J Inf Process Syst, Vol.17, No.2, pp.399~410, April 2021
https://doi.org/10.3745/JIPS.01.0070

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

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

A Facial Expression Recognition Method Using Two-Stream Convolutional Networks in Natural Scenes

Lixin Zhao*

Abstract

Aiming at the problem that complex external variables in natural scenes have a greater impact on facial expression recognition results, a facial expression recognition method based on two-stream convolutional neural network is proposed. The model introduces exponentially enhanced shared input weights before each level of convolution input, and uses soft attention mechanism modules on the space-time features of the combination of static and dynamic streams. This enables the network to autonomously find areas that are more relevant to the expression category and pay more attention to these areas. Through these means, the information of irrelevant interference areas is suppressed. In order to solve the problem of poor local robustness caused by lighting and expression changes, this paper also performs lighting preprocessing with the lighting preprocessing chain algorithm to eliminate most of the lighting effects. Experimental results on AFEW6.0 and Multi-PIE datasets show that the recognition rates of this method are 95.05% and 61.40%, respectively, which are better than other comparison methods.

Keywords

Attentional Mechanism, Confrontational Learning, Double Flow Convolutional Neural Network, Image Preprocessing, Natural Scene Expression Recognition

1. Introduction

Natural scene expression recognition (facial expression recognition in the wild [FERW]), as the name implies, allows computers to automatically recognize the facial expressions of people in natural scenes [1-3]. In this context, the term “natural scene” refers to an indoor or outdoor scene where external conditions are not determined by humans, in contrast to a laboratory environment with controllable conditions. At the same time, the persons themselves are not constrained in the scene, and their emotional state is expressed naturally. Through accurate interpretation of facial expressions, people can ascertain the true feelings of other parties. Computers, on the other hand, can recognize human facial expressions in natural scenes so that they respond effectively to different emotional states of people. It is not only the only way for human-computer interaction to move from the traditional cold input-output interaction mode to a warm and intelligent mode, but also has great applicability in many real scenarios.

As an important research direction in the fields of emotion computing and pattern recognition, facial expression recognition has been receiving more and more attention from researchers since the 1990s

Manuscript received September 4, 2020; first revision October 21, 2020; November 8, 2020.
Corresponding Author: Lixin Zhao (smxzhaolixin@163.com)
* Dean's Office, Sanmenxia Polytechnic, Sanmenxia, Henan, China (smxzhaolixin@163.com)

[4,5]. Traditional facial expression recognition research generally focuses on the extraction, selection, and classifier training of discriminative facial expression features, and uses facial expression databases collected in laboratory scenes under controllable conditions for research. Compared with this type of research, natural scene expression recognition, as a novel expression recognition research direction, is aimed at complex unconstrained environments, and is more suitable for practical applications. In recent years, thanks to the establishment of more and more natural scene expression databases and the promotion of deep learning (DL) methods, more and more valuable research work has emerged around this topic. Based on the current findings, which suggest that complex external variables in natural scenes have a significant impact on facial expression recognition results, in this paper we propose a facial expression recognition method using two-stream convolutional neural network (CNN) models in natural scenes.

2. Related Work

Facial expression recognition based on traditional methods performs well on facial expression databases collected in controllable laboratory environments. However, in real-life natural scenes, factors such as lighting conditions, face poses, object occlusion and environmental background noise are not controlled by humans. These complex external variables have a great impact on traditional facial expression recognition methods, and also pose higher challenges to researchers.

Research on facial expression recognition in natural scenes based on deep learning is currently in the development stage. More and more studies are emerging on the two data formats of pictures and videos in natural scenes. Due to the difficulty of natural scene images with facial expressions, various model fusion methods are often used to improve the overall recognition performance. In [6], multiple convolutional neural network models were trained simultaneously, and decision-level fusion based on the fully connected output of each level of a CNN was performed through exponential weighting. In [7], the probability results of 72 CNN outputs were combined into a tensor, and then a CNN was used to learn fusion weights.

