International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 5, Number 2, Year 2025
DOI: https://doi.org/10.62051/ijcsit.v5n2.12
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Research on Video Emotion Recognition Method Based on Deep Learning

Hui Wang *

Southwest Minzu University, Chengdu, China
* Corresponding Author: Hui Wang

ABSTRACT

Video data has significant temporal and spatial characteristics, and the extraction and recognition of its emotional features is currently a research hotspot in the field of human-computer interaction. This paper proposes an improved model based on P3D 3D residual network (3Res att Network) to address the difficulties in dynamic feature extraction and insufficient spatiotemporal information fusion in video emotion recognition tasks. Firstly, the P3D infrastructure is constructed by decoupling spatiotemporal convolution kernels, effectively reducing model complexity; Secondly, design a 3D Spatial Attention mechanism to dynamically focus on emotionally significant areas and enhance feature discriminability; Finally, by combining residual connections with multi head attention mechanism, an end-to-end 3Res att network is constructed to achieve deep fusion of spatiotemporal features. Experiments on the eNTERFACE '05 dataset show that our method achieves an accuracy of 67.5% in video emotion recognition tasks, which is 18.2% higher than traditional C3D models and 11.4% higher than 3Res att2. The ablation experiments further validated the effectiveness of the attention module, providing a new technological path for video emotion computing.

KEYWORDS

Video Emotion Recognition Method; Deep Learning; P3D.

1. INTRODUCTION

Traditional facial emotion recognition methods rely on manual features such as LBP and Gabor, while deep learning methods significantly improve performance through networks such as VGG and ResNet. For example, Cheng et al. [1]optimized VGG19 and combined it with transfer learning to alleviate the problem of insufficient data; Lu Guanming [2]used residual units to balance the depth and convergence of the network; Zhong et al. [3] introduced SE module and simplified parameters in ResNet; Shahzad et al.[4] enhanced recognition robustness through facial feature localization.

The attention mechanism has further promoted the development of this field: Marrero et al. [5] used U-Net based attention masks to remove irrelevant features; Wang et al. [6] designed a regional attention network to cope with occlusion and pose changes; Gan et al. [7] proposed a multi attention fusion framework; Lan Lingqiang et al.[8] combined bottleneck attention with second-order pooling optimization feature distribution; Vats et al. [9] constructed an attention model based on Swin Transformer.

However, video emotion recognition still faces challenges such as low efficiency in extracting spatiotemporal features and a large number of model parameters. This article proposes an improved 3D convolutional network: firstly, dynamic keyframes are extracted through inter frame difference method; Secondly, based on the P3D concept, the spatiotemporal convolution is decoupled to

construct a lightweight 3D residual architecture; Further introduce a three-dimensional spatial attention mechanism to focus on emotionally significant regions. Experiments have shown that this method significantly improves recognition accuracy compared to benchmark models such as C3D, providing an efficient solution for video emotion computation.

2. METHOD

In the research of video emotion recognition, the research content can be divided into three parts: video data preprocessing, video emotion feature extraction, and video emotion classification.

2.1. Video Preprocessing

Keyframe extraction: Keyframes are frames that can reflect significant features of the video. Extracting keyframes can reduce computational complexity and improve emotion recognition accuracy. To preserve the temporal sequence of video expressions, this paper uses the inter frame difference method to extract keyframes. Due to the short videos in the eENTERFACE'05 dataset, the two frame difference method was used in the experiment to extract keyframes, as shown in Figure 1.

