Facial Emotion Recognition Based on CNN

Shuang Liu
Tianjin Key Laboratory for Control Theory and Applications in Complicated Systems
Tianjin University of Technology
Tianjin,China
liushuang0926@qq.com

Dahua Li*
Tianjin Key Laboratory for Control Theory and Applications in Complicated Systems
Tianjin University of Technology
Tianjin,China
lidah2005@163.com

Qiang Gao
Tianjin Key Laboratory for Control Theory and Applications in Complicated Systems
Tianjin University of Technology
Tianjin,China
gaoqiang@tjut.edu.cn

Yu Song
Tianjin Key Laboratory for Control Theory and Applications in Complicated Systems
Tianjin University of Technology
Tianjin,China
jasonsongrain@hotmail.com

2020 Chinese Automation Congress (CAC) | 978-1-7281-7687-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/CAC51589.2020.9327432

***Abstract*—With the development of artificial intelligence, computers will have not only IQ but also EQ in the future. Affective computing, which makes computers have emotion, has received more and more attention in recent years. Among them, facial expression recognition has become a research hotspot in the field of affective computing. In this paper, facial expression recognition is studied based on Valence-Arousal dimensional emotion model. A facial expression valence dimension prediction system based on convo-lution neural network is designed in this study. The system in-cludes face detection, feature extraction, valence grade prediction and so on. In this system, the annotation of facial expressions is divided into 9 levels. and the probability of each valence dimen-sion is obtained through the output of CNN network, and the final prediction result is equal to the weighted fusion of valence value and its corresponding probability. We use CK+ database and Fer2013 database to complete the training of CNN network model, and verify the performance of the system by recognizing the facial expressions of volunteers when watching video. The results show that the system can correctly predict the emotional effect value of volunteers, and the average RMSE index is 0.0857±0.0064.**

Keywords—emotion recognition, valence-arousal dimensional emotion model, convolution neural network, regression prediction

I. INTRODUCTION

Emotional expression is an effective way of communication and the basis of mutual understanding, unity and cooperation between people. With the development of computer vision and artificial intelligence, the research on human emotion in video and image has become a hot topic in the field of machine learning and pattern recognition. In the future, human-computer interaction will be more intelligent, fluent and fast, and computers can perceive, capture and distinguish human emotions and emotional changes, and based on this to make efficient and intelligent responses, that is, to give machines "brains". Enable machines to perceive and identify emotions to meet the daily needs of human beings.

At present, a large number of researchers have done in-depth research in the field of emotion recognition. Emotion recognition can be divided into facial emotion recognition, speech emotion recognition, language recognition and physiological pattern recognition. Among them, facial expression is one of the most effective, natural and common signals for human beings to express their emotional state and intention [1]. The research direction of emotion recognition includes traditional emotion recognition and dimensional emotion recognition [2]. Traditional emotion recognition is generally treated as a pattern recognition problem, using a classifier to classify emotions into fixed categories of discrete emotion tags, including: happiness, anger, sadness, surprise, calm and so on.

Due to the continuous development of affective computing, people have more in-depth requirements for emotion recognition, and the traditional discrete emotion recognition model can not express the rich emotions of human beings. Therefore, dimensional emotion recognition has received extensive attention in recent years. Dimensional emotion means that emotion is multi-dimensional, which is composed of many dimensions. For example, Figure 1 is called the V-A dimension emotional model. The horizontal axis is the valence dimension, indicating the degree to which the emotion is positive or negative, and the vertical axis is the arousal dimension, indicating the degree to which the emotion is calm or excited. Human emotions can be quantified at some point in the emotional space, indicating complex categories of human emotions, such as joy, panic, surprise and so on. Dimensional emotional space covers any human emotion, so dimensional emotional recognition provides more sufficient and effective support for emotional perception, affective computing, facial expression recognition and other research fields.

