J Inf Process Syst, Vol.17, No.3, pp.556~570, June 2021
https://doi.org/10.3745/JIPS.02.0156

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

A Video Expression Recognition Method Based on Multi-mode Convolution Neural Network and Multiplicative Feature Fusion

Qun Ren*

Abstract

The existing video expression recognition methods mainly focus on the spatial feature extraction of video expression images, but tend to ignore the dynamic features of video sequences. To solve this problem, a multi-mode convolution neural network method is proposed to effectively improve the performance of facial expression recognition in video. Firstly, OpenFace 2.0 is used to detect face images in video, and two deep convolution neural networks are used to extract spatiotemporal expression features. Furthermore, spatial convolution neural network is used to extract the spatial information features of each static expression image, and the dynamic information feature is extracted from the optical flow information of multiple expression images based on temporal convolution neural network. Then, the spatiotemporal features learned by the two deep convolution neural networks are fused by multiplication. Finally, the fused features are input into support vector machine to realize the facial expression classification. Experimental results show that the recognition accuracy of the proposed method can reach 64.57% and 60.89%, respectively on RML and Baum-ls datasets. It is better than that of other contrast methods.

Keywords

Facial Expression Recognition, Multi-Mode Deep Learning, Multiplicative Fusion, Optical Flow Method, Spatial Convolutional Neural Network, Time Convolutional Neural Network

1. Introduction

Facial expression recognition is a biological feature recognition technology that expresses and analyzes human facial expressions and enables computers to recognize and even understand human emotions [1-5]. The research has important application value in the fields of human-computer interaction system, intelligent video surveillance system, and so on [6,7].

The expression recognition system based on video is relatively difficult because dynamic sequence image is not easy to process compared with single image [8]. Video expression recognition usually includes three parts: video preprocessing, expression feature extraction, and expression classification. Video preprocessing mainly detects and extracts human faces from the sequence images in the video [9]. Feature extraction refers to extracting features from facial images in the video that can describe facial expressions. Expression classification is to recognize expressions by inputting these extracted features

Manuscript received May 14, 2020; first revision September 4, 2020; accepted October 4, 2020.

Corresponding Author: Qun Ren (bzrenqun159@163.com)
* Dept. of Electronic and Information Engineering, Bo Zhou University, Bo Zhou, China (bzrenqun159@163.com)

into a classifier [10,11].

To realize video expression recognition, a multi-mode feature fusion multi-mode convolutional neural network (MFFCNN) is proposed. The main innovations are summarized as follows:

1) A multi-mode convolutional neural network is proposed to compress a continuous optical flow sequence to a single ordered optical flow graph and learn the temporal and spatial expression features with discriminative power.
2) Multiplicative feature fusion is proposed. The learned spatial and temporal expression features are fused by compressing the excitation residual block, so that the spatial and temporal information at the feature level can be further learned to better distinguish similar-looking expressions.

2. Related Work

To realize video expression recognition, scholars have proposed many methods. For example, An and Liu [12] proposed an adaptive model parameter initialization method of linear activation function of multi-layer max out network. The method initializes a convolutional neural network (CNN) and a long short-term memory (LSTM) to solve the problems of gradient disappearance and gradient explosion. The authors of [13,14] proposed a facial expression recognition system using a multi-channel deep learning framework and a method of facial expression recognition based on a mixed square diagonal pattern geometric model to improve model efficiency by extracting features of greater importance. However, these methods ignore the edge features of the image.

Li et al. [15] proposed a new face cropping and rotation strategy, which improves the model recognition rate by extracting useful facial features. Several studies [16-18] proposed independent recognition method of micro-expressions based on Wasserstein generative adversarial network, a multi-level uncorrelated DS-GPLVM model (ML-UDSGPLVM) and facial emotion recognition based on multi-layer CNN to improve accuracy and robustness of facial expression recognition. However, the run time of these methods is longer.

Jain et al. [19] proposed emotion recognition based on multi-component deep forest and transfer CNN. Features are extracted through knowledge transfer to improve the accuracy of the model. Kong [20] proposed a facial expression recognition method combining deep CNN and improved LBP feature, which improves the recognition accuracy to a certain extent. However, when the dataset has a lot of noise points, these methods have poor recognition results. Salmam et al. [21] proposed a model of multi-layer deep neural network, which improves the recognition rate of deep learning model. The authors of [22] and [23] proposed a dual-mode spatial and temporal feature representation learning method and an improved complete local ternary pattern (ICTP), respectively, which improve the model recognition rate through feature fusion. However, when the amount of data is large, these methods are prone to overfitting.

Nigam et al. [24] proposed a facial expression recognition method based on directional histogram of wavelet domain. The discrete domain wavelet transform is used to transform the features from the spatial domain to the frequency domain and extract features. It shows better performance. Zia et al. [25] proposed a classifier-based dynamic weighted majority voting mechanism (DWMV). By enhancing the learning ability, all possible expression patterns of features can be learned to improve the model recognition rate. Wang et al. [26] proposed a classifier iterative fusion method based on multi-directional gradient computing features and deep learning features. The model recognition rate is improved by obtaining diagonal gradient information and combining the extracted features. However, these methods ignore the dynamic changes of the human face.

Based on the above analysis, it is clear that deep learning has better modeling and processing capabilities for facial expression images in videos. Many existing deep learning methods focus on the spatial feature extraction of video expression images, but tend to ignore the dynamic changing features of the video sequence. To make full use of the optical flow information, a video facial expression recognition method based on deep CNN is studied in this paper. It mainly takes two steps: firstly, the spatial CNN and temporal CNN are used to extract high-level spatial and temporal features from the video, respectively. On this basis, the spatial and temporal features are multiplicative fused.