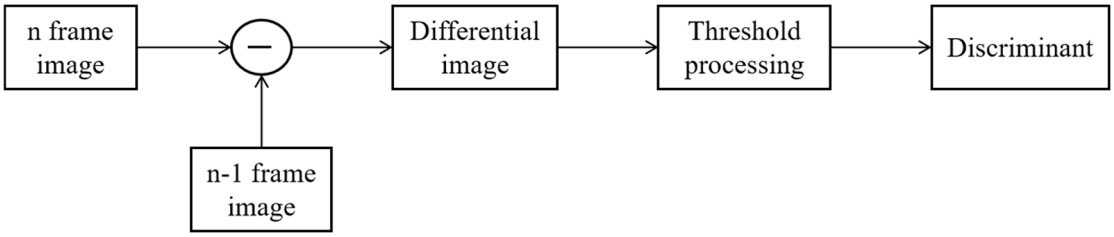


Figure 1. Difference diagram of two frames

Facial detection: The OpenCV library is the most commonly used in facial detection. The file haarcaclad_frontalface-alt.xml contains a trained Haar model that describes the feature values of various parts of the face. The Haar model uses the AdaBoost algorithm to train a strong classifier that distinguishes whether it is a face or not, achieving the task of face detection. The Haar model normalizes the detected facial images and saves them separately.

Facial alignment: The steps for facial alignment are shown in Figure 2. Use face_decognition to align the extracted facial images from the previous section. The face_decognition API first detects 68 keypoints of the face, and uses the center coordinates of the keypoints of the left and right eyes as a reference to rotate and align the face image. Save the aligned facial images one by one as a foundation for further dynamic expression recognition.

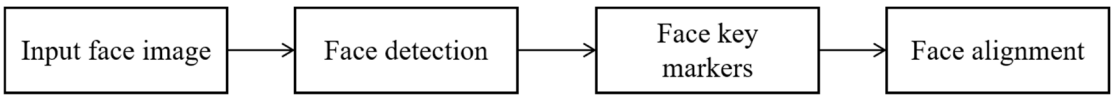


Figure 2. Steps for face alignment

2.2. Feature Extraction Network Architecture

The structure of the C3D network is shown in Figure 3, which consists of 8 3D convolutional layers with a kernel size of 3×3×3 and 5 3D pooling layers. Each convolutional layer is followed by a pooling layer. In order to avoid shortening the temporal length in advance, the size and stride of the first layer pooling kernel are 1×2×2. The final extracted feature map predicts labels through two fully

connected layers and a Softmax loss layer. The number of filters for the 8 convolutional layers is 64, 128, 256, 256, 512, 512, 512, and 512, respectively.

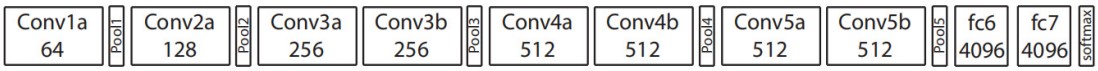


Figure 3. C3D network structure

Before training the C3D network, the video data in the eNTERFACE'05 dataset was preprocessed to extract 16 keyframe image data, resulting in a single video size of [16,128,128, 3]. After processing through eight convolutional layers and five pooling layers in the C3D network, the classifier outputs classification results for six types of emotional labels.

2.3. P3D 3D ResNet-18 Neural Network

To solve the problem of excessive parameters in 3D convolution, Zhaofan Qiu et al. proposed decoupling the traditional $3\times3\times3$ convolution kernel into a combination of $1\times3\times3$ (spatial convolution) and $3\times1\times1$ (temporal convolution). Among them, the $1\times3\times3$ convolution kernel is used to extract spatial features, and the $3\times1\times1$ convolution kernel is used to capture temporal features. Based on this, the author further designed three connection methods (P3D-A, P3D-B, P3D-C), corresponding to cascaded, parallel, and skip connection structures, respectively, to optimize the fusion efficiency of spatiotemporal features. This study adopts the first cascading combination method, which first performs spatial two-dimensional convolution on the feature map, then performs temporal convolution, and finally the result of temporal convolution is combined with the output of the residual block together with the shock. Finally, the Bottleneck structure of the residual network can be obtained as shown in the following figure:

