The research of emotion recognition based on multi-source physiological signals with data fusion

Yuanteng Han*, and *Yong* Xu

South China University of Technology, Electronic Commerce Department, 51006 Guangzhou, China

Abstract. In the context of the increasing application value of emotion recognition and the continuous development of data fusion technology, it is of great significance to study the emotion recognition model based on multi-source physiological signals with data fusion. In this paper, the one-dimensional-convolutional neural network-support vector machine (1D-CNN-SVM) emotion recognition model is constructed to extract the emotional features of multi-source physiological signal data, realize data fusion and complete emotion recognition. Firstly, based on the data level fusion method, dimension splicing for data of each channel is used to compare and analyze different data splicing combinations to explore the best one. Secondly, based on the feature level fusion method, the depth features of each part are fused and extracted by convolutional neural network models. Finally, feature stitching and support vector machine algorithm are used to classify and recognize emotion categories. The experimental results verify the effectiveness of the proposed model in the valence-arousal of the four-class task on DEAP dataset, and the recognition accuracy of the optimal combination can reach 93.10%.

1 Background & summary

Emotion is a precious psychological activity and subjective experience in people's daily life, which seriously affects personal life state and behavior[1]. Although emotion is internal and hidden, various studies show that the change of emotion must be accompanied by physiological changes and behavioral changes[2][3][4][5]. Therefore, personal emotion recognition can be realized accurately and quickly by measuring the observation data related to human changes through sensors and constructing a scientific and reasonable machine learning model. Emotion recognition is an interdisciplinary computer science research, which involves a wide range of research fields, including psychology, physiology, neuroscience, computer science and so on. The key of the research is to select the neurophysiological data obtained by human sensors, such as electroencephalography(EEG) in brain center data[6] and electrocardiogram(ECG), electro-oculogram(EOG), electromyogram(EMG), galvanic skin response(GSR), photoplethysmogram(PPG), respiration(RESP), temperature(TEMP) in peripheral physiological data[7], or behavior performance data, such as facial expression, voice intonation, limb behavior[8], and design appropriate data fusion methods and

* Corresponding author: 201920149089@mail.scut.edu.cn

classification algorithms or prediction algorithms to train and finally build the scientific and accurate emotion recognition model.

Since 1973, data fusion technology has been growing vigorously and the concept of data fusion is becoming more and more comprehensive[9]. Data fusion methods can be divided into three levels: data level fusion, feature level fusion and decision level fusion. Different levels of data fusion methods have different characteristics and are suitable for different application scenarios[10]. Data level data fusion which is a fusion method directly facing the original observation data uses linear, nonlinear estimation and statistical operation methods to fuse and calculate the multi-sensor data. Feature level data fusion is to extract the features of multi-source data, splice or cascade them into high-dimensional feature vectors, and then put them into the classifier for model training. Decision level data fusion coordinates and makes joint decisions on the results of multiple classifiers through specific algorithms or methods to obtain the consistent fusion results[11].

In the field of emotion recognition, data level fusion is rarely applied, while feature level and decision level have been put into practice. Pan et al.(2020) carried out feature engineering processing and feature level data fusion on four kinds of peripheral physiological data, and then trained the emotion recognition model by using support vector machines(SVM) algorithm to realize the recognition of four emotions[12]. Huang et al.(2019) selected EEG data and expression data in the study, applied SVM and convolutional neural network(CNN) to classify and identify emotions respectively, and finally compared the application of enumeration weight method and adaptive enhancement method for decision-making level data fusion to build the final emotion recognition model, which is applied to the identification of valence-arousal emotion two-dimensional mode[13]. Foreign scholars Maaoui et al.(2014) tried to use the data fusion of peripheral physiological data and facial expression data to build an emotion recognition model, and compared the method of feature level fusion and decision level fusion. The experimental results verify the better recognition performance of feature level data fusion method[14].