3. Multi-mode Convolutional Neural Network Model

3.1 Overall Architecture of the Proposed Method

The proposed video expression recognition model framework based on multi-mode deep CNN is shown in Fig. 1. The model consists of three steps: video preprocessing, deep spatiotemporal expression feature extraction, and fusion expression classification. Temporal CNN mainly processes video optical flow signals and extract high-level temporal features. The spatial CNN mainly processes the face image of each frame in the video and extracts high-level spatial features. Then, the extracted temporal features and spatial features are subjected to average pooling, respectively, and the spatial-temporal feature fusion based on DBN is performed on the feature layer. Finally, the fused spatial-temporal features are classified by using support vector machine (SVM) to complete the video expression classification.

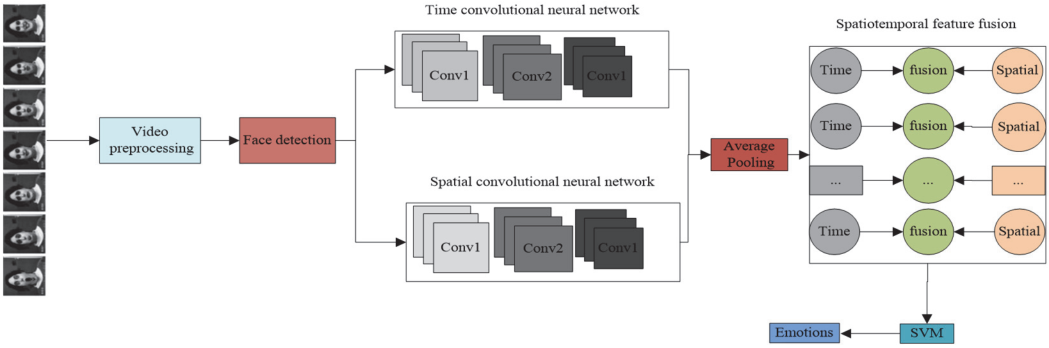


Fig. 1. Video expression recognition model based on multi-mode deep convolutional neural network.

3.2 Video Preprocessing

Assume that L represents the number of frames contained in the video clip. A suitable size L is chosen to ensure the extraction of the temporal information features of the video. If L is too small, the video segment contains insufficient dynamic change information. Conversely, if L is too large, the video clip may contain too much noise and affect the recognition performance.

By performing a sequential search within the range $[2, 20]$ of L, it is found that when $L = 16$, the temporal CNN achieves the best results. Therefore, the paper divides each video into 16-frame-sized fragments. When $L > 16$, this paper discards the front and back $(L - 16)/2$ frames of the video. When $L < 16$, the front and back $(L - 16)/2$ frames of the video are copied. For a video clip with $L = 16$, 15 frames of optical flow images are included, because every two adjacent spatial images will generate one frame of optical flow image. The optical flow image represents the displacement information of the corresponding positions of two adjacent frames. The specific calculation process is as follows: Let two adjacent frames be t and $t + 1$ in the video and the displacement vector d_t represent the displacement information of the video. The optical flow image I_t is composed of $d_t x$ and $d_t y$. $d_t x$ and $d_t y$ are two channels of I_t, which represent the horizontal displacement component and the vertical displacement component of two adjacent image positions in the video, respectively. Considering that the input of $DCNN_S$ is RGB image with three channels, the amplitude component dt_z of the optical flow image I_t is calculated as the third channel of I_t:

$$dt_z = \sqrt{d_t^2 x + d_t^2 y} \tag{1}$$

For the preprocessing of the input images of spatial convolutional neural network, the face detection algorithm proposed in reference [27] is used to extract the face images contained in each frame of the video segment in real time.

3.3 Spatial Feature Extraction

Using full convolutional network to extract the spatial information features of each frame of static expression images in the video and identify the corresponding expressions, a simple convolutional network structure can be designed, as shown in Fig. 2.

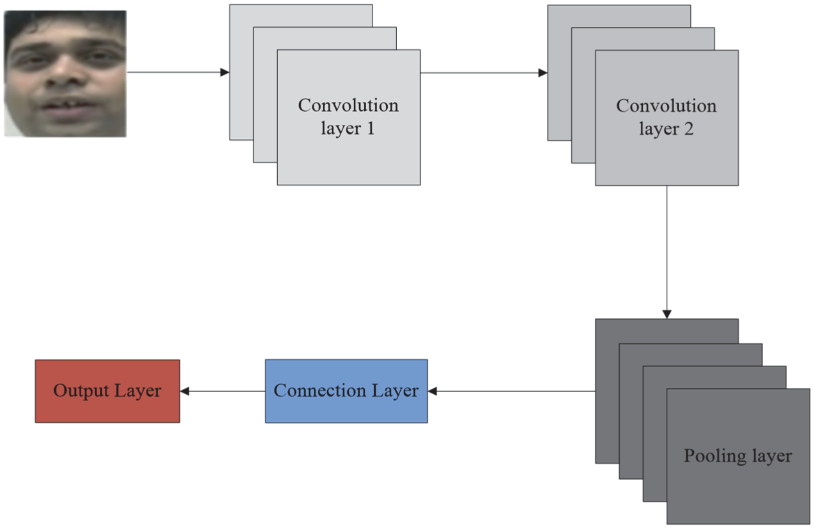


Fig. 2. Convolutional neural network structure for processing a single frame.