However, most of the above algorithms were based on the preconditions of slow illumination changes and rapid reflectivity changes, and are incapable of solving the problems of different brightness and different facial poses in natural scene images. In order these problems, some effective local features of classic images have been used as network inputs. In [8], only the area around the eyebrows, eyes and mouth was considered, and these areas with discriminative expression features were used to train a deep sparse autoencoder based on softmax regression to recognize expressions in images. The authors of [9] combined boosting ideas and deep belief networks to propose a boosted deep belief network (BDBN). First, the face image was divided into blocks with a grid, and the initial feature expression of the image blocks was learned hierarchically based on a bottom-up unsupervised method. Then, the features were further optimized based on a top-down supervised approach, and a series of weak learners was learned iteratively, followed by learning of the multi-layer feature expression of a specific area of the picture through multiple DBNs. However, due to the difficulty of natural scene expression images, methods of this type can easy succumb to overfitting during network training.

In order to strengthen the influence of expression category information on each network level, in [10] supervision information was provided to each convolution level in a CNN supervised scoring ensemble (SSE) module. The output of the module was cascaded to the input of the fully connected layer for final

classification. At the same time, the authors combined the SSE module with three deep network architectures, and the final classification results were obtained through a weighted fusion of the results of each model. Similarly, in [11] a deep supervision CNN was proposed, which introduced a bypass output structure to the different scale feature maps output by different convolutional layers. The supervision information could directly guide each layer of convolution to learn deep features more relevant to the expression category, thereby improving the accuracy of expression recognition. However, judging from the results of facial expression recognition in natural scenes, the performance of deep CNN models is far from satisfactory, and there is still considerable room for improvement. Therefore, in this paper we propose a new facial expression recognition method using a two-stream CNN model in natural scenes. The model can autonomously learn the areas that need attention in the face video frame sequence. By adjusting the attention weights of different areas, the network can focus more on the facial area around the facial action unit (AU). At the same time, uninteresting areas are ignored to suppress their influence on expression recognition results. The main contributions of this model are as follows:

1) Use of the lighting preprocessing chain algorithm for lighting preprocessing to eliminate most of the lighting effects. At the same time, robust illumination edge information can be extracted while retaining the basic appearance details required for recognition.
2) The proposed two-stream CNN model finds some facial regions closely related to various expressions in the static flow in space and dynamic flow in time through learning. Focus is placed on learning the discriminative emotional characteristics of these regions in space and time, while suppressing the influence of other facial regions on expression recognition. This is beneficial to enhance the model's anti-interference ability in complex natural scenes, thereby improving the accuracy and robustness of facial expression recognition effectively.

3. Overall Architecture and Basic Concepts of the Proposed Method

In order to learn the discriminative space-time features related to facial expressions in face data more effectively, in this article we first consider the input and output of the convolutional layer in the two-stream CNN, and introduce the attention mechanism in two aspects. First, a two-stream CNN model based on shared attention mechanism is proposed, the overall structure of which is shown in Fig. 1. It is composed of an M-CNN with blue wire frame and an S-CNN with green wire frame. The feature maps output by the last convolutional layer of the two parts are connected in series as the space-time feature expression of facial texture data. The attention mechanism module after eight convolution input weight matrices is shown in the red box in the middle of M-CNN and S-CNN and the feature map in series.

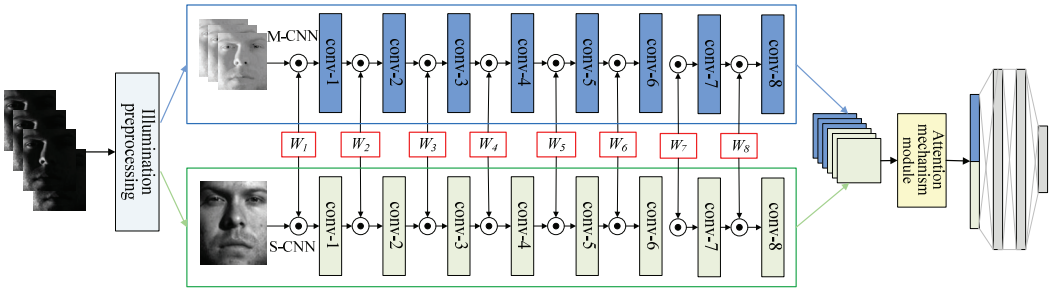

Fig. 1. Facial expression recognition framework using double flow convolutional neural network model.