According to the above literature review, for the purpose of studying and realizing the classification and recognition of emotions, this paper takes the combination of EEG data and peripheral physiological data after data preprocessing as the input. Firstly, the dimensional splicing of multi-source physiological signal data is used for data level data fusion. Secondly, the neural network model of physiological signal feature fusion and extraction is constructed and trained for feature level data fusion. Finally, the feature level data fusion is realized again through feature stitching, and the extracted feature set is input into the SVM classifier to output the category of emotion.

2 1D-CNN-SVM emotion recognition model

2.1 Research Model

The overall research model of emotion recognition in this paper is shown in Fig.1 below.

2.2 Dimension stitching of physiological signal data based on data level fusion

The idea and method of data level fusion have enlightening significance and reference value for the fusion of emotional features in the data layer. The specific operation method of data level fusion in this study is multi-source physiological signal data dimension splicing. In the process of splicing and fusion, different splicing combinations for EEG data and physiological signal data include EOG, EMG, GSR, RESP, PPG and TEMP are studied and analyzed. The acquisition of EEG must be realized by multi-sensor, multi-channel and data fusion, and the data of each channel of EEG is closely related and has common features. EOG and EMG are electrical records of human parts, which may jointly reflect personal emotional changes. GSR, RESP, PPG and TEMP also carry a lot of emotional information, and there are potential close connections through the interaction of nervous system, blood and muscle of the human body. In this study, the data dimension is spliced through data level fusion, and the physiological signal data combination that is scientific and reasonable and most conducive to feature fusion and extraction is selected, so as to give full play to the utility of data level data fusion method in the construction of emotion recognition model.

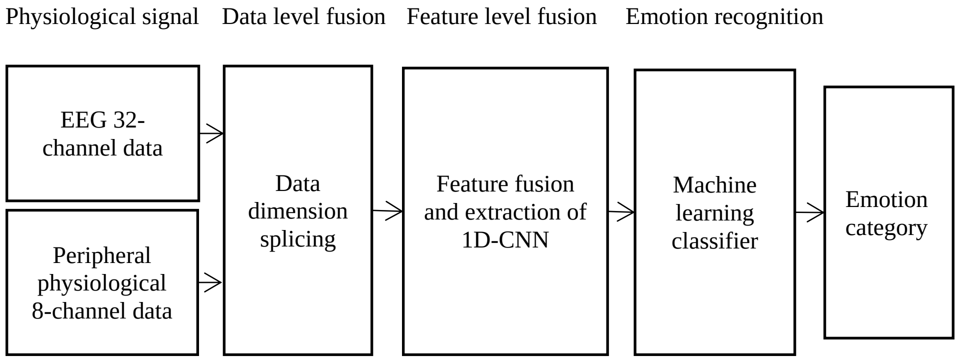


Fig. 1. Emotion recognition model based on multi-source physiological signal with data fusion.

2.3 1D-CNN physiological signal data feature fusion and extraction model based on feature level data fusion

Different from the traditional feature engineering methods, this study applies the deap learning method to fuse and extract the depth features of multi-source physiological signal data on the basis of data dimension stitching. According to the characteristics of physiological signal data, one-dimensional convolutional neural network (1D-CNN) is selected as the basis of model construction. Convolutional neural network is a deep feedforward neural network. Compared with ordinary fully connected neural network, it has fewer parameters, higher training efficiency and lower over fitting degree. It is more suitable for feature fusion and extraction of image data or signal data. Based on the feature-level data fusion method, multi-source physiological signal data is input, and the deep feature fusion extraction model 1D-CNN is built.