For the study of dimensional emotion recognition, the common method is to transform the basic six kinds of emotion categories into three kinds of price-related classification problems, that is, to divide the emotion space into three small emotion spaces: neutral, positive and negative. In 2008, the researcher M.Wollmer [3] used the probability graph model Condition Random Field (CRF) with Markov property to predict the quantitative dimension annotation. In addition, the method of regression model can also be used to predict the

emotional state or dimension attribute value of emotion in continuous dimensional space [4]. Many research methods tend to focus on the continuous and sequential process of emotional expression, which quantifies the emotional space into different levels to predict the quantified annotation. For example, in 2009, Theodoros Giannakopoulos et al [5] extracted voice audio data from the movie, marked the dimension attribute value of emotion in speech, and predicted the emotional coordinate values in the emotional space of dimension A by using the method of KNN. In 2013, Nicolaou [6] proposed the Channel and Space Reliability (CSR) algorithm, which maps the feature space to the dimensional emotional space and studies the correlation of the dimensional space.

In recent years, with the significant improvement of computer processing power and the improved design of network architecture, the field of emotion recognition has begun to turn to deep learning methods. The research shows that the convolution neural network model has achieved excellent results in target detection and facial expression recognition, and with the deepening of the network layer, the recognition effect of the network is further improved, such as AlexNet [7] has 5-layer convolution. Deep learning methods have made considerable achievements in computer vision, pattern recognition, image processing and other aspects.

In this study, a facial expression valence dimension prediction system based on convolution neural network is designed. The generalization ability of the CNN model is improved by the combination of L2 regularization and Dropout, and the CNN network model is optimized by introducing Adam algorithm. We use CK+ database [8] and Fer2013[9] database to complete the training of CNN network model, and verify the performance of the system by recognizing the facial expressions of volunteers while watching the video. The results show that the system can effectively recognize different expressions in the valence dimension.

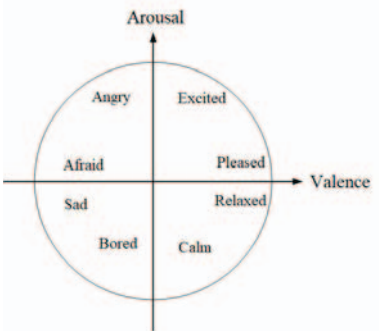


Fig. 1. The V-A dimension emotional space

II. NETWORK STRUCTURE DESING OF CNN

Convolution layer, pooling layer and full connection layer are the key components of convolution neural network. The convolution layer is mainly composed of several convolution kernels, which are used to perceive the features of different parts of the image. The main parameters of convolution layer include the size of convolution kernel, the step size of convolution kernel translation and the mode of padding. The convolution kernel is the key component of the convolution layer. The size of the convolution kernel refers to the number of rows and columns of the convolution kernel matrix. Usually, 3 × 3 and 5 × 5 are common convolution kernel sizes. The translation of the convolution kernel is that the convolution kernel senses the characteristics of each part of the image from the upper left corner of the image through the convolution operation from left to right, and the step size of the convolution kernel translation is set to 1. The Padding method artificially introduces 0 on the outside of the original image data matrix, so that the convolution kernel can perceive the information of the pixels located at the edge of the image for many times. Pooling layer is an important part of convolution network, which is actually a downsampling process, so pooling layer can also be called downsampling layer (Sub-sampling Layer). Average pooling and maximum pooling are the most common pooling modes, similar to the translation of convolution cores. The pooling layer divides the input image into several regions through the window and translation step size of the convolution core, and outputs the maximum or average value in each region. The pooling layer can achieve sparse features and reduce the computational complexity of the whole network, which suppresses the over-fitting phenomenon to a certain extent and improves the generalization ability of the model. Set the window size of pooling to 2 × 2, the translation step to 2, and an example of average pooling and maximum pooling, as shown in Figure 2.

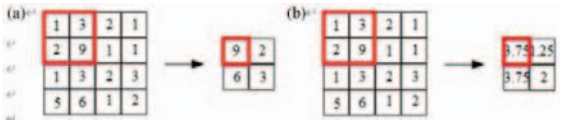


Fig. 2. Schematic diagram of pooled layer operation

(a)maximum pooling (b) average pooling

In this study, a CNN network is designed to predict the valence dimension of facial expression, and this problem is transformed into the problem of making nine classifications in the valence dimension. The probability of each valence dimension of the network output is equal to the weighted fusion of the valence value and its corresponding probability, as shown in Formula (1).Where Vpred represents the final prediction result of the valence dimension, i represents the valence value of 1-9, mapminmax represents the range of values that map the predicted value to the valence dimension [-1,1], and pi represents the prediction probability of a certain valence value.

