recognition. In this way, reducing complexity and cost, and enhancing portability. Therefore, it has been demonstrated that the model obtained by combining EEG signals and facial expression features with time has certain stability, which provides the possibility of using EEG signals and facial expression features for emotion recognition in practice.

D. COMPARATIVE RESULTS OF MULTI-MODAL FUSION EMOTION MODEL EXPERIMENTS

The proposed method is compared with the methods in [22], [26], [28], [30], and [31]. Experiment results are shown in Table 3, where the evaluation standard is the average of emotion recognition rate of discrete emotional states. In order to correctly evaluate the comparison methods, the same data sets were used in all the comparison methods in this study.

TABLE 3. Comparison results of discrete emotion models with multimodal fusion.

	Types of emotion recognition	Average emotion recognition rate	Average running time (s)
Ref.[22]	Positive, negative	62.5%	16.33
Ref. [26]	Happy, sad, angry, relax	74.52%%	19.21
Ref.[28]	Negative, positive, neutral	77.63%	10.29
Ref. [30]	Sad, positive, neutral	79.32%	14.21
Ref.[31]	Happy, positive, neutral	82.36%	9.82
The proposed	Happy, sad, angry, relaxed	85.71%	15.57

From the above table, [22] used brain functional networks and achieved 62.5% average emotion recognition rate on positive and negative emotions. In [26], EEG was used to identify four emotional states, including happiness, sadness, anger, and relaxation, then utilizing wavelet analysis to identify them. The final average emotion recognition rate was 74.52%. [28] recognized four emotion states based on facial expression features, and used SVM for classification. The average emotion recognition rate is 77.63%. [30] proposed a new group sparse canonical correlation analysis (GSCCA) method for synchronizing EEG channel selection and emotion recognition, which divided the entire EEG frequency band into five parts, and then extracted the frequency band from each frequency band of GSCCA characteristics. When distinguishing 3 discrete emotion states, the model achieved an 79.32% average emotion recognition rate. [31] performed multi-modal fusion for EEG and electrooculogram signals. On the feature layer fusion, the feature vectors of the two physiological signals are directly stitched together to form a longer feature vector. On the decision layer fusion, two classifiers are trained with different feature values respectively, and performed the decision layer fusion. The proposed method uses BDAE to fuse EEG signals with facial video features and conducts emotion recognition based on deep learning. Therefore, regardless of the type of discrete emotional state recognition or the average emotion recognition rate, there is a relative increase in average emotion, reaching 85.71%. In terms of the running time of each model, the proposed model is in the middle position, and the average running time is 15.57s.

VI. CONCLUSION

As an important field of artificial intelligence, emotion computing is favored by more and more researchers. At present, most of emotion recognition methodsare based on speech signals, facial expressions, and other bioelectric signals such as electrocardiogram and electroencephalogram. When the emotion signal of a single channel is interfered by other signals, the models usually get lower emotion recognition rate. Therefore, taking expression signals and EEG signals as emotion signals, a multi-modal emotion recognition method combining expression signals and EEG signals is proposed through BDAE feature fusion. Using the constructed video library to carry out supervised learning training experiments on the proposed method, experiment results show that the complementation of EEG signals and facial expression features can effectively improve the ability to recognize four types of emotions. And the application of deep neural networks can greatly improve the ability of multi-modal emotion recognition. Compared with other methods, the recognition accuracy of the proposed method reaches 85.71%.

Currently, most research works take two different emotion signals as target objects, but when humans deliberately disguise emotion signals, the recognition rate of obtained emotions tends to decrease. Therefore, combining more emotion information will improve the construction of emotion recognition system, and structuring a more effective emotion database will be the focus of the next research. In addition, different emotions usually have certain correlation, for example, sad emotions often contain a certain amount of anger. Therefore, it is worth researching that how to construct a more effective emotion recognition model by combining the correlation between emotions. However, the method of classifying video clips proposed in this paper is lack of subjectivity and needs to be further improved. In addition, there is no public database including video and corresponding evoked EEG in the world. In the future, we can consider increasing the type of video, the number of video and the length of video, and collecting EEG signals from more subjects in order to establish a public video EEG database.

REFERENCES

[1] M. S. Hossain and G. Muhammad, "An audio-visual emotion recognition system using deep learning fusion for a cognitive wireless framework," *IEEE Wireless Commun.*, vol. 26, no. 3, pp. 62–68, Jun. 2019.

[2] S. Gica, B. C. Poyraz, and H. Gulec, "Are emotion recognition deficits in patients with schizophrenia states or traits? A 6-month follow-up study," *Indian J. Psychiatry*, vol. 61, no. 1, pp. 45–52, 2019.