The network contains two convolutional layers, a down-sampling layer, a fully connected layer and an output layer. The output layer function is the activation function. Here, the softmax function is selected and expressed as follows.

$$\sigma(z_j) = \exp(z_j) / \sum_{k=1}^{K} \exp(z_k) \tag{2}$$

Its essence is to map an arbitrary real vector of K dimension to a real vector of another K dimension and each element in the vector takes a value between $(0,1)$, where $j = 1,2,\ldots,k$. The z of Eq. (2) represents the output of the fully connected layer, that is, the input of the output layer. The output of the neural network is converted into a probability vector and it can be seen that the magnitude of classification probabilities of different classes through this vector. This layer contains 6 neurons, which means that 6 different expressions are classified. The input of the network is a picture of 128×128 pixel, and the output is a 6-dimensional probability vector. Each dimension represents the probability of expression classification. In the field of image recognition, this structure has been proven to achieve very good results, but it also has limitations. It can only take a single frame of picture as input. Since facial expression is a continuous and dynamic process, if only one frame is used as input, the inter-frame information will be lost. It reflects, therefore, the changing process of the expression is an important classification basis.

3.4 Temporal Feature Extraction

Under the condition of preserving the order information, the continuous optical flow sequence is compressed into a single ordered optical flow map, and the long-term time-domain structure of the video is modeled [28,29]. A CNN containing motion flow is designed. The motion information features of the video are extracted using the ordered optical flow map as input.

Given n frame of continuous optical flow sequence $F = [f_1, f_2, \ldots, f_n]$, where $f_i \in R^{d_1 \times d_2 \times d_3}$, and d_1 and d_2 are the height and width of the optical flow graph respectively. Each frame of the optical flow image is a horizontal and vertical component of two-channel image corresponding to the optical flow, expressed as f_i^x, f_i^y. The corresponding weighted moving average graph of the t frame of the optical flow graph f_t can be defined as:

$$\hat{f} = \sum_{i=1}^{t} \frac{i}{\sum_{j=1}^{t} j} f_i \tag{3}$$

The weighted average method of Formula (3) can reduce the error rate of optical flow estimation and the influence of white noise at the same time. The ordered optical flow diagram on the weighted moving average diagram of the optical flow sequence can be calculated as follows:

$$\min_{G \in R^{d_1 \times d_2 \times 2}, \xi \geq 0} \| G \|^2 + C \sum_{i<j} \xi_{ij} \tag{4}$$

$$s.t. < G, \hat{f}_j > - < G, \hat{f}_i > \geq 1 - \xi_{ij}, \forall i < j \tag{5}$$

where $<.,.>$ is the inner product and C is the compromise parameter between the size of the boundary and the training error, and ξ_{ij} is the relaxation variable. Eq. (5) retains the order information of the optical flow frame. The parameter $G \in R^{d_1 \times d_2 \times d_3}$ obtained by training is used to represent the optical flow sequence. In fact, it is of the same size as the optical flow map, so G is defined as an ordered optical flow map. Eqs. (4) and (5) are equivalent to the unconstrained optimization problem, that is:

$$\min_{G \in R^{d_1 \times d_2 \times 2}} \sum_{i<j} [1 - < G, f_j > + < G, f_i >]_+ + \lambda \| G \|^2 \tag{6}$$

where $[.]_+$ is the function max $(0, x)$ and $\lambda = 1/C$.

It should be noted that the two channels of the optical flow diagram are the channels of the velocity vector, both of which describe the motion vector of the position of each pixel. Therefore, the two are related. The two channels are de-correlated by diagonalizing the matrix. Assume that G_x and $G_y \in R^{d_1 \times d_2}$ are two channels of the horizontal and vertical components of optical flow of ordered optical flow diagram G, respectively, and then Eqs. (4) and (5) can be converted as follows.

$$\min_{G_x, G_y \in R^{d_1 \times d_2 \times 2}, \xi \geq 0} \| G \|^2 + \| G_y \| + C \sum_{i<j} \xi_{ij} \tag{7}$$

$$\begin{aligned} & s.t. < G_x, \hat{f}_j^x > + < G_y, \hat{f}_j^y > - < G_x, \hat{f}_i^x >, \\ & < G_y, \hat{f}_j^y > \geq 1 - \xi_{ij}, \forall i < j \end{aligned} \tag{8}$$

The obtained G_x and G_y are transformed into the range $[0, 255]$ to generate an ordered optical flow map using the minimum-maximum normalization, which is further adopted as the input of the deep network.

When generating the ordered optical flow diagram, several ordered diagrams are generated for each behavioral video to avoid too many compressed optical flow frames to cause information loss. An optical flow sequence is firstly divided into several sub-sequences in unit of w frames in the time dimension with an interval of $w/2$. Then, an ordered optical flow map is created on each subsequence with an input of VGG-16. The input image size is adjusted to 224×224. The $fc6$ layer responses of all ordered optical flow graphs are averaged and L2 normalized to obtain VGG features.

To avoid overfitting due to insufficient labeled samples when training deep network and thus reducing the generalization ability of the network, two strategies are used to enhance the data of the long-term motion stream by 10 times: corner clipping and scale jitter.