$$V_{pred}=mapminmax(\sum_{i=1}^{i=9} i*p_i, -1, 1) \quad (1)$$

The CNN designed in this paper has nine hidden layers, one input layer and one output layer. In the input layer, the input is a 48×48 grayscale facial expression image. Four convolution layers and three pooling layers can be used for local feature perception and sparse features of facial expression images, resulting in 64 6×6 feature map. In this paper, 1×1 convolution kernel is applied in convolution layer 1, and combined with same padding and relu activation function, the introduction of nonlinear expression of image without loss of image resolution can make the network deeper and indirectly improve the

generalization ability of the model. After that, the feature map is expanded and input to the full connection layer for local feature integration, so that the network can learn globally. Considering that the network is difficult to train and may have over-fitting phenomenon due to the large number of parameters in the full-connection layer, parameter initialization optimization strategy and regularization strategy are added to the full-connection layer to improve the possible over-fitting phenomenon. Finally, the probability of taking a value of 1-9 is output through the softmax layer. The structure diagram and parameter information of each layer of the network are shown in Figure 3 and Table I:

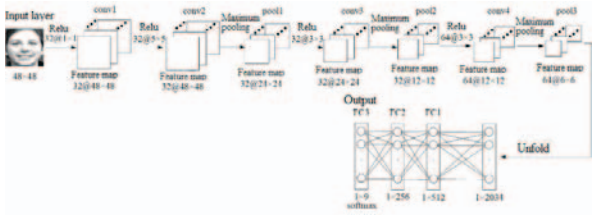


Fig. 3. CNN network structure diagram

TABLE I. TABLE TYPE STYLES

Network layer	Number of parameters	Optimization strategy
Input layer	------	Input feature normalization
Convolution layer 1 (Relu, 32@1×1)	32×1×1×1+32=64	Same padding
Convolution layer 2 (Relu, 32@5×5)	32×5×5×32+32=25632	Same padding
Pooled layer 1 (maximum pooling, 2 × 2)	------	------
Convolution layer 3 (Relu, 32@3×3)	32×3×3×32+32=9248	Same padding
Pooled layer 2 (maximum pooling, 2 × 2)	------	------
Convolution layer 4 (Relu, 32@3×3)	64×3×3×32+64=18496	Same padding
Pooled layer 3 (maximum pooling, 2 × 2)	------	------
Fully connected layer 1 (Relu, 512)	576×512+512=1180160	L2 regularization (parameter 0.05); dropout (0.4); weight matrix w random normal distribution initialization
Fully connected layer 2 (Relu, 512)	512×256+256=131328	L2 regularization (parameter 0.05); dropout (0.4); weight matrix w random normal distribution initialization
Full connection layer 3 (output layer) (Softmax, 9)	1024×9+9=2313	Tags: one-hot coding Cross entropy (cross-entropy) loss function

Deep network should not only perform well through the training set, but also ensure its good generalization performance. In the construction of machine learning model, the regularization method is used to reduce the error of the test set. It can be seen from Table I that two regularization methods (L2 regularization and Dropout) are applied to improve the generalization ability of CNN.

Weight parameter attenuation is a direct and effective regularization strategy in machine learning, and L2 regularization is a regularization strategy of weight parameter attenuation. This method adds a regularization term Ω(w) of weight L2 norm to the loss function, which can be expressed by formula (2):

$$\Omega(w)=\frac{\alpha}{2}\|w\|_2^2 \tag{2}$$

Where a represents the regularization coefficient, which is a super parameter. Let the original ground loss function be J, and the ground loss function after adding regularization term is Ĵ, which can be expressed as:

$$\hat{J}(w;X,y)=\frac{\alpha}{2}\|w\|_2^2+J(w;X,y) \tag{3}$$

Where X represents the characteristics of the sample, y represents the annotation of the sample, and the corresponding gradient can be expressed as:

$$\nabla_w\hat{J}(w;X,y)=\alpha w+\nabla_w J(w;X,y) \tag{4}$$

Suppose the learning rate is ϵ, and the weight parameter update is expressed as:

$$w \leftarrow w - \epsilon\alpha w+\nabla_w J(w;X,y) \tag{5}$$

Dropout is also a regularization strategy, which makes the nodes of the hidden layer be discarded according to a certain probability, that is, "random inactivation". Dropout provides a parallel training method of multi-subnetworks with low computation to solve the over-fitting problem.