[3] C. Ferrari, C. Papagno, A. Todorov, and Z. Cattaneo, "Differences in emotion recognition from body and face cues between deaf and hearing individuals," *Multisensory Res.*, vol. 32, no. 6, pp. 1–21, 2019.

[4] S. Zhao, A. Gholaminejad, G. Ding, Y. Gao, J. Han, and K. Keutzer, "Personalized emotion recognition by personality-aware high-order learning of physiological signals," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 1s, pp. 1–18, Feb. 2019.

[5] H. Pa Pa Win, M. University of Computer StudiesHpa-an, and P. Thu Thu Khine, "Emotion recognition system of noisy speech in real world environment," *Int. J. Image, Graph. Signal Process.*, vol. 12, no. 2, pp. 1–8, Apr. 2020.

[6] L. Bhole and M. Ingle, "Estimating range and relationship of EEG frequency bands for emotion recognition," *Int. J. Comput. Appl.*, vol. 178, no. 13, pp. 16–21, May 2019.

[7] J. R. Zhuang, Y. J. Guan, H. Nagayoshi, K. Muramatsu, K. Watanuki, and E. Tanaka, "Real-time emotion recognition system with multiple physiological signals," *J. Adv. Mech. Des. Syst. Manuf.*, vol. 13, no. 4, pp. 75–85, 2019.

[8] K. Nalbant, B. M. Kalayci, D. Akdemir, S. Akgül, and N. Kanbur, "Emotion regulation, emotion recognition, and empathy in adolescents with anorexia nervosa," *Eating Weight Disorders*, vol. 24, no. 4, pp. 1–10, 2019.

[9] Y. F. Zhou and N. Chen, "The LAP under facility disruptions during early post-earthquake rescue using PSO-GA hybrid algorithm," *Fresenius Environ. Bull.*, vol. 28, no. 12A, pp. 9906–9914, 2019.

[10] L. Singh, S. Singh, and N. Aggarwal, "Improved TOPSIS method for peak frame selection in audio-video human emotion recognition," *Multimedia Tools Appl.*, vol. 78, no. 5, pp. 6277–6308, Mar. 2019.

[11] C. Westby, "Emotion recognition by children with hearing loss," *Word Mouth*, vol. 31, no. 1, pp. 5–7, Sep. 2019.

[12] S. A. M. Al-Sumaidaee, M. A. M. Abdullah, R. R. O. Al-Nima, S. S. Dlay, and J. A. Chambers, "Multi-gradient features and elongated quinary pattern encoding for image-based facial expression recognition," *Pattern Recognit.*, vol. 71, pp. 249–263, Nov. 2017.

[13] A. S. Dandoti and S. M. Sangve, "Emotion identification between POMS and multinomial naive Bayes algorithm using Twitter API," *Int. J. Comput. Sci. Eng.*, vol. 7, no. 7, pp. 14–19, Jul. 2019.

[14] Y. Huang, K. Tian, A. Wu, and G. Zhang, "Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition," *J. Ambient Intell. Humanized Comput.*, vol. 10, no. 5, pp. 1787–1798, May 2019.

[15] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial–Temporal recurrent neural network for emotion recognition," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 839–847, Mar. 2019.

[16] A. Sun and Y. M. Huang, "A traffic balance scheme of group emotion recognition by using the service function chain," *Int. J. Commun. Syst.*, vol. 32, no. 14, pp. 1–17, 2019.

[17] W. Xiaohua, P. Muzi, P. Lijuan, H. Min, J. Chunhua, and R. Fuji, "Two-level attention with two-stage multi-task learning for facial emotion recognition," *J. Vis. Commun. Image Represent.*, vol. 62, pp. 217–225, Jul. 2019.

[18] K. Xia, T. Hu, and W. Si, "Editorial for the special issue on 'Research on methods of multimodal information fusion in emotion recognition,'" *Pers. Ubiquitous Comput.*, vol. 23, nos. 3–4, pp. 359–361, Jul. 2019.

[19] S. Jirayucharoensak, S. Pan-Ngum, and P. Israsena, "EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation," *Sci. World J.*, vol. 2014, pp. 1–10, Sep. 2014, doi: 10.1155/2014/627892.

[20] D.-D. Wu, S.-H. Li, J. He, W. Su, and H.-B. Chen, "Emotion recognition in patients with parkinson disease," *Cognit. Behav. Neurol.*, vol. 32, no. 4, pp. 247–255, Dec. 2019.

[21] H. Yang, J. Han, and K. Min, "A multi-column CNN model for emotion recognition from EEG signals," *Sensors*, vol. 19, no. 21, p. 4736, Oct. 2019, doi: 10.3390/s19214736.