3.5 Spatial-temporal Feature Fusion

To better understand the characteristics of the spatial-temporal flow network, the space compression excitation residual network and the time compression excitation residual network [30] are multiplied. The two compressed excitation residual blocks are connected bidirectionally, and the multiplication fusion is shown in Fig. 3. The outputs of the compressed excitation residual block of the time stream and the spatial stream are processed with element-level multiplication, so that the information of the residual unit through the spatial stream is adjusted by the time signal. Similarly, the outputs of the compressed excitation residual block of the spatial stream and the time stream are multiplied, and the information of the temporal stream is adjusted by the spatial signal. Through the multiplication fusion of time flow and space flow, the spatial-temporal information at the feature level can be learned, which is conducive to distinguishing similar-looking expressions.

For multiple models generated by different strategies, direct average method and weighted average method are used for integration. Suppose there are N models to be integrated. For the test sample D, the test result is N vector of M dimensions $q_1, q_2, \cdots q_N$ (M is the size of the labeled space of the data). The calculation formulas corresponding to the direct averaging method and the weighted averaging method are shown in Formula (8) and Formula (9), respectively.

$$Score = \frac{\sum_{i=1}^{N} q_i}{N} \tag{9}$$

$$Score = \frac{\sum_{i=1}^{N} w_i q_i}{N} \tag{10}$$

where w_i corresponds to the weight of the *i* model, $w_i \geq 0$ and $\sum_{i=1}^{N} w_i = 1$.

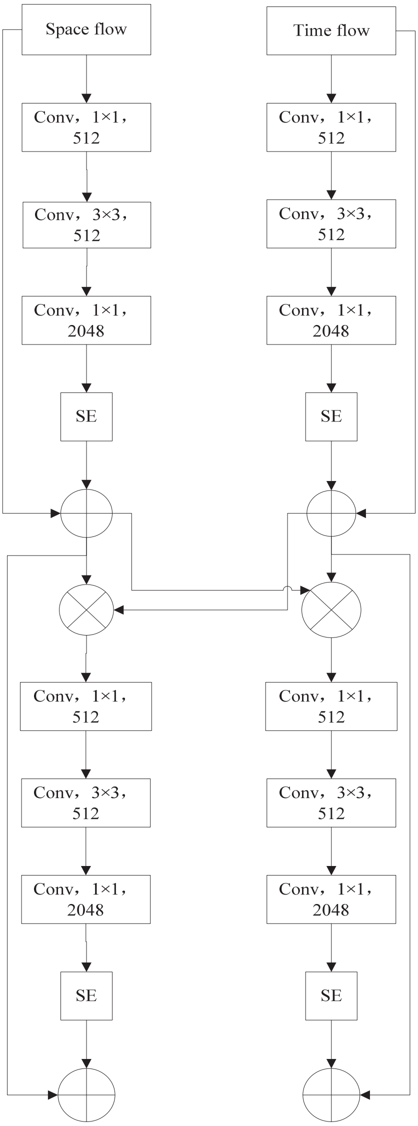


Fig. 3. Schematic diagram of multiplication fusion.

3.6 Network Training

The experiment uses cross entropy function (CEF) as the loss function of the training network.

$$C(\theta) = -\frac{1}{n}\sum[y\ln a + (1-y)\ln(1-a)] \tag{11}$$

where θ is the training parameter of the neural network, y is the expected output, a is the actual output of the neuron, and n is the number of training samples. The purpose of neural network training is to optimize this loss function. In the training process, the momentum algorithm is selected for training, and the formula is as follows:

$$\Delta\theta_t = \rho\Delta\theta_{t-1} - \eta g_t \tag{12}$$

ρ is the attenuation coefficient, indicating the degree to which the original update direction is retained. Its value is between $(0, 1)$. η represents the learning rate, and the value range is between $(0.01\sim0.1)$.

The Adadelta algorithm is selected for training, and its effect is significant. The formula is as follows:

$$E[g^2]_t = \rho E[g^2]_{t-1} + (1-\rho)g_t^2 \tag{13}$$

$$\Delta\theta_t = -\eta/(E[g^2]_t + \varepsilon)^{1/2} \tag{14}$$

where g_t is the gradient of x at time t, ρ is the attenuation coefficient, and ε is the error. Through the attenuation coefficient ρ, g_t of each moment decays with ρ, which is equivalent to only using the information of g_t closer to the current moment. Therefore, the parameters can still be updated after a long time.

4. Experimental Results and Analysis

To verify the effectiveness of the proposed multi-mode CNN and multiplication feature fusion video expression recognition method, a sufficient experimental evaluation is performed on the RML and BAUM-ls video sentiment datasets. The comparison and analysis are based on different feature fusion methods, times and positions, and the impact of parameter changes on recognition performance is analyzed. The confusion matrix on the two data sets is given and the proposed method is compared with several other new facial expression recognition methods. These methods are implemented in Python 3.0 using the image processing toolbox.

4.1 Experimental Dataset

The RML dataset has 720 videos consisting of the expressions of 8 people. There are 6 expressions on this dataset, namely angry, disgusted, scared, happy, sad and surprised. The average length of each video sample is about 5 seconds. The size of each image in the video is 720×480×3.

The BAUM-ls dataset has 1,222 videos consisting of the expressions of 31 people. There are 8 expressions on this dataset. This paper only studies 6 of them, namely angry, disgusted, afraid, happy, sad and surprised. And they are obtained from 521 video samples. The size of each image in the video is 720×480×3.