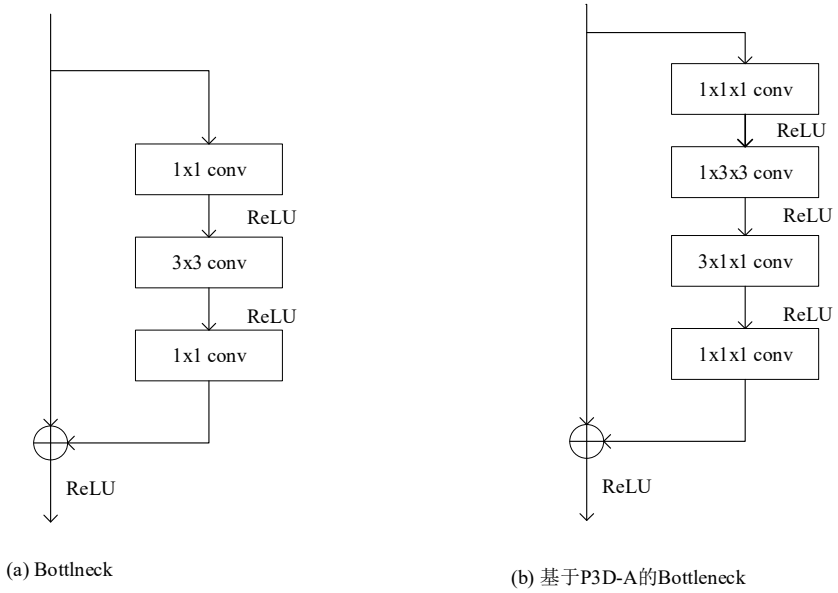


Figure 4. P3D-based jumper module

Based on this idea, we replace all 3×3 2D convolutional layers in the 2D ResNet-18 network with cascaded combinations. In the convolutional block structure of the ResNet-18 network, the stride of some convolutional layers is set to 2×2, while in the convolutional block structure of the 3D network, the stride is set to 2×2×2. After introducing the spatial and temporal convolutional layers in P3D, their stride is decoupled into a stride of 1×2×2 for the spatial convolutional layer and a stride of 2×1×1 for the temporal convolutional layer, achieving feature extraction in both spatial and temporal directions. The convolutional block structure is shown in Figure 5.

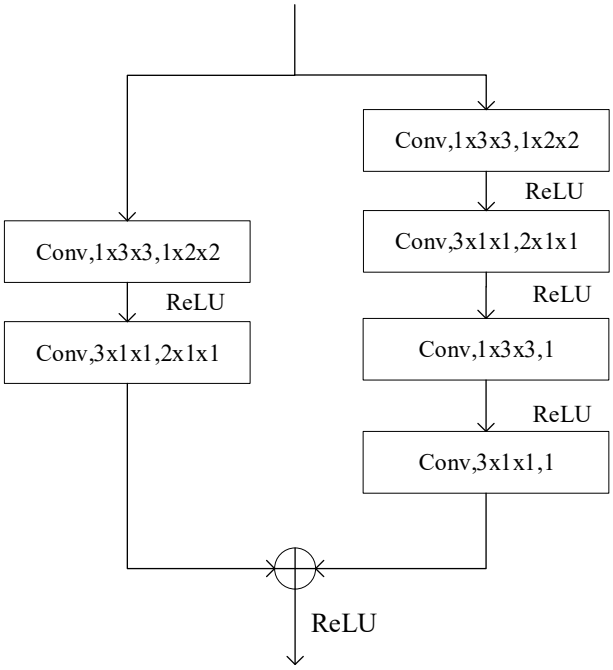


Figure 5. 3d convolutional block structure

In the 3D ResNet-18 network architecture based on P3D, the input layer consists of 64 convolutional kernels with a size of 1×7×7 and a stride of 1×2×2. Pass through the Max pooling layer again with a stride of 2 and a pooling kernel of 3×3×3. The specific network structure of 3D ResNet-18 is shown in Figure 6.