[22] W.-C. Fang, K.-Y. Wang, N. Fahier, Y.-L. Ho, and Y.-D. Huang, "Development and validation of an EEG-based real-time emotion recognition system using edge AI computing platform with convolutional neural network System-on-Chip design," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 4, pp. 645–657, Dec. 2019.

[23] J. Arunnehru, G. Chamundeeswari, and S. P. Bharathi, "Human action recognition using 3D convolutional neural networks with 3D motion cuboids in surveillance videos," *Procedia Comput. Sci.*, vol. 133, pp. 471–477, 2018.

[24] R. O. Bustillos, R. Z. Cabada, M. L. B. Estrada, and Y. H. Pérez, "Opinion mining and emotion recognition in an intelligent learning environment," *Comput. Appl. Eng. Edu.*, vol. 27, no. 1, pp. 90–101, Jan. 2019.

[25] M. Ren, W. Nie, A. Liu, and Y. Su, "Multi-modal correlated network for emotion recognition in speech," *Vis. Informat.*, vol. 3, no. 3, pp. 150–155, Sep. 2019.

[26] A. T. Wieckowski and S. W. White, "Attention modification to attenuate facial emotion recognition deficits in children with autism: A pilot study," *J. Autism Develop. Disorders*, vol. 50, no. 1, pp. 30–41, Jan. 2020.

[27] L. A. B. Prieto and Z. K. Oplatkova, "Emotion recognition using autoencoders and convolutional neural networks," *Mendel*, vol. 24, no. 1, pp. 113–120, Jun. 2018.

[28] L. A. Rutter, D. J. Norton, and T. A. Brown, "The impact of self-reported depression severity and age on facial emotion recognition in outpatients with anxiety and mood disorders," *J. Psychopathology Behav. Assessment*, vol. 42, no. 1, pp. 86–92, 2020.

[29] S. Angadi and V. S. Reddy, ''Hybrid deep network scheme for emotion recognition in speech,'' *Int. J. Intell. Eng. Syst.*, vol. 12, no. 3, pp. 59–67, Jun. 2019.

[30] V. Doma and M. Pirouz, ''A comparative analysis of machine learning methods for emotion recognition using EEG and peripheral physiological signals,'' *J. Big Data*, vol. 7, no. 1, pp. 1–21, Dec. 2020.

[31] P. Wei and Y. Zhao, ''A novel speech emotion recognition algorithm based on wavelet kernel sparse classifier in stacked deep autoencoder model,'' *Pers. Ubiquitous Comput.*, vol. 23, nos. 3–4, p. 59, 2019.

[32] A. T. Wieckowski, D. M. Swain, A. L. Abbott, and S. W. White, ''Task dependency when evaluating association between facial emotion recognition and facial emotion expression in children with ASD,'' *J. Autism Developmental Disorders*, vol. 49, no. 6, pp. 1–8, 2019.

[33] S. Wang, H. Chi, Z. Yuan, and J. Geng, ''Emotion recognition using cloud model,'' *Chin. J. Electron.*, vol. 28, no. 3, pp. 30–34, 2019.

[34] M. Sharma, A. S. Jalal, and A. Khan, ''Emotion recognition using facial expression by fusing key points descriptor and texture features,'' *Multimedia Tools Appl.*, vol. 78, no. 12, pp. 16195–16219, Jun. 2019.

[35] M. Ghosh, T. Kundu, D. Ghosh, and R. Sarkar, ''Feature selection for facial emotion recognition using late hill-climbing based memetic algorithm,'' *Multimedia Tools Appl.*, vol. 78, no. 18, pp. 25753–25779, Sep. 2019.

[36] S. B. Wankhade and D. D. Doye, ''Deep learning of empirical mean curve decomposition-wavelet decomposed EEG signal for emotion recognition,'' *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 28, no. 01, pp. 153–177, Feb. 2020.

[37] Y. Zhou, H. Yu, Z. Li, J. Su, and C. Liu, ''Robust optimization of a distribution network location-routing problem under carbon trading policies,'' *IEEE Access*, vol. 8, pp. 46288–46306, 2020.

[38] 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.

[39] Y. Ye, X. Zhang, Y. Lin, and H. Wang, ''Facial expression recognition via region-based convolutional fusion network,'' *J. Vis. Commun. Image Represent.*, vol. 62, pp. 1–11, Jul. 2019.

[40] J. Yan, H. Kuai, J. Chen, and N. Zhong, ''Analyzing emotional oscillatory brain network for valence and arousal-based emotion recognition using EEG data,'' *Int. J. Inf. Technol. Decis. Making*, vol. 18, no. 04, pp. 1359–1378, Jul. 2019.