4.2 Effect of the Way, Frequency and Position of Feature Fusion on Recognition Performance

By using the same multiplication fusion method, that is, a multiplication fusion method from time flow to space flow, the effects of fusion number and position on recognition performance are experimentally analyzed. The experimental results are shown in Table 1. The recognition accuracy obtained after training on RML and BAUM dataset are reported. "C_2_1_R+C_2_1" indicates that the C_2_1 layer of the time stream is connected to the C_2_1_R layer of the spatial stream for multiplication fusion, and so on.

Table 1. Recognition accuracy of fusion from time stream to space stream at different times and positions on RML and BAUM-ls datasets

Fusion times	Fusion position	RML (%)	BAUM-ls (%)
1	C_2_1_ R+C_2_1	62.67	56.33
	C_3_1_ R+C_3_1	61.45	55.14
	C_4_1_ R+C_4_1	64.04	58.63
	C_5_1_ R+C_5_1	63.21	58.32
2	C_4_1_ R+C_4_1; C_5_1_ R+C_5_1	65.02	60.21

3	C_3_1_ R+C_3_1;
	C_4_1_ R+C_4_1;
	C_5_1_ R+C_5_1
		64.78	59.33

4	C_2_1_ R+C_2_1;
	C_3_1_ R+C_3_1;
	C_4_1_ R+C_4_1;
	C_5_1_ R+C_5_1
		63.81	57.43

As can be seen from Table 1, for single fusion, the features learned by the higher convolutional layers are complete and discriminative. In "single fusion, two fusions, three fusions, and four fusions," the fusion of different times shows that the underlying convolution layer fusion learns more shallow features such as color and edges and has not learned the discriminative semantic features of the upper layer. The fusion of the bottom layer convolution layer and other relatively high-level convolution layers reduces the recognition accuracy to a certain extent.

To explore the specific impact of different fusion methods on recognition performance, further experimental analysis is performed to set the fusion method to multiplication fusion from space stream to time stream under different fusion times and positions. The recognition accuracy rates on RML and BAUM-ls datasets are shown in Table 2.

The experimental results in Table 2 show that the fusion effect using "time flow to space flow" is better than that using "space flow to time flow." When using the fusion method of "space flow to time flow," with the number of fusions increasing, the recognition accuracy of the two, three, and four fusions gradually decreases. Compared to spatial stream network, time stream network has a stronger learning ability and the learned features are more discriminative. Injecting the characteristics of the spatial stream network with relatively weak learning feature capabilities into the temporal stream network will, feature learning of time flow network will be interfered to a certain extent. With the increase of the number of fusions, it may have a negative impact, resulting in the decrease of recognition rate.

Table 2. Recognition accuracy of different fusion methods on RML and BAUM-ls datasets

Fusion times	Fusion position	Fusion method	RML (%)	BAUM-ls (%)
1	C_2_1_R+C_2_1	T→S	62.67	56.33
		S→T	61.35	55.47
	C_5_1_ R+C_5_1	T→S	63.21	58.32
		S→T	62.03	57.34
2	C_4_1_ R+C_4_1, C_5_1_ R+C_5_1	T→S	65.02	60.21
		S→T	64.57	58.35
3	C_3_1_ R+C_3_1, C_4_1_ R+C_4_1, C_5_1_ R+C_5_1	T→S	64.78	59.33
		S→T	64.07	58.67
4	C_2_1_ R+C_2_1, C_3_1_ R+C_3_1, C_4_1_ R+C_4_1, C_5_1_ R+C_5_1	T→S	63.81	57.43
		S→T	62.73	56.82

4.3 Performance Analysis of Parameters

Without loss of generality, the number of feature maps of the first layer of the convolutional layer is set to 8 and the number of feature maps of the second layer of the convolutional layer is set to 16. The number of neurons in the fully connected layer is set to 64. A three-frame CNN structure is used, and the fusion layer is connected using four different convolution kernel sizes. Experiments are performed on the RML and BAUM-ls datasets. The experimental results are shown in Table 3.

Table 3. Recognition rates of different convolution kernel sizes on RML and BAUM-ls datasets

Case#	Convolution layer		RML (%)	BAUM-ls (%)
	First layer	Second layer		
1	5×5	3×3	63.52	59.32
2	11×11	9×9	64.66	60.81
3	17×17	15×15	65.36	62.34
4	35×35	33×33	63.21	61.03

It can be seen from Table 3 that with the increase of the convolution kernel, the recognition rate gradually increases, but there will be a threshold. If the convolution kernel is too large, it will reduce the recognition rate significantly.

When calculating the ordered optical flow diagram, the optical flow sequence of the behavioral video is first divided into several sub-sequences in the unit of w frames, and then the diagram is calculated on each sub-sequence. If the number of sub-sequence frames is too small, it will not achieve the purpose of modeling a long-term time-domain structure. If it is too large, some motion information will lose. Therefore, a reasonable sub-sequence length needs to be determined. Fig. 4 shows the recognition results of different sub-sequence lengths w on two datasets when using long-term motion flow alone for behavior recognition.

It can be seen from Fig. 4 that when w is 28, the recognition results obtained on the RML and BAUM-ls datasets are relatively high. Therefore, the length of the subsequence in the experiment is 28 frames.