In each training, batch training is adopted and a small learning rate is set, and then part of the neuron nodes are "randomly inactivated" according to the probability set in each layer, so as to eliminate the joint adaptability of the nodes as far as possible and enhance the generalization ability of the model. the amount of computation is reduced. Therefore, this paper chooses the combination of L2 regularization and Dropout to improve the generalization ability of the constructed CNN model. In addition, this paper selects Adam algorithm to optimize CNN.

III. IMPLEMENTATION OF CNN

In this paper, the implementation of CNN is based on the Keras deep learning framework. The advantage of Keras is that it is user-friendly and modular. Developers call the integrated network layer module to build the network, which reduces the development cost and the code is easy to debug. In this paper, the CNN network is built on the Python platform under the Windows system, the Python version is 3.5.2, and Tensoflow-gpu is selected as the back-end of Keras, and the Tensorflow version is 1.7.0. And Tensorflow supports using GPU to realize the operation of network training, so this paper uses NVIDIA GeForce GTX 950M to complete the operation of network training and configures the engine CUDA 9.0 of GPU operation.

As shown in Table I, the activation functions of convolution layer, full connection layer 1 and full connection layer 2 are all set as relu activation functions, and the pooling layer adopts the maximum pooling of 2 × 2 window size, and cross entropy is used as the loss function. Adam as the optimization algorithm of the network, the learning rate is 0.001. The first moment attenuation coefficient of the gradient and the second moment attenuation coefficient of the gradient are set to 0.9 and 0.999 respectively. The network uses batch training of 32 samples per group, and the epoch is set to 30.

IV. IMAGE PREPROCESSING

The image preprocessing work in this paper mainly includes the valence dimension annotation and face detection of the image. First of all, the valence dimension of the image is annotated. 10 annotationers average the images of the training set and the test set based on the SAM system[11]. Each annotator needs to annotation each picture with a valence dimension of 1-9 based on the SAM system. Because the valence dimension of this paper is mapped in the range of [-1,1], the 1-9 annotation corresponds to the nine levels (L1-L9) defined by the valence dimension of the SAM system, and the average annotation rounding is finally used as a annotation for classification tasks. Considering that the film clips mainly involve three kinds of emotions: calm, sadness and happiness, the training set and test set are mainly composed of facial expression images in CK+ database and Fer2013 database. There are a total of 1141 photos in the training set and 110photos in the test set. The quantitative distribution and annotation examples of each valence dimension image are shown in Figure 4 and Figure 5.

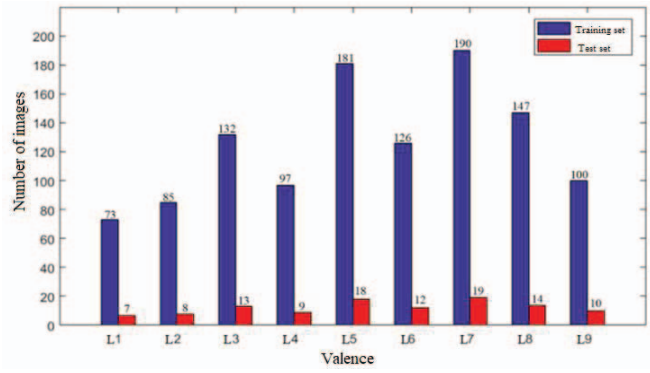


Fig. 4. Data distribution of training set and test set

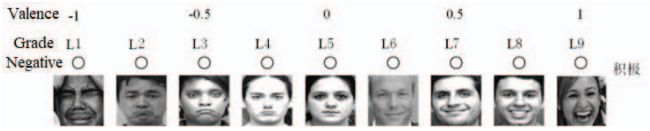


Fig. 5. Example of picture annotation

The second step of image preprocessing is face region detection, which sets the face region to a uniform size. this step is completed with the help of open source computer vision library (Open Source Computer Vision Library, OpenCV) on Python platform. The version of OpenCV used in this paper is 2.7.0.