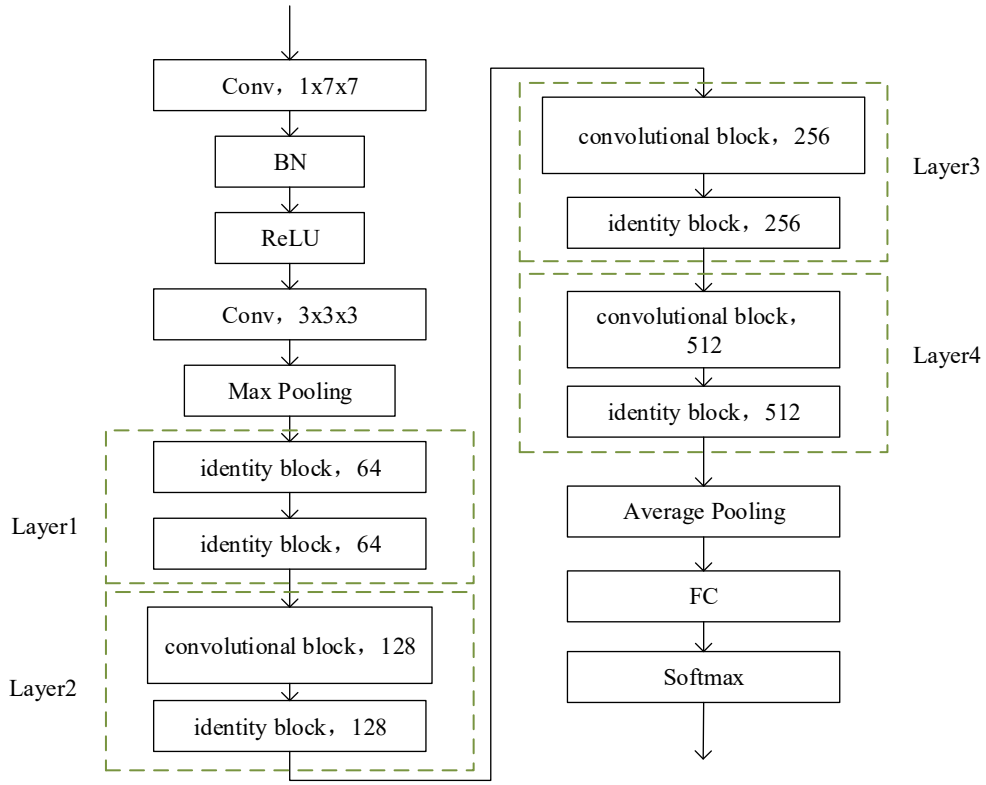


Figure 6. P3D-based ResNet-18 network

2.4. Res-att

To further extract emotional features from videos, this study proposes to make the spatial attention mechanism three-dimensional and introduce it into the residual blocks to construct a three-dimensional Res-A module. The aim is to train the network to focus more on emotional features and ignore redundant information.

The structure of spatial visual attention mechanism is shown in Figure 7. The spatial attention module first performs max pooling and average pooling operations on the input feature map by channel. Maximum pooling is to retain the maximum value of the pooling window. Average pooling is the process of preserving the overall data features of the pooling window and calculating the average value. Connect the two pooling results in parallel channels, express them as a whole, and then use 1×1 convolutional layers to extract features through convolution. Then, use an activation function to convert the obtained single-layer feature map into a probability distribution, and obtain the attention mechanism score for each element of the input features. The higher the attention score, the more valuable the corresponding element information is in recognizing emotional categories. Conversely, the lower the attention score, the less influential the corresponding element information is in recognizing emotional categories and can be ignored. Finally, the attention score matrix is dot multiplied with the input values to adjust the model's emphasis on different spatial positions of the input.