From the point of view of input static video frame and optical flow sequences, we expect that the network will highlight the region of facial muscles, that is, the area around the AUs. By enhancing the input information of these regions, the features of these regions will always be retained in the data stream of the network. Naturally, a weight is assigned to each position of the network input to represent its contribution to the expression. Then, the weight is multiplied with the input of the corresponding position to enhance or suppress the input information at that position. In the same manner, this enhancement operation can be used not only for the initial input of the network, but also for the feature map of each subsequent layer of convolution input. Especially after the pooling operation, the size of the convolutional input will be reduced.

In two-stream CNN, both the M-CNN and S-CNN use eight groups of convolutional layers similar to the VGG-M structure, and the length and width of the static frame and optical flow image sequence input to the network are both 224×224. Therefore, the inputs of each group of convolutional layers in the static and dynamic flow networks are the same size. In addition to the first group of convolutional layers, the input feature map size of each group of convolutional layers is half the size of the previous layer. At the same time, facial muscles have a limited range of facial expressions. The position of the muscle movement area in the optical flow sequence is not much different from that in static video frames. Therefore, the weight matrix W_i marked by the five red boxes in Fig. 1 is shown. Each group of convolutional layers of the M-CNN and S-CNN in the TSCN-SA can share the same weight matrix to enhance or suppress the current input signal before input.

At the same time, in order to prevent the network from resetting the weight of important areas to 0 at the initial stage, all the information on the area is lost. In this paper, the e index operation is performed on each value in the weight matrix W_i, so that the ownership value is greater than 0. Then, e^{W_i} is multiplied with the input information of the corresponding position, so that even if the information of some important areas is suppressed in the initial stage, the subsequent network will have the opportunity to recover it. Therefore, in this paper we refer to this weight matrix shared between the M-CNN and the S-CNN as an exponentially-enhanced shared convolution input weight. In this manner, different size inputs of different convolutional layers have their spatial position importance explicitly defined. The weight matrix introduces a very limited amount of parameters in the network and because only exponential functions are used, the process of derivation during back propagation is not complicated.

3.1 Illumination Preprocessing

Illumination preprocessing is performed before feature extraction, and entails a series of stages designed to remove the effects of lighting changes, local shadows and highlights and to preserve the basic elements of visual appearance. Fig. 2 shows the three main stages and their impact on typical images containing faces.

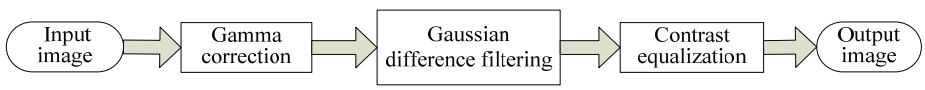

Fig. 2. Three stages of illumination preprocessing.

Gamma correction is a non-linear grayscale transformation, which compresses and expands a certain grayscale range. After correcting the image using the algorithm, the differences between elements

containing image information become more pronounced. The basic form of gamma correction is:

$$s = c \cdot r^{\gamma} (r \in [0,1]) \tag{1}$$

where c takes the normal number, while γ is the gamma coefficient. The value of the gamma coefficient determines the mapping relationship between the input and the output images. When the value of the gamma coefficient is less than 1, the contrast of areas with a low gray value is be enhanced. When the value of the gamma coefficient is greater than 1, the gamma correction can increase the contrast of darker areas. When the gamma coefficient is 1, the original image will not change.