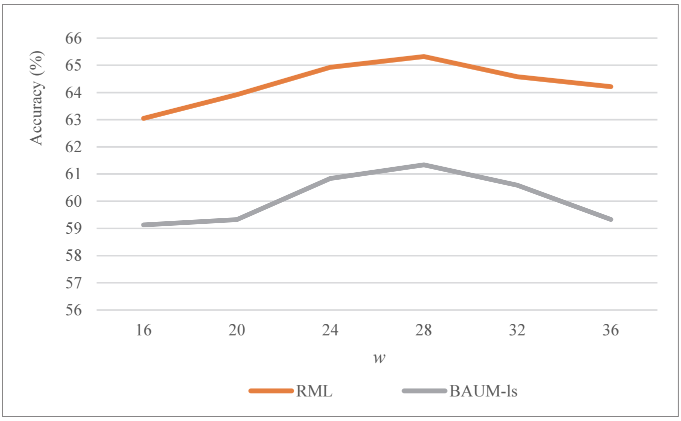


Fig. 4. Recognition accuracy of different sub-sequence lengths on RML and BAUM-ls datasets.

During network training, in order to verify the value of attenuation coefficient ρ and learning rate η, experiments are carried on RML and BAUM-ls datasets. In the experiments, the value range of attenuation coefficient is 0.1–1, and the value range of learning rate is 0.01–0.1. The experimental results are shown in Figs. 5 and 6.

As can be seen from Figs. 5 and 6, when the attenuation coefficient and learning rate are 0.2 and 0.02, respectively, the recognition rate on the RML and BAUM-ls datasets can reach the highest value. Therefore, in the next experiment, the attenuation coefficient is set to 0.2, and the learning rate is set to 0.02.

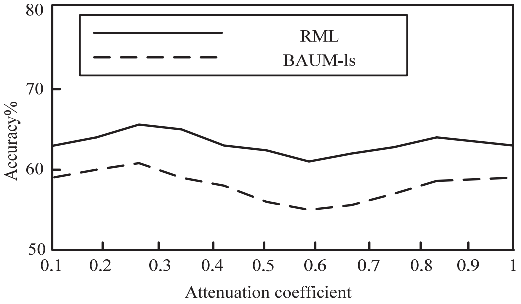


Fig. 5. Attenuation coefficient value analysis.

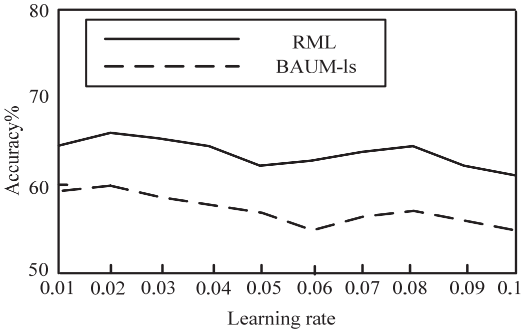
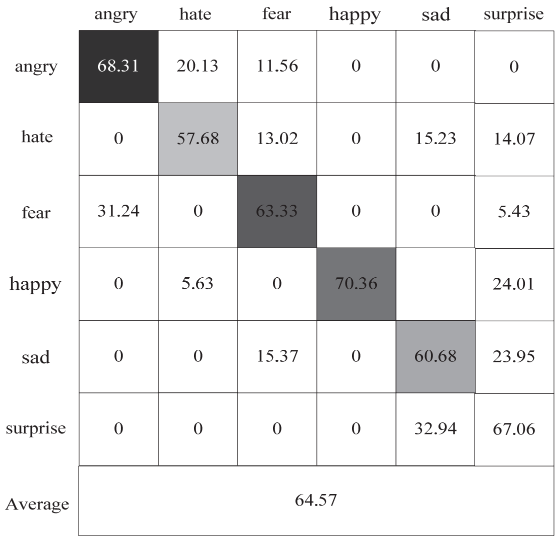


Fig. 6. Learning rate value analysis.

4.4 Confusion Matrix

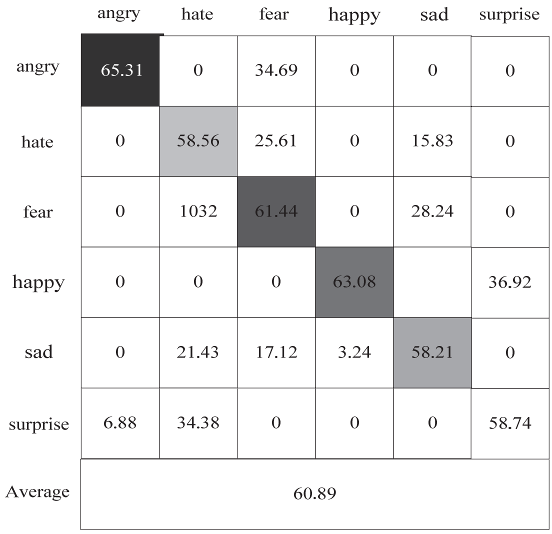
Figs. 7 and 8 show the fuzzy matrices to illustrate the recognition of each expression using MFFCNN when the spatial-temporal features are fused on the RML and BALUM-ls datasets to obtain the best performance. It indicates that MFCNN can learn the spatial-temporal features with discriminative power better and thus improve the accuracy of the model.