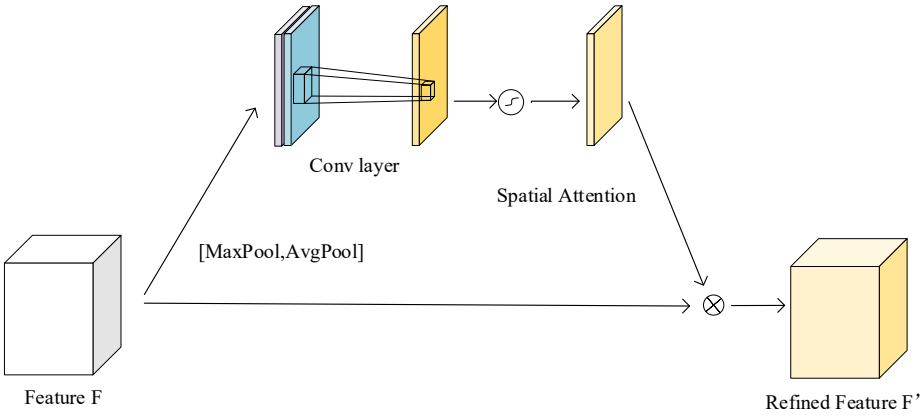


Figure 7. Spatial attention

Improve it into a three-dimensional spatial attention mechanism for easy application in three-dimensional convolutional networks. The 3D spatial attention module first performs 3D max pooling and 3D average pooling operations on the input feature map according to the channel dimension, and obtains the sum. Connect the two pooling results in parallel by channel, then use the 1×3×3 convolutional layer and 3×1×1 convolutional layer for feature extraction through convolution. Next, use the activation function to obtain the probability distribution and obtain the attention mechanism score for each element of the input feature. Finally, the attention score matrix is dot multiplied with the input values to adjust the model's emphasis on different spatial positions of the input.

Introduce the 3D attention mechanism into the residual network based on P3D, and construct the 3D Res-A module by incorporating spatial attention mechanism into the identity residual block, as shown in Figure 8 (b). Propose a 3Res att network model based on the 3D ResNet-18 model, with the model structure shown in Figure 8 (a). The input of the 3Res att network is 16 video frames, so after the fourth identity residual block and convolution residual block, the value of the time dimension decreases to 2. To further extract information in the time dimension, the data will also be processed through a convolutional layer with a stride of 1×2×2 and a three-dimensional convolutional layer with a stride of 1×3×3, a three-dimensional convolutional layer with a stride of 2×1×1 and a convolutional layer with a stride of 3×1×1, and a normalization layer. Finally, six types of video

sentiment classification are completed through a fully connected layer consisting of a global average layer and an activation function.