Gaussian differential filtering is used to extract robust lighting edge information, and is essentially achieved using a Gaussian filter obtained by subtracting two different Gaussian filters. The specific definition is as follows:

$$DOG(x, y, \sigma_1, \sigma_2) = G_{\sigma_1}(x, y) - G_{\sigma_2}(x, y) \tag{2}$$

In Eq. (2), the two-dimensional expression of the Gaussian function is:

$$G_{\sigma}(x, y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \tag{3}$$

where σ is the standard deviation of the Gaussian function, which determines the smoothness of the Gaussian filter. Larger values of σ, correspond to a wider frequency band of the Gaussian filter and yield a better smoothing effect. The result $g_{12}(x, y, \sigma_1, \sigma_2)$ of the image $f(x, y)$ filtered by $DOG(x, y, \sigma_1, \sigma_2)$ is given by:

$$g_{12}(x, y, \sigma_1, \sigma_2) = DOG(x, y, \sigma_1, \sigma_2) * f(x, y) \tag{4}$$

Here, $*$ is the convolution operator and $f(x, y)$ is the original image. When $\sigma_1 > \sigma_2$, Gaussian differential filtering is equivalent to band-pass filtering the image. When the band-pass frequency is consistent with the frequency band of the shadow part of the face, the shadow part is strongly highlighted, and high-frequency noise and low-frequency background information are suppressed.

Contrast equalization: The final stage of lighting preprocessing is to readjust the overall contrast of the image, which reduces the intensity difference of the edge information of the face image. Eqs. (5) and (6) are the two stages of contrast equalization, which can be used to realize normalization in a computationally efficient manner.

$$I(x, y) \leftarrow \frac{I(x, y)}{\left(mean\left(\left|I(x', y')\right|^{a}\right)\right)^{1/a}} \tag{5}$$

$$I(x, y) \leftarrow \frac{I(x, y)}{\left(mean\left(\min\left(\delta, \left|I(x', y')\right|\right)^{a}\right)\right)^{1/a}} \tag{6}$$

Here, a is the compression index, whose role is to reduce the influence of larger values. $\delta$is a threshold used to truncate larger values after the first stage of standardization. Although the processed image has a

good zoom effect, it still contains extreme values. To reduce their impact on subsequent processing stages, a nonlinear mapping is applied to compress these excessive values. Eq. (7) uses the hyperbolic tangent function to limit the image pixel values to the range of $(-\delta, \delta)$.

$$I(x,y) \leftarrow \delta \tanh\left(\frac{I(x,y)}{\delta}\right) \tag{7}$$

3.2 Network Structure and Parameters

Fig. 3 shows the structure and parameters of the CNN used in this article. Since processing a single frame of video image is similar to the structure of the trained and optimized VGG-M model, the input layer is the original video frame image with a size of 640×360×3. After the first convolution operation (with a convolution kernel of 7×7×96), local response normalization is carried out, and the result is input into the maximum pooling layer with size 3×3 and a step size of 2 to obtain the processing results. These results are then input into the second convolutional layer (with a 5×5×256 convolution kernel) and the second maximum pooling layer (size of 3×3, step size is 2), and the convolution operations are continued. The size of each convolution kernel is shown as the convolution layer 3–8 in Fig. 1. Get the result of the video frame after the convolution operation. It is worth noting that except for the last convolution layer, each of the remaining convolution layers is immediately followed by a rectified linear unit (ReLU) for processing as an activation function. Finally, the deconvolution operation is performed with a kernel size of 8×8×1 and a step size is 4. The final output is the same size as the input video frame image, and reflects the area that attracts human visual attention in the video frame.