It can be seen from Figs. 7 and 8 that in the RML dataset, it is difficult to identify aversion and in the BALUM-ls dataset, it is difficult to identify sadness. The correct recognition rates are 57.68% and 58.21%, respectively. The reason is the characteristics of these expressions are similar to those of other expressions, causing confusion in recognition.



	angry	hate	fear	happy	sad	surprise
angry	68.31	20.13	11.56	0	0	0
hate	0	57.68	13.02	0	15.23	14.07
fear	31.24	0	63.33	0	0	5.43
happy	0	5.63	0	70.36		24.01
sad	0	0	15.37	0	60.68	23.95
surprise	0	0	0	0	32.94	67.06
Average	64.57					

Fig. 7. Confusion matrix diagram of RML dataset using MFFCNN.



	angry	hate	fear	happy	sad	surprise
angry	65.31	0	34.69	0	0	0
hate	0	58.56	25.61	0	15.83	0
fear	0	1032	61.44	0	28.24	0
happy	0	0	0	63.08		36.92
sad	0	21.43	17.12	3.24	58.21	0
surprise	6.88	34.38	0	0	0	58.74
Average	60.89					

Fig. 8. Confusion matrix diagram of BALUM-ls dataset using MFFCNN.

4.5 Comparison with Other Methods

To verify the effectiveness and superiority of the algorithm in this paper, the comparison experiments were performed on RML and BALUM-ls based on ML-UDSGPLVM proposed by Nguyen et al. [17], ICTP proposed by Luo et al. [23], and DWMV and MFFCNN proposed by Zia et al. [25]. All comparisons are made when ensuring that the training set and test set are irrelevant. SVM is used as the classifier. The experimental results are shown in Table 4.

Table 4. Accuracy (%) obtained by several expression recognition methods

Datasets	ML-UDSGPLVM [17]	DWMV [25]	ICTP [23]	MFFCNN
RML	57.66	58.64	60.07	64.57
BAUM-ls	53.06	54.09	55.29	60.89

It can be seen from Table 4 that under the same classifier, compared with several other expression recognition methods, MFFCNN can achieve a higher recognition accuracy rate, since the proposed MFCNN fully considers the dynamic changes of the video sequence to learn the spatiotemporal features with discriminative ability. At the same time, by multiplying temporal and spatial streams, the spatial-temporal information at the feature level can be learned, which enhances the distinguishing ability of similarly-looking expressions. Therefore, the proposed MFFCNN can effectively improve the recognition accuracy of CNN model.

5. Conclusion

In this paper, a new video expression recognition method based on multi-mode CNN and multiplicative feature fusion is proposed. The training process of the proposed method is mainly divided into two stages. Firstly, the spatiotemporal CNN is fine-tuned on the target video expression dataset to learn the discriminative spatiotemporal expression features. Secondly, the learned spatiotemporal features are fused in feature layer, and finally classified by classifier. The results on RML and Baum-ls datasets show that the proposed MFFCNN fully considers the dynamic variation characteristics of video sequences, and can better learn the discriminative spatiotemporal features. At the same time, by multiplying and fusing the temporal and spatial streams, the spatiotemporal information at the feature level is learned, which is helpful to distinguish similar facial expressions. Deep learning model usually contains very complex network parameters, which requires a lot of computing resources. Therefore, the next step is to study how to reduce the model parameters of MFFCNN, so as to further improve the operation speed of deep learning under the premise of ensuring the recognition accuracy.

Acknowledgement

This work was supported by the Anhui Province University excellent young talents support project (No. GXYQ2019119), the Key natural science projects of Anhui Province (No. KJ2018A0820), and Bozhou University scientific research (No. BYZ2018C05).

References

[1] S. Maity, M. Abdel-Mottaleb, and S. S. Asfour, “Multimodal biometrics recognition from facial video with missing modalities using deep learning,” *Journal of Information Processing Systems*, vol. 16, no. 1, pp. 6-29, 2020.

[2] O. Agbolade, A. Nazri, R. Yaakob, A. A. Ghani, and Y. K. Cheah, “3-Dimensional facial expression recognition in human using multi-points warping,” *BMC Bioinformatics*, vol. 20, no. 1, article no. 619, 2019. https://doi.org/10.1186/s12859-019-3153-2

[3] K. Talele and K. Tuckley, “Facial expression recognition using digital signature feature descriptor,” *Signal, Image and Video Processing*, vol. 14, pp. 701-709, 2020. https://doi.org/10.1007/s11760-019-01595-1

[4] H. Sadeghi and A. A. Raie, “Human vision inspired feature extraction for facial expression recognition,” *Multimedia Tools and Applications*, vol. 78, no. 21, pp. 30335-30353, 2019.

[5] M. Sajjad, S. Zahir, A. Ullah, Z. Akhtar, and K. Muhammad, “Human behavior understanding in big multimedia data using CNN based facial expression recognition,” *Mobile Networks and Applications*, vol. 25, pp. 1611-1621, 2020. https://doi.org/10.1007/s11036-019-01366-9

[6] S. Nestler, “Safety-critical human computer interaction,” *it-Information Technology*, vol. 61, no. 1, pp. 67-70, 2019.

[7] P. Loslever, T. Guidini Gonçalves, K. M. de Oliveira, and C. Kolski, “Using fuzzy coding with qualitative data: example with subjective data in human-computer interaction,” *Theoretical Issues in Ergonomics Science*, vol. 20, no. 4, pp. 459-488, 2019.

[8] U. A. Shaikh, V. J. Vishwakarma, and S. S. Mahale, “Dynamic scene multi-exposure image fusion,” *IETE Journal of Education*, vol. 59, no. 2, pp. 53-61, 2018.

