A method for facial expression recognition of image sequence based on spatial features

Xin Li[*] and *Xiangyu* Zheng

Department of Computer Science University of Shantou Shantou City, Guangdong Province, China

Keywords: Facial expression recognition, Image sequence, Attention mechanism, Spatial features, Convolutional neural network.

Abstract. Facial expression recognition is the foundation of human emotion recognition, which has become a hot topic in the field of artificial intelligence in recent years. To some extent, facial expressions can be regarded as the process of facial muscle changes. Image sequence contains richer expression contents compared with a single image. So the expression recognition based on image sequence can yield more accurate results. A new method for facial expression recognition is proposed in this paper, which is based on spatial feature for image sequence. The method consists of four steps. Firstly, Siamese neural network is used to construct an evaluation model for changes of expression intensity, which extracts an appropriate image sequence from the video. Secondly, a convolutional neural network with attention mechanism is designed and trained, which is used to extract spatial features from each image in the sequence. Then, the spatial features of multiple images are fused. Finally, the fusion results are put into a convolutional neural network to recognize the facial expressions. This method is validated on CK+ dataset and the experimental results show that it's more accurate than several other methods.

1 Introduction

Facial expressions are one of the most important ways to reflect human emotions. Researches show that over half of information exchange between people is transmitted through facial expressions [1]. With the increasing demand for human-machine intelligent interaction, problems of insufficient emotional interaction abilities in intelligent products are becoming prominent. So, facial expression recognition, as a key issue in emotional interaction, has become a hot topic in AI researches. The progress in this field will effectively improving the effectiveness of intelligent applications.

The emergence of deep learning has greatly promoted the research of facial expression recognition. In 2017, Chollet proposed Xception model [2] which effectively reduced the number of model parameters by replacing ordinary convolutions with deep separable ones. In 2018, Hu J migrated the attention mechanism from natural language processing to

[*] Corresponding author: lixin@stu.edu.cn

computer vision direction and proposed SENet (Squeeze-and-Excitation Networks) [3]. At the same time, Woo proposed CBAM (Convolution Block Attention Module) [4], which reweighted the extracted features based on spatial and channel information. Jinyuan Ni introduced channel attention mechanism into neural network so that the network could recognize facial expression of partially erased faces [5]. Aouayeb constructed a vision transformer network based on self attention mechanism and channel attention [6].

Technically, facial expression recognition can be divided into two routes: facial expression recognition of single image and facial expression recognition of image sequence from the video. Relatively, methods of image sequence can express the process of facial changes, which includes more facial information, so that the recognition can be more accurate. How to process the information from the video quickly and how to improve the accuracy of the recognition are key issues in the research. Dae Kim used convolutional neural networks to extract features from each frame of the video and used long-short term memory networks to extract temporal features between the frames [7]. Ling Log took changes in facial motion units as the input of a graph convolutional network to extract association features between facial motion units [8].

Lots of expression recognition methods for image sequence use cyclic neural networks to process feature correlations between the image frames. The structural characteristic of cyclic neural networks requires that the processing of the next image frame depends on the processing of the previous one. So, the execution efficiency of cyclic neural networks will inevitably be affected. In this paper, a new recognition method of image sequence is proposed. Compared to previous methods, it focuses more on the following three aspects: (1) Reducing the number of images in the sequence through expression intensity evaluation, thereby significantly reducing computational burdens. (2) Constructing a model for facial expression recognition based on spatial features, which uses attention mechanism to improve recognition accuracy. (3) Fusing spatial features from multiple images to process them as a whole, which can accelerate the recognition remarkably.

The rest of this paper is organized as follows: the second part is the methodology and the third part is empirical study from CK+ (Cohn-Kanade+) dataset. The last part is the conclusion.

2 Methodology

2.1 Extracting image sequence

The method for facial expression recognition in this paper can be divided into four steps, which is shown in Figure 1.

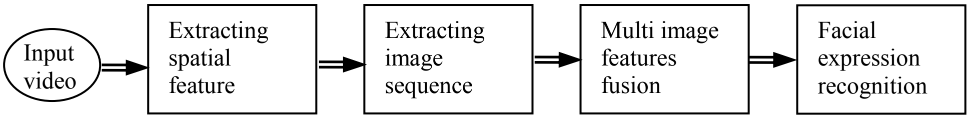
Input video
Extracting spatial feature
Extracting image sequence
Multi image features fusion
Facial expression recognition

Fig. 1. The framework of facial expression recognition of video.

Facial expression can be regarded as the process of facial muscle changes. A video used for facial expression recognition typically contains over ten to twenty image frames. The difference between some adjacent frames sometimes may be very small. Therefore, partial images can be extracted to replace the entire video for the recognition. With extensive experimental verification, five appropriate images which are selected from a video of about twenty frames can usually achieve good result. The evaluation of the intensity is the basis for extracting image sequence. Siamese neural network is used to construct an evaluation model for the change of two images, which is shown in Figure 2.