Conv-1：7 × 7 × 96
Local response normalization
MaxPool-1: 3 × 3, Stride: 2
Conv-2：5 × 5 × 256
MaxPool-2: 3 × 3, Stride: 1
Conv-3：3 × 3 × 512
Conv-4：5 × 5 × 512
Conv-5：5 × 5 × 512
Conv-6：7 × 7 × 256
Conv-7：11 × 11 × 128
Conv-8：13 × 13 × 1
Deconvolution operation: 8 × 8 × 21, Stride: 4

Fig. 3. Structure and parameters of CNN used in this paper.

Optical flow is an important method for the analysis of moving images at present, as it can extract change information from the target's motion in the image. Suppose two adjacent video frames are obtained at t and $t + \Delta t$, respectively. Assume that the gray value of a point at moment t is $I(x, y, t)$. When the point moves to $(x + \Delta x,\ y + \Delta y)$ at moment $t + \Delta t$, its gray value becomes $I(x + \Delta x,\ y + \Delta y,\ t + \Delta t)$ and remains unchanged for a short time interval Δt, i.e.,

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \tag{8}$$

Expanding the above equation using Taylor's formula, we obtain:

$$\frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial t} = 0 \tag{9}$$

In the x and y directions, assuming that $u = dx/dt$, $v = dy/dt$, I_x, I_y and $I(x, y, t)$ represent the partial derivative of coordinates x, y and time coordinate t, respectively, the above equation can be abbreviated as:

$$I_x u + I_y v + I_t = 0 \tag{10}$$

3.3 Attention Mechanism

In the last set of the neural network's convolution layers, the feature maps output by the M-CNN in time and the S-CNN in space are stacked together to form a deep space-time feature expression of facial texture. However, for facial expressions, different expressions are composed of different AUs. If the space-time features in the area that are more discriminative to the expression category are to be highlighted, the facial area around the AU corresponding to the expression category must be determined and any external interference introduced by occlusion in the natural scene must be suppressed. Therefore, an attention mechanism module is integrated in the series of feature maps to learn the importance of deep space-time features at different positions and map the deep features of the image based on this fused feature.

The feature map output by the CNN passes through two fully connected layers to obtain the attention weight matrix, and then weighted fusion is applied on the matrix and the original CNN feature map to obtain the deep space-time features with attention.

Assume that the dimension of the combined feature map is $N \times N \times C$, where C is the total number of channels, and the space-time feature vector of coordinate (i, j) is $F_{i,j}$. Then, the initial attention weight $a_{i,j}$ at this position obtained after two layers of full connection is:

$$\begin{aligned} a_{i,j} &= \sigma\left(W_2^T \sigma\left(W_1^T F_{i,j} + b_1\right) + b_2\right) \\ i, j &= 1, 2, \ldots, N \end{aligned} \tag{11}$$

where σ is the activation function, while W_1, b_1 and W_2, b_2 are the fully connected parameter matrix and offset of the first and the second layer, respectively. Therefore, the attention weight matrix represented by the red solid line box in the figure can be obtained, and then the attention weight matrix is normalized using a softmax function:

$$A_{i,j} = \frac{e^{a_{i,j}}}{\sum_{i=1}^{N}\sum_{j=1}^{N} e^{a_{i,j}}} \tag{12}$$

The attention weight matrix A is multiplied with the original feature map at the corresponding position, and the deep space-time features at different positions are weighted to obtain the deep space-time feature vector b with attention, as shown in Eq. (13), where ⊙ indicates the multiply operation at the corresponding position.

$$b = \sum_{i,j} A \odot F \tag{13}$$

4. Experimental Results and Analysis

In this section, we verify the proposed facial expression recognition method using the two-stream CNN model on two expression libraries, and compare with other methods. The expression databases used in this section are the Multi-PIE database of expressions collected in the laboratory and the natural scene database AFEW6.0.