[9] Y. Jiang, K. Zhao, K. Xia, J. Xue, L. Zhou, Y. Ding, and P. Qian, “A novel distributed multitask fuzzy clustering algorithm for automatic MR brain image segmentation,” *Journal of Medical Systems*, vol. 43, article no. 118, 2019. https://doi.org/10.1007/s10916-019-1245-1

[10] F. Ramdani, M. T. Furqon, B. D. Setiawan, and A. N. Rusydi, “Analysis of the application of an advanced classifier algorithm to ultra-high resolution unmanned aerial aircraft imagery: a neural network approach,” *International Journal of Remote Sensing*, vol. 41, no. 9, pp. 3266-3286, 2020.

[11] N. Zikiou, M. Lahdir, and D. Helbert, “Hyperspectral image classification using graph-based wavelet transform,” *International Journal of Remote Sensing*, vol. 41, no. 7, pp. 2624-2643, 2020.

[12] F. An and Z. Liu, “Facial expression recognition algorithm based on parameter adaptive initialization of CNN and LSTM,” *The Visual Computer*, vol. 36, no. 3, pp. 483-498, 2020.

[13] R. Ramya, K. Mala, and S. S. Nidhyananthan, “3D facial expression recognition using multi-channel deep learning framework,” *Circuits, Systems, and Signal Processing*, vol. 39, no. 2, pp. 789-804, 2020.

[14] C. Xu, Y. Cui, Y. Zhang, P. Gao, and J. Xu, “Person-independent facial expression recognition method based on improved Wasserstein generative adversarial networks in combination with identity aware,” *Multimedia Systems*, vol. 26, no. 1, pp. 53-61, 2020.

[15] K. Li, Y. Jin, M. W. Akram, R. Han, and J. Chen, “Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy,” *The Visual Computer*, vol. 36, no. 2, pp. 391-404, 2020.

[16] S. Kumar, M. K. Bhuyan, and Y. Iwahori, “Multi-level uncorrelated discriminative shared Gaussian process for multi-view facial expression recognition,” *The Visual Computer*, vol. 37, no. 1, pp. 143-159, 2021.

[17] H. D. Nguyen, S. Yeom, G. S. Lee, H. J. Yang, I. S. Na, and S. H. Kim, “Facial emotion recognition using an ensemble of multi-level convolutional neural networksm,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 11, article no. 1940015, 2019. https://doi.org/10.1142/S0218001419400159

[18] X. Liu, X. Yin, M. Wang, Y. Cai, and G. Qi, “Emotion recognition based on multi-composition deep forest and transferred convolutional neural network,” *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 23, no. 5, pp. 883-890, 2019.

[19] N. Jain, S. Kumar, and A. Kumar, “Effective approach for facial expression recognition using hybrid square-based diagonal pattern geometric model,” *Multimedia Tools and Applications*, vol. 78, no. 20, pp. 29555-29571, 2019.

[20] F. Kong, “Facial expression recognition method based on deep convolutional neural network combined with improved LBP features,” *Personal and Ubiquitous Computing*, vol. 23, no. 3, pp. 531-539, 2019.

[21] F. Z. Salmam, A. Madani, and M. Kissi, “Fusing multi-stream deep neural networks for facial expression recognition,” *Signal, Image and Video Processing*, vol. 13, no. 3, pp. 609-616, 2019.

[22] X. Zhu and Z. Chen, “Dual-modality spatiotemporal feature learning for spontaneous facial expression recognition in e-learning using hybrid deep neural network,” *The Visual Computer*, vol. 36, pp. 743-755, 2020.

[23] Y. Luo, X. Y. Liu, X. Zhang, X. F. Chen, and Z. Chen, “Facial expression recognition based on improved completed local ternary patterns,” *Optoelectronics Letters*, vol. 15, no. 3, pp. 224-230, 2019.

[24] S. Nigam, R. Singh, and A. K. Misra, “Efficient facial expression recognition using histogram of oriented gradients in wavelet domain,” *Multimedia Tools and Applications*, vol. 77, no. 21, pp. 28725-28747, 2018.

[25] M. S. Zia, M. Hussain, and M. A. Jaffar, “A novel spontaneous facial expression recognition using dynamically weighted majority voting based ensemble classifier,” *Multimedia Tools and Applications*, vol. 77, no. 19, pp. 25537-25567, 2018.

[26] H. Wang, S. Wei, and B. Fang, “Facial expression recognition using iterative fusion of MO-HOG and deep features,” *The Journal of Supercomputing*, vol. 76, no. 5, pp. 3211-3221, 2020.

[27] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. P. Morency, “Openface 2.0: Facial behavior analysis toolkit,” in *Proceedings of 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, Xi'an, China, 2018, pp. 59-66.

[28] Y. Zhou and N. Chen, “The LAP under facility disruptions during early post-earthquake rescue using PSO-GA hybrid algorithm,” *Fresenius Environmental Bulletin*, vol. 28, no. 12A, pp. 9906-9914, 2019.

[29] J. Jian, Y. Guo, L. Jiang, Y. An, and J. Su, “A multi-objective optimization model for green supply chain considering environmental benefits,” *Sustainability*, vol. 11, no. 21, article no. 5911, 2019. https://doi.org/10.3390/su11215911

[30] Y. Ren, T. Cheng, and Y. Zhang, “Deep spatio-temporal residual neural networks for road-network-based data modeling,” *International Journal of Geographical Information Science*, vol. 33, no. 9, pp. 1894-1912, 2019.

Qun Ren **https://orcid.org/0000-0001-9578-1302**

She is born in Bozhou City, Anhui Province, China. She has got Master of Computer Science from Anhui University in 2011. She is currently an associate professor in Bozhou University. Her research interests include artificial intelligence and big data..