In this study, the "haarcascade_frontalface_alt" cascade classifier in OpenCV library is selected to detect face regions. The steps of face detection are as follows. First, import the image of the face region to be detected and convert the imported image into grayscale image, then import the cascade classifier model for face detection. Then the method of multi-scale detection is called to detect the face in the grayscale image containing human face, and then the coordinates of the face detection point (x, y), the width w of the face and the height h of the face are obtained for the annotationing of the face region. Annotation and display the face region, and then the scale of the face region is converted into a uniform scale of 48 × 48 and input to CNN. Finally save the image to complete the face detection of the current image; cycle the above steps until the completion of face detection in all images. An example of face detection is shown in Figure 6:

Fig. 6. Face detection example based on OpenCV

V. EMOTION RECOGNITION RESULT BASED ON CNN

After data processing, we train the proposed CNN. The loss of training set and test set and the classification accuracy of training set and test set are shown in Figure 7. From the graph, we can see that the loss of the training set first decreases and then converges, and after the 15th epoch, it basically converges around 0.7. It can be seen that CNN improves the classification ability through the learning of the training set. Although the loss of the test set also decreases at first, the loss of the test set tends to increase after the 20th epoch with the increase of epoch. In terms of accuracy, the accuracy of the training set first increases and then converges, and the accuracy of the test set converges around 0.9 at the 15th epoch, and the accuracy of the test set converges around 0.7 at the 15th epoch. Although we have adopted regularization methods such as L2 regularization and dropout, as well as other optimization methods, there is still some over-fitting in the trained CNN network.

Considering that what CNN completes is the classification of emotional valence dimension, it can be found that there is no significant difference in facial expressions between adjacent valence dimensions in actual emotion annotation, and it is more difficult to classify valence dimensions than the task of identifying different types of emotions, which is one of the reasons why CNN does not achieve the desired effect in the test set. On the other hand, the prediction results of the valence dimension are obtained according to the weighted fusion of the prediction probability and effect value of each emotion level output by CNN, which makes the prediction results closer to the actual situation. Generally speaking, the CNN network

gains the ability to classify the valence dimension through the training of the training set, and the trained CNN will be used for the facial expression recognition of the subjects.

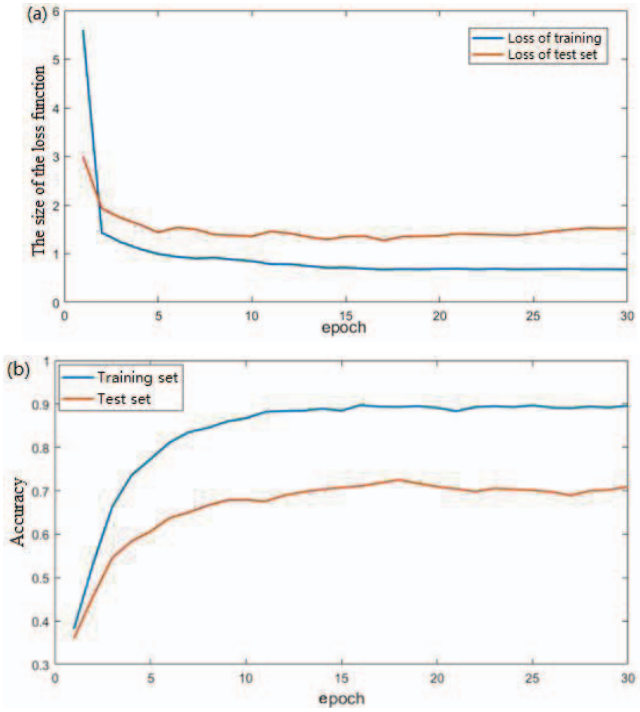


Fig. 7. Performance of training set and test set

(a)Loss of training set and test set (b)The accuracy of training set and test set

Ten healthy subjects, including five men and five women, took part in the experiment. The subjects sat in a comfortable chair in the 50cm directly in front of the computer screen, and the camera was supported by a bracket and placed close to the right side of the screen. Each subject needs to complete the experiment of watching all six movie clips in turn, and a total of 60 films are completed. The subjects' facial expressions were recorded by the camera during the short film.

We selected movie clips from the SEED dataset [11] to stimulate the emotion of the subjects. These film clips are from Chinese movies. Choosing Chinese movies as stimuli can eliminate the differences caused by culture and customs, make the emotions conveyed by stimuli more clear, and thus induce the expected emotions of the subjects. Movie clips are divided into three categories: happy, calm and sad. Two different movie clips are selected from each category of movie clips, and Table II shows the details of the selected movie clips. The duration of each film clip is about 200 seconds (average = 212.83s, standard deviation = 8.71s).