4.1 Experimental Dataset

The Multi-PIE dataset contains face images of 337 people in 15 poses and 19 lighting conditions. It is a dataset mainly used for facial recognition tasks. However, because some of the data have emoticon tags, the database is also widely used in facial expression recognition tasks. In order to facilitate comparison with the experimental results of existing methods, the same experimental settings as those in [12] were used. A total of 270 people was used in five different poses (-30°, -15°, -0°, 15°, 30°) and 6 expressions (neutral, disgusted, surprised, happy, screaming and squinting). The AFEW6.0 data set is a natural scene database. For the experiments of this section, we used 534 training samples and 278 validation samples from the AFEW6.0 database to analyze the performance of the model by comparing the accuracy of various methods on the validation set.

4.2 Experimental Setup

Both expression datasets used in the experiment contained background information unrelated to the face. The size of the cropped face image was normalized to 224×224 to adapt to the input of the network. To ensure that the expression intensity in the static video frame input of the S-CNN was sufficiently large, 5–10 frame images at the beginning and end of each sequence were discarded depending on the length of the video sample. One frame was randomly extracted from the remaining frame sequence as the input of S-CNN. Meanwhile, using this frame as the starting point, the optical flow sequence corresponding to a frame length of L=10 was taken backward and input into the M-CNN. To expand the sample size, for each video sample in the training set we extracted a combination of static frames and optical flow sequences multiple times depending on the number of frames and we performed data augmentation in the form of flips for all training samples. For the test data, three groups were randomly selected from the middle of the video sample. We used TensorFlow to train and test the networks of the experiment. The initial learning rate of the network was set to 0.002, and the Adam optimizer was used to train all network structures.

4.3 Display of Recognition Results

4.3.1 Analysis of facial expression recognition results on Multi-PIE

Table 1 shows the expression recognition rates of the proposed method under each expression and each posture on the Multi-PIE dataset. It can be seen from the table that the average recognition rate of the proposed method on the Multi-PIE dataset was 95.05%, while the last column of Table 1 shows that,

except for the "fear" expression, the average recognition rate of the other expressions exceeded 94%. Among the six expressions, the highest recognition rate was achieved for "happiness," with a recognition rate of 97.04%. The recognition rate of this expression exceeded 95% for all poses. Among the six expressions, the one with the lowest recognition rate was the "fear" expression, with a recognition rate of only 92.31%.

Fig. 4 shows the confusion matrix of the facial expression recognition results of the proposed method on the Multi-PIE dataset, where it can be seen that the low recognition rate of the "fear" expression is mainly due to the higher confusion rate between them and the "surprised" expressions. 7.11% of the fear expression images were recognized as "surprised" expressions, and 7.84% of the "surprised" expressions were recognized as "hate" expressions. The confusion between them was mainly due to the similar texture changes of the two expressions around the eyes.

Fig. 5 shows the results of a comprehensive comparison between the proposed method and other existing methods on the Multi-PIE dataset. It can be seen that the proposed method performed significantly better than the methods based on manual features. Compared with the existing methods, except for the -15° pose, the expression recognition rate of the proposed method was higher. It can also be seen from Fig. 5 that the recognition rate of the other methods is usually low on frontal pose images, but the proposed method has a greatly improved facial expression recognition rates in these poses.

Table 1. Expression recognition results under different poses on the Multi-PIE dataset (unit: %)

Expressions	-30°	-15°	0°	15°	30°	Average
Neutral	96.21	92.81	92.72	93.93	95.73	94.28
Hate	97.14	95.48	94.81	96.89	98.63	96.59
Surprise	97.28	92.81	93.28	94.43	96.15	94.79
Happy	97.81	96.31	95.67	97.16	98.25	97.04
Fear	93.68	91.38	89.84	92.43	94.22	92.31
Strabismus	97.86	93.14	93.51	94.62	97.17	95.26
Average	96.66	93.66	93.31	94.91	96.69	95.05