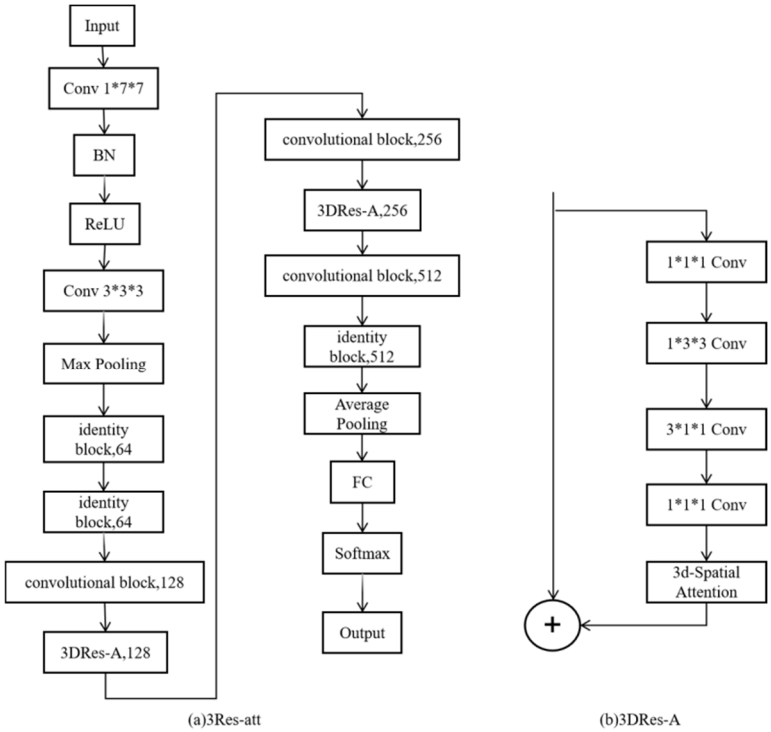


Figure 8. 3Res-att and 3d Res-A

3. EXPERIMENT AND RESULTS

3.1. Dataset and Experimental Setup

The eNTERFACE'05 dataset consists of 42 participants who listen to short stories and use them to trigger specific emotional recordings. Two human experts determine whether the specific emotional audio and video triggered clearly express the corresponding emotions. This dataset contains a total of 1166 sets of English audio and video data, with a data length of approximately 1-2 seconds. The eNTERFACE'05 dataset contains six emotions: anger, fear, happiness, disgust, sadness, and surprise.

This experiment used the Keras deep learning framework to build an emotion classification network model. Keras supports parallel computing and utilizes multiple GPUs to accelerate model training. Based on the Ubuntu operating system, train each model on two 16GB GeForce GTX 2080Ti.

The dataset is randomly divided into 80% training set and 20% testing set. The initial learning rate of the experimental setup is set to 0.0008, and the cross entropy function is used as the loss function. The BP algorithm and Adam optimizer continuously optimize the learning network parameters, with a total of 25 iterations of training.

3.2. Comparative Experiment

Build four network models: C3D network model, C3D+attention network model, 3D ResNet, and P3D based 3D ResNet network model. Through comparative experiments on four types of networks in video emotion classification tasks, analyze and explore the feasibility of P3D 3D ResNet network in this task.

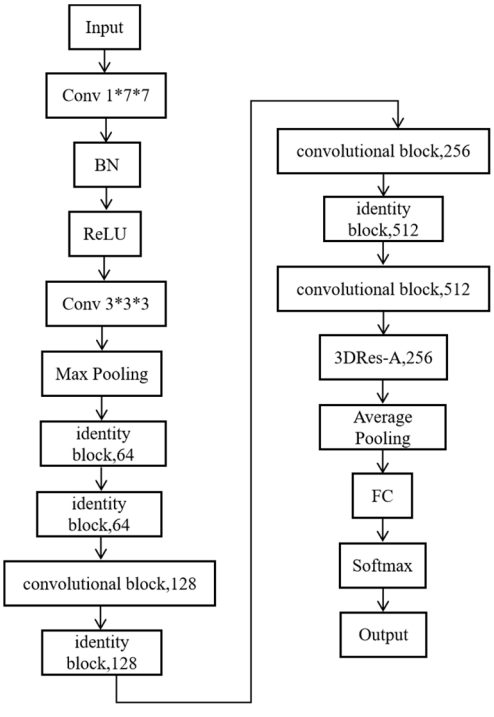
Table 1. C3D, 3D ResNet and P3D-based ResNet video sentiment classification

Experimental group	data set	network model	accuracy
1	eENTERFACE′05	C3D	49.3%
2		C3D+attention	52.3%
3		3D ResNet	55.4%
4		3D ResNet based on P3D	56.1%