Two important operations in 1D-CNN are convolution and pooling. The core layer of convolutional neural networks [15] is the convolution layer, which acts to retain and fuse the extracted features of the signal data. The key of one-dimensional convolution is the dot product operation between the local region of input data and convolution kernel (also known as filter). The convolution kernel with the number of n and scale of k is used to perform matrix dot product operation on the input data of each channel from left to right according to the step s, so as to obtain the calculated value of the corresponding position and output it. Suppose the input data is $x_1, x_2, \cdots, x_{samples}$, convolution kernel parameter is $\omega_1, \omega_2 \cdots, \omega_k$. The mathematical expression of this process can be simply expressed as:

$$\mathrm{y_i} = \sum_{\mathrm{j}=1}^{\mathrm{k}} \upomega_{\mathrm{j}}\, \mathrm{x}_{\mathrm{j}+(\mathrm{i}-1)\mathrm{s}}, \mathrm{i} = 1,2,\cdots \tag{1}$$

The pooling layer is usually added periodically after a certain number of convolution layers. Its function is to reduce the dimension of the convoluted characteristic matrix, so as to reduce the number of model parameters, shorten the model training time, overcome the over fitting of the model and so on. The most common pool operations are maximum pool and average pool, that is, the maximum or average value in the pool window is retained. Similar to the convolution layer, the pool window size and pool step can be set in the network parameters. Different from the convolution layer, there are no specific parameters in the pooling window. Just slide the pooling window on the characteristic matrix from left to right according to the pooling step, and simply output the maximum or mean value in the window, and finally output the fused feature matrix.

In addition to convolution layer and pooling layer, batch normalization layer can be added after each convolution layer in 1D-CNN model, which aims to make each convolution output data have the same distribution. At the end of the model, flatten layer, dropout layer and dense layer are set to jointly extract the features of data applied to emotion classification and recognition.

2.4 Emotion recognition model based on feature stitching and SVM algorithm

In order to fully integrate and make use of the deep emotional features of the extracted multiple physiological data, firstly, all feature data are spliced into a group, and a complete set of feature data set is obtained as the input data of emotion classification model. Finally, the study select an appropriate machine learning classification algorithm, input the feature dataset into the model for training, and build a scientific and accurate emotional recognition model. In this study, the model construction effects of several machine learning algorithms are analyzed and compared, including Support Vector Machine (SVM)[16], K-NearestNeighbor (KNN)[17], RandomForest (RF)[18], Logistic Regression(LR)[19], Extreme Gradient Boosting (XGBoost)[20]. This paper describes the SVM-based emotional recognition model. SVM is a powerful and comprehensive machine learning model that can perform linear or non-linear classification and regression tasks, and is particularly suitable for small and medium-sized complex datasets. The process of solving the optimal model in SVM algorithm, that is, the process of solving support vector, which can be expressed as solving the constraint problem:

$$\max \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \, K(x_i x_j) \tag{2}$$

$$\text{s.t.} \qquad \alpha_i \geq 0, \quad i = 1,2,\cdots,n \tag{3}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{4}$$

3 Experimental settings & analysis of results

3.1 Data preprocessing

The experimental data for this study are adopted from DEAP, a publicly available data set experimentally collected by Koelstra et al. at Queen Mary's college, University of London, London, UK[21]. The data set collected EEG and peripheral physiological signals at a sampling rate of 512 Hz generated by 32 subjects during viewing 40 1-minute audio-video sessions as emotional triggers, and collected 1-9 ratings on arousal, validity, dominance and liking dimensions after each video session to reflect the emotional state during viewing different videos.

This study selects the data set as the research data after 128Hz down sampling pretreatment, including 32-channel EEG data, 2-channel EOG data, 2-channel EMG data, and 1-channel GSR, RESP, PPG, TEMP data. Each sample of each channel data contains 3s resting baseline data and 60s emotion-induced data, totaling 63×128=8064 sample points. In terms of data preprocessing, in order to reduce the interference of objective factors such as environmental equipment and better reflect the changes of physiological signals and emotional characteristics after emotional induction, this research subtracts the experimental data of the last 60s from the mean of baseline data of the previous 3s of each sample, and retains 8064-3×128=7680 sample points. Then, min-max normalization is performed on all channel signal data to fit the training and improve the effect of network model. Finally, each 60s sample data is divided into 10s segments, and each sample retains 7680/6=1280 sample points, and the number of samples is increased from 1280 to 7680. Before model training, data dimension splicing of related channel signals is performed, and the original sample sequence is disrupted, and then training set and test set are divided according to the ratio of 9:1.