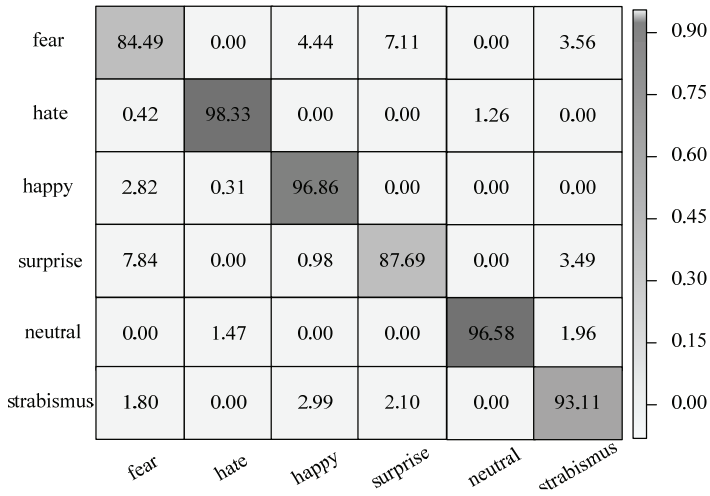

Fig. 4. Confusion matrix of expression recognition results on Multi-PIE dataset.

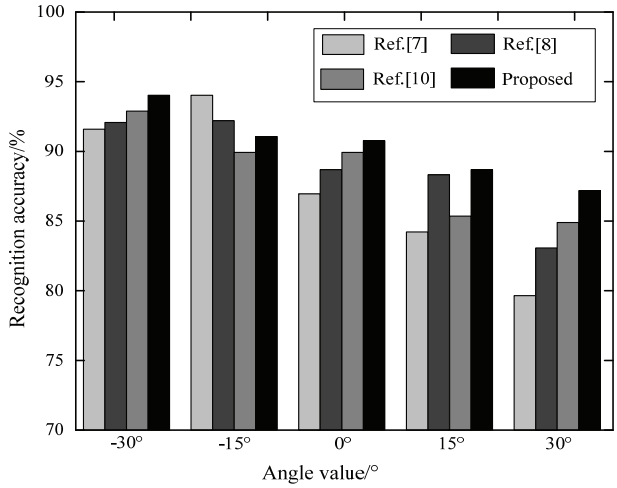

Fig. 5. Comparison with existing methods on Multi-PIE database.

4.3.2 Analysis of facial expression recognition results on natural scene database AFEW6.0

Table 2 lists the recognition rates of various methods on the validation set. Due to the difficulty of facial expression recognition in natural scenes, the recognition success rate of most methods is relatively low, namely less than 60%. The facial expression recognition method proposed in this paper using the two-stream CNN model achieves a recognition rate of 61.40% on the AFEW6.0 verification set, which is slightly higher than the other state-of-the-art methods.

Table 2. Comparison of recognition rates of various methods on AFEW6.0 dataset

Method	Recognition accuracy (%)
Pons and Masip [7]	56.48
Chen et al. [8]	58.53
Hu et al. [10]	55.09
Proposed method	61.40

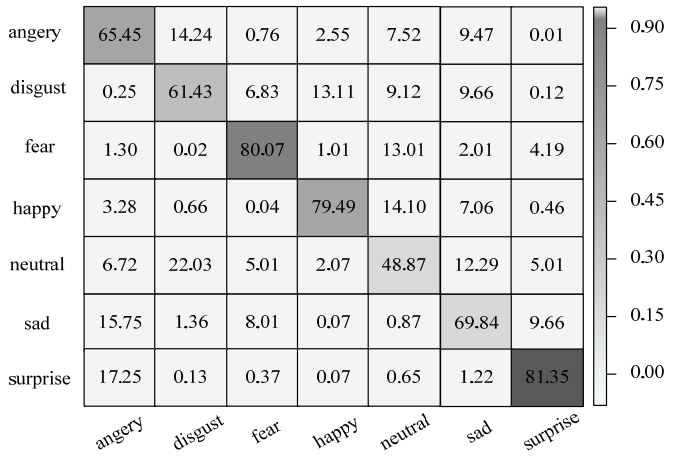

Fig. 6. Confusion matrix of expression recognition results on AFEW6.0 dataset

The corresponding confusion matrix is shown in Fig. 6. The recognition rates of "fear" and "surprise" was higher, while that of "disgust" and "neutral" was lower. Due to the large difference between "surprise" and other expressions, the attention mechanism can distinguish it from other expressions by focusing on the area around any facial AU, and thus the highest recognition rate of 81.35% is achieved. At the same time, due to a certain degree of sample imbalance in the training set, a considerable number of samples were mistakenly classified as "angry."