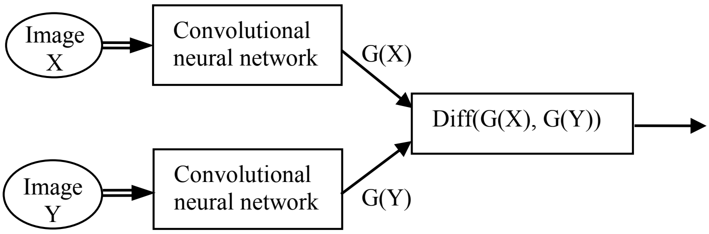

Fig. 2. Siamese neural network model.

If Diff(G(x), G(Y))>*strthd*, image Y is picked out, otherwise, image Y is discarded. Due to space limitations, the detailed structure of the convolutional neural network in Figure 2 is omitted here.

2.2 Extracting spatial feature

Literature [4]-[6], etc. have detailed elaborated the application of attention mechanism in image feature extraction. The attention model in this paper references the researches of above literatures. The important component in the model is CBAM (Convolution Block Attention Module). The structure of CBAM is shown in Figure 3. Feature vectors as the input pass through those two attention modules sequentially. The extracted features will become more discriminative through the attention modules.

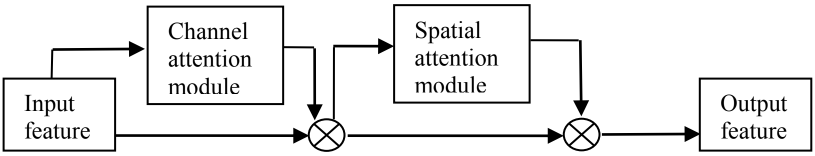

Fig. 3. The structure of attention module.

The input of channel attention module is a feature graph with the size of C×H×W, where C is the number of channels, H and W are the length and width of the graph. The output of channel attention module is a new feature graph. The product of the output and initial feature graph is taken as the input of spatial attention module. The structure of spatial attention module is shown in Figure 4.

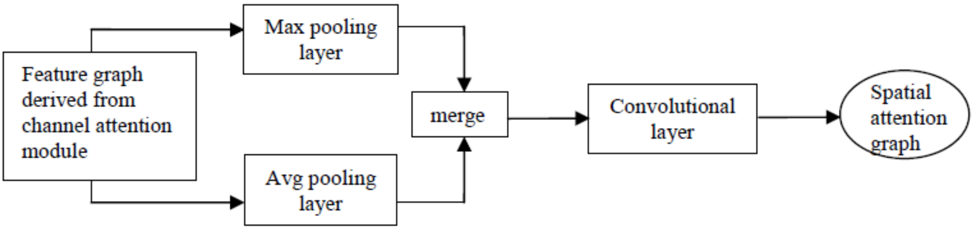

Fig. 4. The structure of spatial attention module.

Then, as shown in the Figure 3, the spatial feature is multiplied by previous channel feature. The product is taken as the output of whole attention module.

2.3 Features fusion of multiple images and Facial expression recognition

After extracting spatial features from each image, the next step is to use the features of images in the sequence to recognize facial expression. Different from cyclic neural network where the feature of each image is put into the network one by one to process, the features of all images are input and processed as a whole in this paper. Firstly, the relationship between multiple images should described appropriately. Then, the features of all the images are fused according to that relationship.

Obviously, the images in the sequence have a temporal order relationship which can be represented by a directed graph. Two matrices are constructed: matrix A and matrix D.

$$A=\begin{vmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{vmatrix} \qquad D=\begin{vmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{vmatrix}$$

Matrix A represents the temporal order and matrix D represents the penetration of each node in the graph. Then, Define Matrix G,

$$G=(D+I)^{-1}*(A^T+I) \tag{1}$$

Here, I is the identity matrix. Thus, the value of G can be obtained through the calculation. Matrix G is the fusion matrix for the features of multiple images.

Partial structure of the network for facial expression recognition is shown in Figure 5.

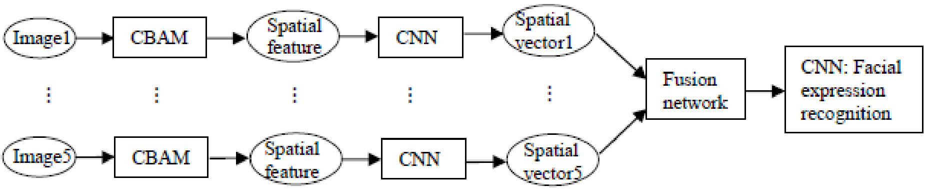

Fig. 5. Partial structure of the network for facial expression recognition.

The output of the recognition CNN is a vector of facial emoticon types. There are seven types of emoticon data in this paper, which are anger, contempt, disgusted, fear, happy, sadness and surprise. The form of emoticon vector Emo is shown as follows.

$$Emo=[emo_1, emo_2 \ldots\ldots emo_7] \tag{2}$$

Then, Softmax function is used to normalize Emo. Due to space limitations, the detailed structures of the CNNs in Figure 5 are omitted here.

