Classroom monitoring system based on facial expression recognition

Boxuan Zhang,Dandan Wei,Qianying Zhang,Wenyu Si,Xiang Li,Quanyin Zhu*
Faculty of Computer&Software Engineering,Huaiyin Institute of Technology,Huaian,China
*Corresponding author's e-mail:hyitzqy@qq.com

***Abstract*—Facial expressions are important information that reflects human emotions. Recognizing dynamic expressions of students in class, 8 kinds of emotions are selected for application: positive emotions: "happy"; negative emotions: "disgust , Sadness, doubts, contempt, anger"; neutral emotion: "focus, surprise". [1] In this design, the classroom performance scoring system in normal hours is split into four functions: wireless network list acquisition and verification, face recognition, emotion analysis, and scoring record storage. On this basis, SVM and Softmax are used. Facial expression recognition and a data storage database is designed to realize the function of an intelligent scoring system for classroom performance points. To solve this problem, an expression recognition method combining pyramid convolutional neural network and attention mechanism is proposed.**

Keywords-component; facial expression recognition; pyramid convolution; attention mechanism

I. INTRODUCTION

For the past few years, the continuous expansion of colleges and universities has resulted in a shortage of teaching resources to some extent, and the workload of schools is heavy, in the process of college education and student work management, the final exam is an important content, and class performance scores also account for a part of it. However, some problems have also arisen, such as incomplete recognition by teachers, the inability of teachers to grade students' class status during class, and the long time consuming scoring. This led to some mistakes in the scoring of teachers 'classroom performance. Therefore, a classroom monitoring system based on facial expression recognition and behavior prediction was developed to meet the actual needs of the current final assessment, such as schizophrenia and post-traumatic stress disorder.[2]

Therefore, we propose a facial expression recognition network based on the attention model to achieve the best detection effect by letting the system learn the important parts and ignore the unimportant parts. We will verify the superiority of this method in commonly used data sets.

This project is a classroom monitoring system based on facial expression recognition and behavior prediction.

II. RELATED WORK

A. Expression recognition from the classroom perspective

Facial expression,which is the most direct and effective emotion recognition mode, has many application scenarios. For instance, people use facial expression to detect the fatigue driving. However, we apply this technology to students' classrooms, and use this to help teachers make more accurate ending points. The basic expressions that people put forward at the beginning are:angry, scared, Disgust, happy, sad, surprised and neutral.

Traditional hand-designed features and even shallow learning features can no longer adapt well to various interference factors that are not related to expressions in the real world, such as light changes, different head postures, and facial blocking.

On this basis,the study of attention mechanism started to be applied to facial expression recognition.

For facial expression recognition in real scenes, direct occlusion caused by objects or indirect occlusion caused by factors such as illumination and posture changes is one of the inherent challenges of expression recognition. Because the occlusion in the real scene is complex and diverse, the feature reconstruction method relies on a large number of training data under different occlusion conditions to have better results, and the reconstruction of face details is not ideal. An expression recognition method that does not rely on the detection of key points on the face is needed to recognize expressions in real scenes.

B. Attention mechanism

An important feature of the human visual system is that it does not try to process the entire scene at once, but selectively focuses on the target area that needs attention to obtain the detailed information of the target, while suppressing other useless information, which greatly improves the visual information processing Rate.

Zhang et al. [3] proposed an attention layered bilinear pooling residual network for expression recognition. This method uses the channel attention mechanism to explicitly model the importance of each channel, and assigns different output feature maps. The weight of, locate the salient area according to the weight value. Li et al. [4] used the spatio-temporal attention mechanism to recognize facial micro-expression, and the temporal attention module was used to learn the motion information of the expression sequence, focusing on the more discriminative frames in the expression sequence.Wang et al. [5] proposed a new type of regional attention network, which extracts the features of each region through the backbone convolutional network, and weights the attention feature information to improve the accuracy of facial expression recognition under occlusion and posture changes. Amir et al. [6] proposed a deep attention center loss (Deep Attentive Center Loss, DACL) method. The proposed DACL integrates an attention mechanism and uses the intermediate spatial feature map extracted by CNN as the context to estimate and feature Importance-related attention weights to adaptively select a subset of important feature elements to enhance discrimination.