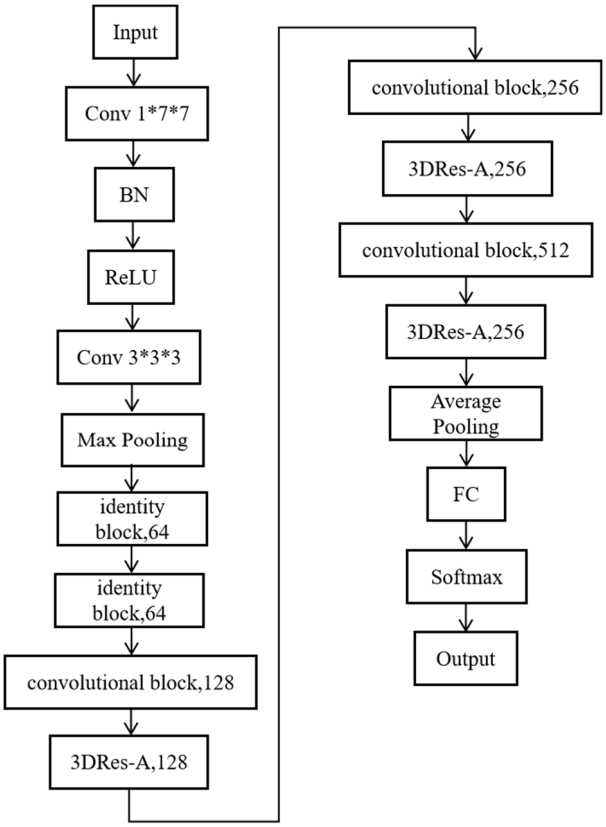
Table 1 shows the sentiment classification accuracy of C3D, 3D ResNet, P3D based 3D ResNet, and C3D+attention network on the eNTERFACE '05 dataset. The experimental results show that the C3D+attention network improves the accuracy by 3.0% compared to C3D, verifying the effectiveness of the 3D spatial attention mechanism in extracting emotional features from videos. In addition, 3D ResNet is significantly better than C3D, indicating that increasing network depth helps to learn deep spatiotemporal features. The 3D ResNet based on P3D reduces model parameters while improving by 0.5% compared to ordinary 3D ResNet, further verifying the feasibility of the P3D structure in video emotion recognition tasks.

3.3. Ablation Experiment

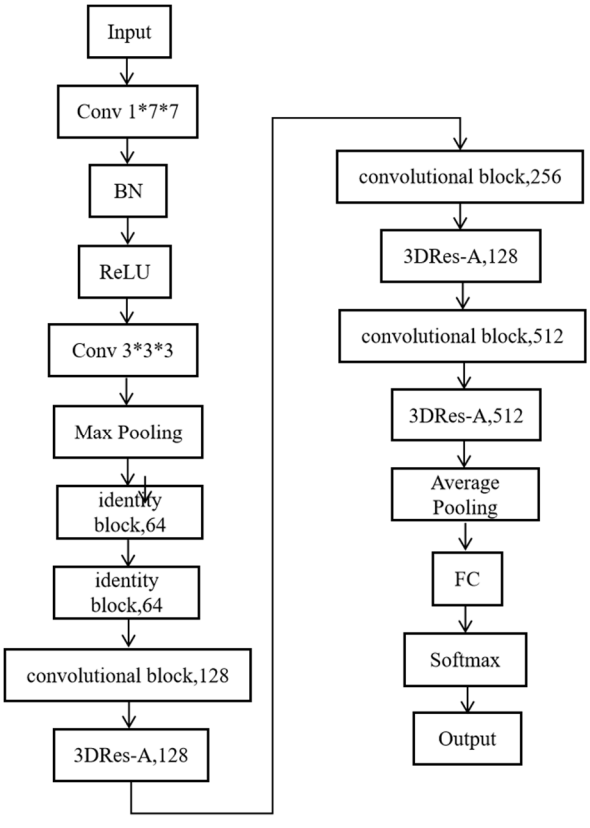
To evaluate the impact of the 3D Res-A module on model performance, this paper used the P3D based 3D ResNet as a benchmark and constructed network models incorporating 3Res-att1, 3Res-att2, and 3Res-att3 Res-A modules, as shown in Figure 9. The experimental results show that the recognition accuracy of 3Res-att1 and 3Res-att2 is improved by 7.6% and 11.4% respectively compared to the benchmark model, verifying the effectiveness of the Res-A module in video emotion classification tasks. However, the accuracy of 3Res-att3 decreased to 64.4%, indicating that excessive attention modules may lead to overfitting. Therefore, the number and location of Res-A modules have a significant impact on model performance, and the configuration of two-layer modules achieves the best balance between feature extraction and generalization ability.