In terms of processing the output label, referring to the two-dimensional evaluation system of emotional arousal-valence, considering that the evaluation range of these two dimensions on DEAP data set is 1-9, the emotional tags are divided into four categories based on the threshold of 5. The four emotional output tags are high valence high arousal (HVHA), low valence high arousal(LVHA), low valence low arousal(LVLA) and high valence low arousal(HVLA).

3.2 Model training parameters

The model training parameters of this study were continually optimized and adjusted during model trial building and training, culminating in setting that learning rate is 0.0001, model optimizer is Adam, and sample batch is 128. The specific network hierarchy of the model is shown in Fig.2, with two one-dimensional convolutional layers (Conv1D) setting the number of convolution kernels to 64, the size of the convolution kernel to 5, the convolution step to 3, the filling method to adopt the complementary 0 strategy, the initialization method to use the he normal distribution initialization method, and the regularization method to use l2 regularization. Batch normalization layer is added after each of the two convolution layers, and the activation layer adopts the relu activation function. The two pooling layers (MaxPooling1D) both adopt the maximum value pooling strategy, and the pooling window size is 4. At the end of the model, a flatten layer, a dense layer and a dropout layer are set. Flatten layer converts the input data format into the input requirements of the dense layer. Dense layer is set with 30 neuron nodes, that is, the network fuses and extracts 30 depth features. The initialization method adopts he normal distribution initialization method, the regularization method adopts l2 normalization, and the dropout layer sets the random deactivation coefficient to 0.2. Finally, the output layer outputs four kinds of emotion classifications predicted by the models to evaluate the effect of feature fusion extraction, and the activation function is softmax.

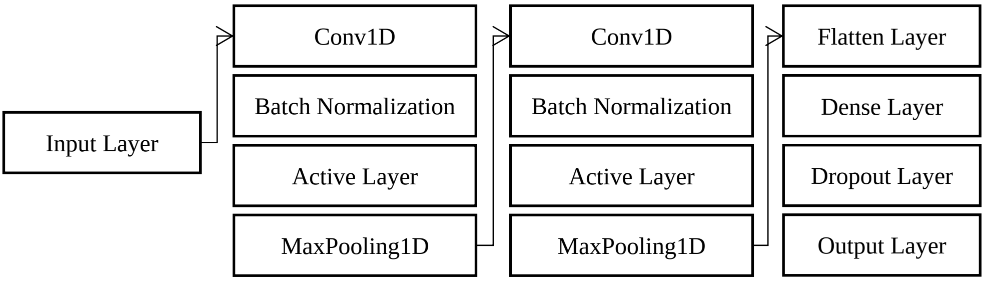


Fig. 2. Hierarchical structure of 1D-CNN feature fusion and extraction network.

The kernel function of each SVM classifier is set to Radial Basis Function (RBF), the multi-classification strategy is OVR(One Vs Rest). The hyperparameters C and Gama make different optimization adjustments and different settings for the fusion feature data extracted from different physiological signal data splicing combinations, so as to show the better emotion recognition effect under different data splicing combinations.