Based on the method proposed by Wang [5], this paper proposes a global attention module. The attention

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00034

mechanism allows more prominent features to be selected according to needs. The global attention module in this paper can better solve the problem of real scene expression recognition.

C. Pyramid Convolution

In the training process of deep neural networks, image feature information is extracted through convolution operations, and the spatial features that can be learned by convolution kernels of different sizes are not the same. For small targets and targets with noise, detailed feature information is very important, and pixel-level deviations often lead to errors in recognition. Pyramidal Convolution (PyConv)[6]can process input information through multiple convolution kernels of different scales. The main advantage of PyConv is multi-scale processing, with different spatial resolutions and depths. Compared with standard convolution, PyConv can expand the receptive field of the convolution kernel without increasing additional costs. Pyramid convolution can capture the diversity of expression features and the variability of their scales, and maintain the continuity of facial action units.

III. EASE OF USE

A. Sub-image generation

The neurons in each layer of the convolutional neural network are arranged in three dimensions: width, height, and depth. The width and height are well understood, because the convolution itself is a two-dimensional template. However, the depth in the convolutional neural network refers to the third dimension of the activated data volume, not the depth of the entire network. The depth of the entire network refers to the number of layers of the network.

We will see that the neurons in the layer will only be connected to a small area in the previous layer, instead of being fully connected.

B. Pyramid Convolutional Network

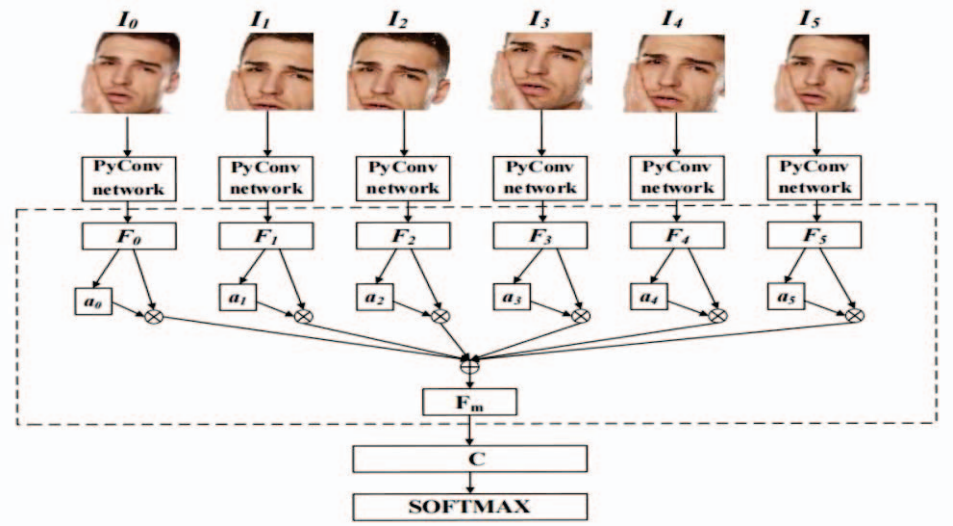

Fig.1 PyConv-Attention network mod

The key problem of facial expression recognition is to find the expression feature areas with prominent expression changes. The attention mechanism is widely regarded as a method to help solve this type of problem. Studies have shown that the mouth, eyes, eyebrows and nose are formed Different action units, the combination of these facial units forms face expressions. PyConv is a pyramid convolution unit. As shown in Figure 2, it is composed of convolution kernels of different sizes and different depths. As the size of the convolution kernel increases, its depth decreases accordingly. These convolution kernels can capture Different levels of detail features in the image. For the centi-word tower convolution unit, due to the use of convolution kernels of different depths, a grouped convolution method is adopted. The input feature map is split into several parts, and the convolution kernels of different depths are used for each group of input feature maps. Feature extraction. When the grouping is 1, it is a standard convolution, where the depth of the convolution kernel is equal to the number of channels of the input feature map.