3 Experiments

The experimental data used in this paper is CK+ dataset. The experiment selects 70% of the data as training set and the other 30% as test set. Dlib library is used to detect the position of the face for each image and the face image is cropped out with the size of 64*64. The process of accuracy change of test set for facial expression recognition in image sequence is shown in Figure 6. It can be seen that the accuracy improvement gradually tend to stability after 170 epochs and it reaches about 89.57%.

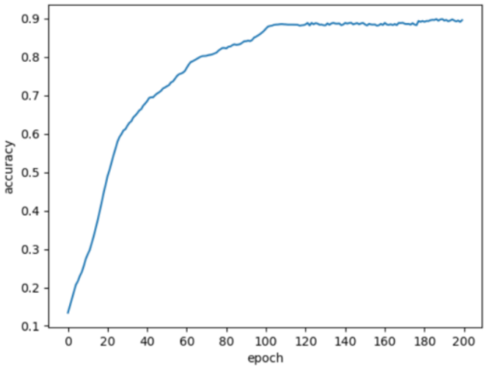

Fig. 6. Accuracy change of test set for facial expression recognition in image sequence.

Table 1 compares the recognition accuracy of our method with other methods.

Table 1. The accuracy of different expression recognition method.

Method	Accuracy
Our method	89.57%
LBP-TOP[9]	88.99%
3D SIFI[10]	81.35%
3DCNN[11]	85.9%
OCNN+LSTM[12]	88.4%
GCNN+LSTM[12]	89.8%

It can be seen that the accuracy of our method is higher than those based on traditional feature extraction operators such as LBP and SIFI. It also has approximate accuracy compared to the methods based on convolutional neural networks which are combined with cyclic neural network. As discussed earlier, it is difficult for cyclic neural network to achieve high efficiency.

4 Conclusions

Benefit from the development of convolutional neural networks, researches on facial expression recognition have made significant progress in recent years. Compared to a single image, the video contains more abundant information and wider application scenarios. However, video oriented method also has some disadvantages, for example, large amount of data and big time consumption. In this paper, a video oriented method for facial expression recognition is proposed, which investigates some bottleneck issues in video processing. The contributions of this paper are listed as follows,

(1) Extracting a image sequence from the video by calculating the change in expression intensity. This step significantly reduces the number of images and thus reduces processing burdens.

(2) Extracting image spatial features by CBAM modules which adopt attention mechanism. The extracted spatial feature graphs can better highlight facial expressions than original image.

(3) Fusing the features of multiple images based on temporal relationships between them. Thus, multiple features can be processed as a whole, which improves recognition efficiency remarkably.

Experiments show that the method in this paper has good recognition performance. In the future, sound information in the video can be considered to process in combination with image information since the changes in the tone also reflect human emotions. Comprehensively processing sound and image information will bring better results.

References

1. Mehrabian A. Communication without words [M]. Communication theory. Routledge, 2017, pp. 193-200.
2. Chollet F. Xception: Deep learning with depthwise separable convolutions[C]. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2017, pp. 1800-1807.
3. Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks[C]. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018, pp. 7132-7141.
4. Woo S, Park J, Lee J Y. CBAM: Convolutional block attention module[C]. In Proceedings of the European conference on computer vision (ECCV). 2018, pp. 3-19.
5. Jinyuan Ni, Jianxun Zhang, Xinyue Zhang. Facial expression recognition based on deep wide residual network attention mechanism [J]. Journal of Chongqing University of Technology, 2023, vol. 37, No. 1, pp. 9-14.
6. Aouayeb M, Hamidouche W, Soladie C, et al. Learning vision transformer with squeeze and excitation for facial expression recognition[J]. arXiv: 2107.03107, 2021.
7. Kim D H , Baddar W J , Jang J , et al. Multi-Objective Based Spatio-Temporal Feature Representation Learning Robust to Expression Intensity Variations for Facial Expression Recognition[J]. Affective Computing, 2019, vol. 10, No. 2, pp. 223-236.
8. Ling L, Xie H, Shuai H, et al. MER-GCN: Micro Expression Recognition Based on Relation Modeling with Graph Convolutional Network[C]. In Proceedings of 2020 IEEE Conference on Multimedia Information Processing and Retrieval, 2020.
9. Zhao G, Pietikainen M. Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007, vol. 29, pp. 915-928.
10. Scovanner P. A 3-dimensional sift descriptor and its application to action recognition [J]. ACM Multimedia, 2007.
11. Ji S, Xu W, Yang M. 3D Convolutional Neural Networks for Human Action Recognition [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, vol. 35, no. 1, pp. 221-231.
12. Min Hu, Yong Gao, Hao Wu. Fusion of Edge Detection and Recurrent Neural Network for Video Expression Recognition [J]. Journal of Electronic Measurement and Instrumentation, 2020, vol. 7, no. 34, pp. 103-111.