(a) 3Res-att1



（b）3Res-att2



（c）3Res-att3

Figure 9. Network structure of 3Res-att1, 3Res-att2, and 3Res-att3

Table 2. Experimental results based on RES-A1 ablation experiment

Experimental group	data set	network model	accuracy
1	eENTERFACE′05	3D ResNet based on P3D	56.1%
2		3Res-att1	63.7%
3		3Res-att2	**67.5%**
4		3Res-att3	64.4%

Table 2 shows the test results of the Res-A module ablation experiment. The experiment showed that the addition of Res-A module significantly improved the model performance: 3Res-att1 and 3Res-att2 increased by 7.6% and 11.4% respectively compared to the benchmark model, verifying the effectiveness of Res-A module in video emotion feature extraction. However, when the number of Res-A modules increased to 3 layers, the accuracy of 3Res-att3 decreased to 64.4%, indicating that excessive attention mechanisms may lead to overfitting. Therefore, this article chooses the 3-layer Res-A module's 3Res-att2 as the backbone model, achieving the best balance between feature extraction and generalization ability.

4. CONCLUSION

Through experiments, the superiority of the 3D ResNet network model based on P3D has been verified, effectively reducing the model size while improving the network recognition rate, with an accuracy improvement of 0.5% compared to ordinary 3D ResNet networks. The Res-A module based on identity residual blocks is extended to three-dimensional convolution, and the spatial attention mechanism is upgraded to enable its application to three-dimensional convolutional networks. The effectiveness of the Res-A module was verified through ablation experiments, and the experimental results showed that the Res-A module can effectively improve the video emotion recognition rate. However, the model performed the best when the number of Res-A modules was 2.

REFERENCES

[1] CHENG S,ZHOU G H.Facial expression recognition method based on improved VGG convolutional neural netw ork[J].International Journal of Pattern Recognition and Artificial Intelligence,2020,34(7):2056003.

[2] Lu Guanming, Zhu Hairui, Hao Qiang, etc Facial expression recognition based on deep residual network [J]. Data acquisition and processing,2019,34(1):50-57.

[3] ZHONG Y,QIU S,LUO X,et al.Facial expression recognition based on optimized ResNet[C]//Proceedings of 2020 2ndWorld Symposium on Artificial Intelligence(WSAI).IEEE,2020:84-91.

[4] SHAHZAD T,IQBAL K,KHAN M A,et al.Role of zoning in facial expression using deep learning[J].IEEE Access,2023,11:16493-16508.

[5] MARRERO F P D,GUERRERO P F A,REN T,et al.Feratt:Facial expression recognition w ith attention net[J].arXiv preprint arXiv:1902.03284,2019.

[6] WANG K,PENG X,YANG J,et al.Region attention networks for pose and occlusion robust facial expression recognition[J].IEEE Transactions on Image Processing,2020,29:4057-4069.

[7] GAN Y,CHEN J,YANG Z,et al.Multiple attention network for facial expression recognition[J].IEEE Access, 2020, 8:7383-7393.

[8] Lan Lingqiang, Liu Qiyuan, Lu Shuhua Facial expression recognition based on attention mechanism and feature correlation [J]. Journal of Beihang University, 2022, 48 (1): 147-155.

[9] VATS A,CHADHA A.Facial emotion recognition[J].arXivpreprint arXiv:2301. 10906, 2023.