3.3 Results and analysis

In order to complete the task of emotion recognition and classification more accurately and efficiently, 1D-CNN-SVM multi-source physiological signal emotion recognition model is constructed by using the physiological signal data such as EEG, EOG, EMG, GSR, RESP, PPG and TEMP from DEAP data set as the input data of research. Specifically, firstly, simple multi-source data dimension splicing is achieved for data level fusion, and different data splicing combinations are divided. Then, 1D-CNN model is constructed to extract the fusion features of multi-source physiological signal data respectively. Finally, feature splicing is carried out and emotion recognition is realized by SVM classification algorithm. The model complicates the four-class tasks under the arouse-valence two-dimensional emotional evaluation model. In order to reflect the advantages of multi-source data fusion recognition, this study explores the optimal data splicing combinations. Based on the model in the research, experimental comparisons are made among various data splicing combinations. The classification accuracy of four-class tasks is achieved by the fusion features of different data splicing combinations as shown in Fig.3 below. The classification accuracy of the emotional recognition models constructed by single-source data such as EEG, EEGa (the first 16 channels of EEG), EEGb (the last 16 channels of EEG), PHY (the 8 channels of peripheral physiology), EOG, EMG are 85.94%, 85.16%, 75.13%, 65.36%, 62.50%, 56.25%, while that of the multi-source data combinations such as EEG+PHY, EEG+EOG+EMG, EEGa+EEGb+EMG, EEGa+EEGb+PHY are 89.19%, 90.89%, 91.93%, 93.10%, respectively. The experimental results fully reflect the advantages of multi-source data fusion in emotional recognition and verify the validity of the 1D-CNN-SVM emotional recognition model in this paper. The classification accuracy of the feature fusion model of EEG data and peripheral physiological data is generally higher than that of single source data. To compare with the highest accuracy, the multi-source EEGa+EEGb+PHY is 7% higher than single source EEG, breaking through to more than 90%.

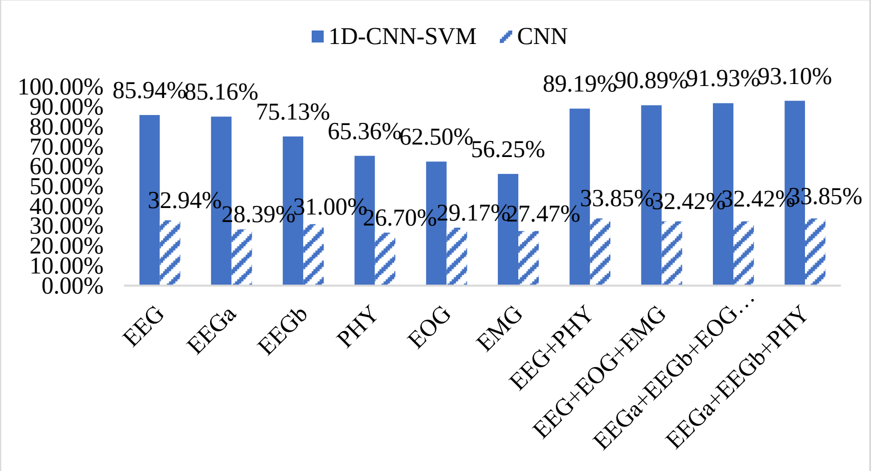


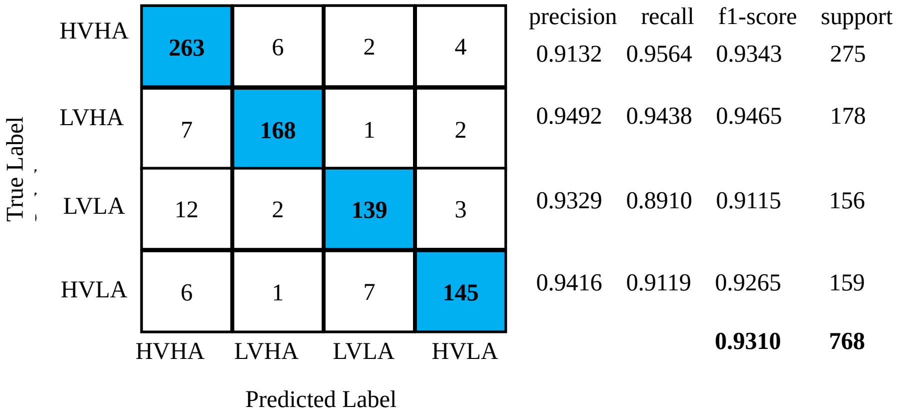
Fig. 3. Comparison of accuracy of emotional recognition models with different data splicing combinations.