After completing the experiment, we need to annotate the valence dimension of the facial emotional response of the subjects. In the visual emotion annotating interface combined with the joystick operation, 10 annotators completed the continuous annotation of 60 facial expressions of the subjects, and the value range of emotional annotation was customized to -1 to 1, as shown in figure 8. The visual interface is based on the custom design of DARMA, and DARMA is a multimedia annotation program running on MATLAB platform [12]. In this work, the sampling frequency of the rocker position is set to 20Hz and the output frequency is 1Hz. The annotation results can be displayed and saved together with the subjects' facial expression response video files to facilitate statistics and analysis. Finally, we extract a facial expression image every 2 seconds in each annotated facial expression video for CNN emotion recognition, and take 100 images for each video.

TABLE II. THE DESCRIPTION OF SEED DATASET FILM CLIPS

No.	Annotation of the film clips	Title of film clips	The length of time
1	positive	Lost in Thailand	1:04:57-1:08:20
2	positive	Flirting Scholar	1:19:57-1:23:23
3	neutral	World Heritage in China I	0:02:59-0:06:40
4	neutral	World Heritage in China II	0:00:50-0:04:36
5	negative	Aftershock	0:20:10-0:23:35
6	negative	Back to 1942	2:01:10-2:04:46

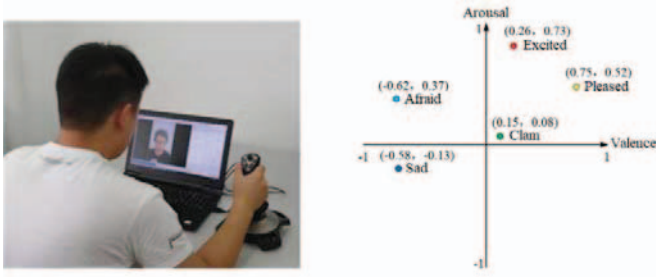
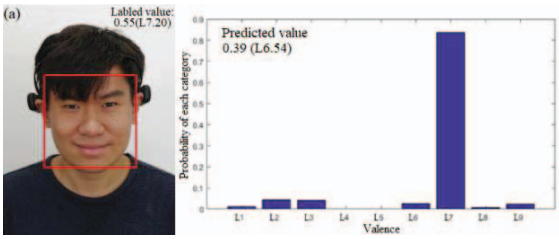


Fig. 8. The annotation environment and annotation examples Performance

Finally, the trained CNN is used to recognize the facial expressions during watching the movie. Figure 9 shows the representative results of the subjects' facial expression recognition of the happy film clip "Lost in Thailand", the calm film clip "World Heritage in China I " and the sad film clip " Aftershock ". From the results, it can be seen that the prediction of CNN network can correctly distinguish different emotion categories and give a more correct valence dimension prediction, which also proves the feasibility of facial expression emotion recognition based on CNN proposed in this paper. After that, CNN recognized the emotion of the images in the facial expression videos of ten subjects and achieved an average RMSE index of 0.0857±0.0064, as shown in Figure 9. Although facial emotion recognition based on CNN has achieved good results, CNN network needs a lot of data to learn in order to achieve better performance, but a large number of valence dimension annotated facial expression images are difficult to obtain, which is one of the reasons why CNN does not achieve the desired results.



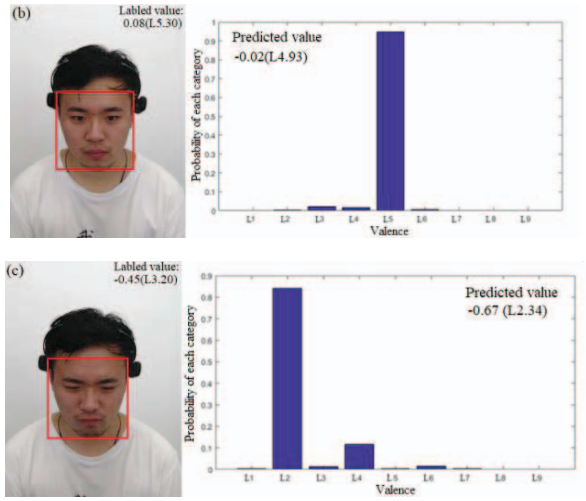


Fig. 9. Recognition results of typical expressions based on CNN

(a) Subject's expression when watching “Lost in Thailand” (b) Subject's expression when watching "World Heritage record of China I " (c) Subject's expression when watching “Aftershock”