5. Conclusion and Prospects

Compared with static images, expression videos of natural scenes contain complicated facial texture timing changes. Applying common facial expression methods directly on the frame images will result in a loss of information on timing changes. Based on the deep learning method and introducing the attention mechanism, in this paper, we propose an expression recognition method based on the two-stream CNN with a shared attention mechanism. Results on expression video libraries in laboratory and natural scenes show that the accuracy of expression recognition is significantly improved through the influence of the shared attention mechanism. The proposed method is significantly better than facial expression recognition methods based on manual features. At present, the sample size of existing natural scene expression databases is relatively limited, especially when it comes to natural scene video databases. Deep learning methods generally require large amounts of training data. In response to this shortcoming, the use of data augmentation or transfer learning ideas to introduce other related databases and other methods can play a limited role. In order to solve this problem fundamentally, in the future, a large-scale expression database with rich diversity of natural scenes can be established through crowdsourcing and other methods.

References

[1] N. P. Gopalan, S. Bellamkonda, and V. S. Chaitanya, "Facial expression recognition using geometric landmark points and convolutional neural networks," in *Proceedings of 2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, Coimbatore, India, 2018, pp. 1149-1153.

[2] Y. He and X. He, "Facial expression recognition based on multi-feature fusion and HOSVD," in *Proceedings of 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Chengdu, China, 2019, pp. 638-643. https://doi.org/10.1109/ITNEC.2019.8729003

[3] S. Wang, B. Pan, H. Chen, and Q. Ji, "Thermal augmented expression recognition," *IEEE Transactions on Cybernetics*, vol. 48, no. 7, pp. 2203-2214, 2018. https://doi.org/10.1109/TCYB.2017.2786309

[4] J. Ueda and K. Okajima, "Face morphing using average face for subtle expression recognition," in *Proceedings of 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, Dubrovnik, Croatia, 2019, pp. 187-192. https://doi.org/10.1109/ISPA.2019.8868931

[5] N. Song, H. Yang, and P. Wu, "A gesture-to-emotional speech conversion by combining gesture recognition and facial expression recognition," in *Proceedings of 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, Beijing, China, 2018, pp. 1-6. https://doi.org/10.1109/ACIIAsia.2018.8470350

[6] B. K. Kim, J. Roh, S. Y. Dong, and S. Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 173-189, 2016.

[7] G. Pons and D. Masip, "Supervised committee of convolutional neural networks in automated facial expression analysis," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 343-350, 2018. https://doi.org/10.1109/TAFFC.2017.2753235

[8] L. Chen, M. Zhou, W. Su, M. Wu, J. She, and K. Hirota, "Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction," *Information Sciences*, vol. 428, pp. 49-61, 2018.

[9] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 1805-1812. https://doi.org/10.1109/CVPR.2014.233

[10] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, Glasgow, UK, 2017, pp. 553-560.

[11] Y. Fan, J. C. Lam, and V. O. Li, "Video-based emotion recognition using deeply-supervised neural networks," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, Boulder, CO, 2018, pp. 584-588.

[12] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 189-204, 2014. https://doi.org/10.1109/TIP.2014.2375634

Lixin Zhao **https://orcid.org/0000-0002-1210-5905**

He was born in 1981. He holds a Master in Electronic Information Engineering, and is currently an Associate professor. He graduated from Henan University of Science and Technology in 2013, and has worked in Sanmenxia Polytechnic. His major research interests include computer networks.