In addition, in order to verify the validity of the combination of CNN and SVM, this study constructs a separate CNN model for emotional four-classification by selecting various data splicing combinations, and compares the classification accuracy of the 1D-CNN-SVM model constructed in this study with that of the CNN model. As shown in Figure 3 above, the accuracy of the CNN model is only about 30%, which is much lower than the emotion recognition model constructed in this study, which fully verifies the accuracy advantage of the model in the research.

Finally, in order to compare the classification accuracy of emotion recognition achieved by different machine learning classifiers, comparative experiment is conducted on the EEGa+EEGb+PHY, the best data splicing combination in the 1D-CNN-SVM model. Fusion features are extracted by using 1D-CNN model and input feature data sets are obtained. Classifiers based on KNN, RF, LR and XGBoost are constructed respectively. The classification accuracy of each model is shown in table 1 below. Experiments show that the unique advantages of SVM classifier in this model are sufficient to verify the significant improvement of classification accuracy brought by the application of SVM. Fig.4 below shows the emotion four-class performance of the 1D-CNN-SVM data-fused emotional recognition model with EEGa+EEGb+PHY as the multi-source input data in the test set of 768 data samples. The left part is a 4×4 square matrix consisting of four types of vertical real labels and four types of horizontal model predictive labels. The four blue squares diagonally from the top left to the bottom right in the matrix represent the number of correct classification of the four types of emotional labels achieved by the model respectively. The right part is the specific classification numerical indicators of the emotion recognition model in the four types of emotion labels, including precision, recall, f1 score and support.

Table 1. Classification accuracy with different machine learning classifiers.

Classifier	SVM	KNN	RF	LR	XGBoost
Precision	93.10%	88.93%	85.29%	87.76%	90.49%



True Label	HVHA	LVHA	LVLA	HVLA	precision	recall	f1-score	support
HVHA	**263**	6	2	4	0.9132	0.9564	0.9343	275
LVHA	7	**168**	1	2	0.9492	0.9438	0.9465	178
LVLA	12	2	**139**	3	0.9329	0.8910	0.9115	156
HVLA	6	1	7	**145**	0.9416	0.9119	0.9265	159
							0.9310	**768**

Predicted Label

Fig. 4. Specific classification performance of 1D-CNN-SVM emotion recognition model.

4 Conclusions

In this paper, the construction of emotion recognition model based on 1D-CNN-SVM is proposed. Firstly, the model uses the concept of data level fusion to splice the data dimensions of multi-source physiological signal data, and then input different data splicing combinations into 1D-CNN feature fusion and extraction model to realize feature level data fusion and extract the fusion features of data. Finally, the feature data set is obtained by feature splicing and input into SVM classifier to achieve output of emotion category.

The proposed emotion recognition model based on 1D-CNN-SVM with data fusion of multi-source physiological signals EEGa, EEGb and PHY achieves excellent result of four-class tasks on public affective dataset DEAP. The validity of this model has been preliminarily verified by experiments. The innovations of this paper are as follows. At the data level, the less studied combination of EEG data and peripheral physiological data is considered. At the feature level, a more efficient deep learning method is considered to construct the CNN feature fusion extraction network. At the fusion level, the less applied combination of data level data fusion and feature level data fusion is considered. At the model level, the combination of CNN and SVM is proposed to construct 1D-CNN-SVM emotion recognition model.