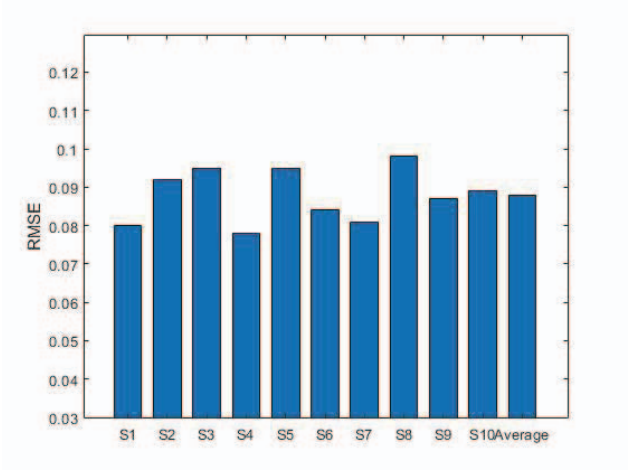


Fig. 10. Performance of emotion prediction based on CNN

VI. DISSCUTION

The facial emotion recognition system based on CNN designed in this paper achieves a RMSE index of 0.0857±0.0064. In this method, the cascade classifier is used to capture the face region, and then input to the CNN trained by CK+ and Fer2013 expression database to get the prediction results. The emotion recognition method is relatively stable and will not be affected by the facial physiological differences between the subjects. On the other hand, considering that it is difficult to classify the valence dimensions of the nine grades, that is, the difference in the expression of the adjacent valence dimensions is not obvious, this method does not directly select the valence grade of the maximum probability as the output, but chooses the weighted fusion of each effect value and the prediction probability. Although L2 regularization and dropout regularization are applied in this method, there is some over-fitting in the training and testing phase, which may be due to the insufficient number of facial expression images with valence dimensions. Although this method can correctly distinguish the emotional categories (sadness, serenity and happiness), it still needs to improve the recognition performance of the valence dimension.

REFERENCES

[1] Tian. Y. I., T. Kanade, and J. F. Cohn, “Recognizing action units for facial expression analysis,”IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001,pp. 97–115.

[2] Baltrusaitis, Tadas, N. Banda, and P. Robinson, “Dimensional affect recognition using Continuous Conditional Random Fields.” Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on IEEE, 2013.

[3] M. Wllmer, et al. "Abandoning Emotion Classes -- Towards Continuous Emotion Recognition with Modelling of Long-Range Dependencies," Asian International Conference on Interspeech, Conference of the International Speech Communication Association, 2008.

[4] M. Wollmer, B. Schuller, F. Eyben, and G. Rigoll, ″Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening, ″ IEEE Journal of Selected Topics in Signal Processing, 2010, pp:867–881.

[5] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, “A dimensional approach to emotion recognition of speech from movies,” IEEE International Conference on Acoustics, Speech and Signal Processing,2009.

[6] M. A. Nicolaou, H. Gunes, and M. Pantic, “Output-associative RVM regression for dimensional and continuous emotion prediction,” Image and Vision Computing, 2012, pp:186–196.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Communications of the ACM, 2017, pp:84–90.

[8] P. Lucey, J. F. Cohn, T. Kanade, “The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, USA, 2010.

[9] I. J. Goodfellow, D. Erhan, P. L. Carrier, “Challenges in representation learning: A report on three machine learning contests,” Neural Networks. 2015, pp: 59-63.

[10] M. M. Bradley, P. J. Lang, “Measuring emotion: The self-assessment manikin and the semantic differential,” Journal of Behavior Therapy and Experimental Psychiatry. 1994, pp: 49-59.

[11] I. J. Goodfellow, D. Erhan, P. L. Carrier, “Challenges in representation learning: A report on three machine learning contests,” Neural Networks. 2015, pp: 59-63.

[12] J. M. Girard, A. G. C. Wright, “DARMA: Software for Dual Axis Rating and Media Annotation,” Behavior Research Methods, 2017, pp: 902-909.