C. Attention module

The proposed global attention module is shown in the dashed box in Figure 1. After the original image and sampled sub-images are extracted from the centripetal convolutional network, they are sent to the global attention module to calculate the image's value.

D. Attention module

The proposed global attention module is shown in the dashed box in Figure 3. After the original image and sampled sub-images are extracted from the centripetal convolutional network, they are sent to the global attention module to calculate the image's value through a fully connected layer and Sigmod activation function. The feature weight ai is finally weighted and summed to obtain a global feature representation Fm. We denote the original image as I_0, the sub-images as I_0, I_1, ..., I_k, and

the backbone network as $r(I^*;\theta)$. The feature set X of the image I* is defined as the formula (1):

$$X\,[F_0, F_1, ..., F_k] = [r(I_0;\theta), r(I_1;\theta), \cdot, r(I_k;\theta)] \quad (1)$$

It is sent to the global attention module to use the fully connected layer and the Sigmoid activation function to calculate the attention weight.

The meaning weight is expressed as formula (2):

$$a_i = f(F_i^T, q) \quad (2)$$

Since there is an important feature information area in an expression image, we set the attention loss function to limit the attention weight of the original image and the sub-image. The loss function requires that the attention weight from the sub-image, yi, should be greater than the weight of the original expression image. The attention loss function formula is shown in(3):

$$L = \max\{0, \partial - (\mu_{max} - \mu_0)\} \quad (3)$$

Among them is the hyperparameter, which is set to 0.03 in this paper, 0 is the attention weight of the original image, and max represents the maximum weight of all sub-images.

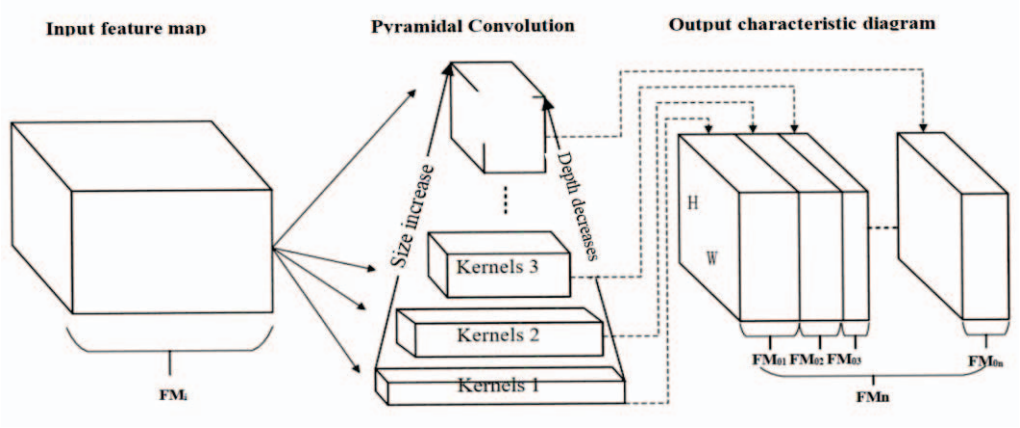

Fig.2 Pyramidal Convolution

IV. Experiments and Results

A. Pretreatment

Because the image sizes of different data sets are different, before training the model, the data needs to be preprocessed to adjust the size of all images to 224×224×3. Due to the differences in samples in different data sets, the convergence speed of the model on different data sets is different. In this experiment, the number of iterations on the CK+, RAF-DB, and AffectNet data sets are 100, 200, and 200, respectively. In the CK+ training process, the learning rate of the 40th and 80th rounds is attenuated with a decay rate of 0.9; in the training process of RAF-DB and AffectNet, the learning rate is attenuated with a decay rate of 0.9 every 50 rounds.