Future research perspectives are as follows. Firstly, the network hierarchy of the 1D-CNN feature fusion extraction model for each part of the multisource physiological signal data is consistent, and the differences of different physiological signal data are not fully taken into account. The differences of adaptivity for the same feature fusion extraction model may result in large differences in model accuracy. Feature fusion models of different network hierarchies need to be explored more deeply to achieve more optimal model effects. Secondly, the emotional recognition model proposed in this paper has been preliminarily validated in DEAP dataset, and other emotional datasets can be tried to validate the model in the future, in order to seek more ideas for model optimization and adjustment. Finally, this study mainly considers the data fusion of physiological signal closely related to personal emotion. In fact, there are many other types of data related to emotion recognition, such as facial image data and voice data, etc. Implementing data fusion of different types of data to improve the practicability and efficiency of emotion recognition model is the direction of future research.

References

1. Qiao Jianzhong. Emotional Research: Theory and Method[M]. Nanjing: Nanjing Normal University Press, 2003:1-10 (2003)
2. William James. What is an Emotion?[J]. Mind, 9(34) (1884)
3. Lange C.G. The emotions: a psychophysiological study[J]. The emotions, 33-90 (1885)
4. Canon W. The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory[J]. The American Journal of Psychology, 39(1/4) (1927)
5. Bard P. Emotion: I. The Neuro-humoral Basis of Emotional Reactions[J]. A handbook of general experimental psychology, 264–311 (1934)
6. Nie Yi, Wang Xiaopang, Duan Ruo Male, Lv Baoliang. A review of emotional recognition based on EEG[J]. Chinese Journal of Biomedical Engineering, 31(04): 595-606 (2012)
7. Dunlina, Wang Xiaoliang. A review of emotional recognition based on physiological signals[J]. Internet of Things Technology, 11(07): 33-41 (2021)
8. Pan Jiahui, He Zhipeng, Li Zina, Liang Yan, Qiu Lina. A review of multi-modal emotional recognition[J]. Journal of Intelligent Systems, 15(04): 633-645 (2020)
9. He You, Lu Daju, Peng Yingning. Overview of multi-sensor data fusion algorithms[J]. Fire and Command Control, 1996 (01): 12-21 (1996)
10. Xu Xiaoqin. Overview of multi-sensor data fusion target recognition algorithms[J]. Infrared and Laser Engineering, 2006 (S4): 321-328 (2006)
11. Zhang Baomei.Research on Data Fusion Methods at Data Level and Feature Level [D]. Lanzhou University of Technology (2005)

12. Pan Lizheng, Yin Zeming, She Shigang, Yuan E-E, Zhao Lu. Study on emotional recognition based on FCA-ReliefF fusion physiological signals [J]. Computer Measurement and Control, 28(02): 179-183 (2020)
13. Yongrui Huang,Jianhao Yang,Siyu Liu,Jiahui Pan. Combining Facial Expressions and Electroencephalography to Enhance Emotion Recognition[J]. Future Internet, 11(5) (2019)
14. C. Maaoui,F. Abdat,A. Pruski. Physio-visual data fusion for emotion recognition[J]. IRBM, 35(3) (2014)
15. Backpropagation Applied to Handwritten Zip Code Recognition Y.LeCun, B.Boser, J.S.Denker, D.Henderson, R.E.Howard, W.Hubbard, and L.D. Jackel Neural Computation 1989 1:4, 541-551 (1989)
16. Cortes, C.; Vapnik, V. Support-vector networks. Machine Learning. 20(3):273–297 (1995)
17. Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 46 (3): 175–185 (1992)
18. Tin Kam Ho, "Random decision forests," Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, pp. 278-282 (1995)
19. Verhulst P.F. Recherches mathematiques sur la loi d'accroissement de la population. Nouveaux Memoires de l'Academie Royale des Sciences,des Lettres et des Beaux-Arts de Belgique, 18, 1-38 (1845)
20. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[C]. the 22nd ACM SIGKDD International Conference. ACM (2016)
21. Koelstra S, Muhl C, Soleymani M, et al. DEAP: a database for emotion analysis; using physiological signals[J]. IEEE Transactions on Affective Computing, 3(1): 18-3 (2011)