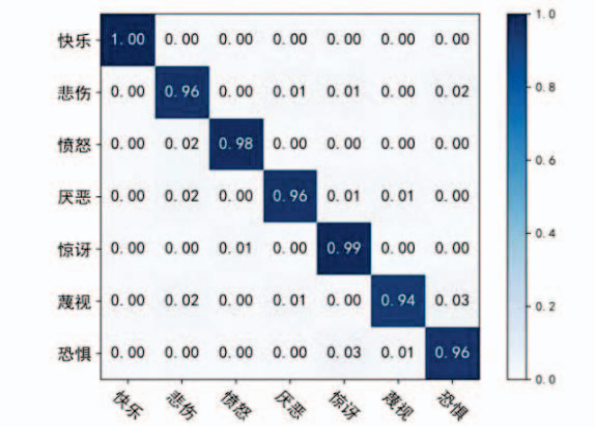

Fig.3 Confusion matrix of CK+

B. CK+ experiment

The confusion matrix of the proposed model on the CK+ data set is shown in Figure 5. For the six expressions of happiness, sadness, anger, disgust, surprise and fear, the recognition accuracy rate is 95%.

The expression "contempt" with the lowest recognition accuracy rate also reached 94%, and the recognition rate of the two expressions with obvious characteristics of "happy" and "surprise" reached more than 99%. The proposed model is compared with other mainstream methods on the CK+ data set, and the experimental results are shown in Table 1. It can be seen that the accuracy of the model proposed in this paper on the CK+ data set is 98.46%. Compared with the four methods Gabor [7], WLS-RF [8], PACNN [9] and SCAN [10], it is 6.27%, 4.16%, 1.43%, 1.15% higher, respectively.

TABLEI. Experimental Comparison Of CK+

Method	Accuracy(%)
Gabor	92.19
WLS-RF	94.3
pACNN	97.03
SCAN	97.31
PyConv-Attention Network	98.46

This method realizes facial expression recognition of feature maps with less overhead through channel dimension reduction and expansion convolution. The experimental results show that the method in this paper significantly improves the accuracy of facial expression recognition on the CK+ expression data set.This paper uses the residual network as the basic framework to design a facial expression recognition model that combines pyramid convolution and global attention.Pyramid convolution can learn multi-scale feature information and improve the nonlinear expression ability of the model; the attention mechanism can make the network pay more attention to important feature information and suppress noise interference. The proposed model has achieved 98.46% accuracy on the CK+ public expression data set.

REFERENCES

[1] AlNatour Ahlam,Gillespie Gordon Lee,Alzoubi Fatmeh. "We cannot stop smoking": Female university students' experiences and perceptions.[J]. Applied nursing research : ANR,2021,61:

[2] Yunxin Huang, Fei Chen, Shahe Lv, Xuedong Wang. Facial Expression Recognition: A Survey [J]. Symmetry, 2019,11(10).

[3] ZHANG A M, XU Y. Attention Hierarchical Bilinear Pooling Residual Network for Expression Recognition [J].Computer Engineering and Application, 2020, 56(23):161- 166.

[4] Gera D, Balasubramanian S. Landmark Guidance Independent Spatio-Channel Attention and Complementary Context Information based Facial Expression Recognition[J].Pattern Recognition Letters, 2021, 145:58-66.

[5] LI G H, YUAN Y F, Ben X Y, ZHANG J P. Spatiotemporal attention network for micro- expression recognition[J].Journal of Image and Gra-phics,2020,25(11):2380-2390.

[6] Duta I C, Liu L, Zhu F, et al.Pyramidal convolution: rethinking convolutional neural networks for visual recognition[EB/OL].(2020-6-20)[2021-5-9]

[7] Adil B, Nadjib K M, Yacine L. A novel approach for facial expression recognition[C]//2019 International Conference on Networking and Advanced Systems (ICNAS). IEEE, 2019: 1-5.

[8] Dapogny A,Bailly K,Dubuisson S. Confidence -weighted local expression predictions for occlusion handling in expression recognition and action unit detection[J]. International Journal of Computer Vision, 2018, 126(2): 255- 271.

[9] Li Y, Zeng J, Shan S, et al. Occlusion aware facial expression recognition using CNN with attention mechanism[J]. IEEE Transactions on Image Processing, 2018, 28(5): 2439-2450.

[10] Gera D, Balasubramanian S. Landmark Guidance Independent Spatio-Channel Attention and Complementary Context Information based Facial Expression Recognition[J]. Pattern Recognition Letters, 2021, 145:58-66.