HIGH TECHNOLOGY LETTERS | Vol. 29 No. 2 | Jun. 2023 | pp. 213-222

doi:10.3772/j.issn.1006-6748.2023.01.012

Micro-expression recognition algorithm based on graph convolutional network and Transformer model①

WU Jin(吴 进)②, PANG Wenting, WANG Lei, ZHAO Bo

(School of Electronic Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, P. R. China)

Abstract

Micro-expressions are spontaneous, unconscious movements that reveal true emotions. Accurate facial movement information and network training learning methods are crucial for micro-expression recognition. However, most existing micro-expression recognition technologies so far focus on modeling the single category of micro-expression images and neural network structure. Aiming at the problems of low recognition rate and weak model generalization ability in micro-expression recognition, a micro-expression recognition algorithm is proposed based on graph convolution network (GCN) and Transformer model. Firstly, action unit (AU) feature detection is extracted and facial muscle nodes in the neighborhood are divided into three subsets for recognition. Then, graph convolution layer is used to find the layout of dependencies between AU nodes of micro-expression classification. Finally, multiple attentional features of each facial action are enriched with Transformer model to include more sequence information before calculating the overall correlation of each region. The proposed method is validated in CASME II and CAS(ME)^2 datasets, and the recognition rate reached 69.85%.

Key words: micro-expression recognition, graph convolutional network (GCN), action unit (AU) detection, Transformer model

0 Introduction

Human facial expressions are very rich. Through the analysis of micro-expressions, people's inner emotional activities can be understood. Micro-expressions are unconscious movements of the face, and the duration is very short, between 0.04 s and 0.2 s. Different from macro expressions, which may mislead human emotion recognition, micro-expressions are mostly unconscious expressions. A series of special expressions have very significant application prospects and values in daily life. Therefore, micro-expression recognition has become one of the important areas of research[1-2]. The widespread application of deep learning has made significant progress in micro-expression recognition. However, due to problems such as difficulty in data capture and small samples, the current micro-expression recognition still has great difficulties when faced with small sample datasets. Applying the training model to the dataset test will cause the capability of micro-expression recognition to decrease. To solve this problem, it is essential to build a special network for small sample models, so the research of micro-expression recognition has become a hot issue.

Traditional methods mainly include histogram of oriented gradient[3], local binary pattern[4], local binary patterns from three orthogonal planes (LBP-TOP)[5] when extracting micro-expression features. As well as the combination of LBP-TOP features and light flow, the above methods have very good results in the extraction of salient feature points, but they all have the problem of relatively simple feature description. Because of the particularity of micro-expression recognition, the above methods cannot precisely and quickly refine the original characteristics of micro-expression. Simultaneously, micro-expression is based on video frame sequences[6]. The complete micro-expression usually contains a lot of video sequences. The streaming method can complete the feature tracking of two adjacent frames, but it cannot complete the extraction operation of dozens of video frame sequences. Therefore, the application of these methods for micro-expression recognition cannot achieve the expected results. Before

① Supported by Shaanxi Province Key Research and Development Project (2021GY-280) and the National Natural Science Foundation of China (No. 61834005, 61772417, 61802304).

② To whom correspondence should be addressed. E-mail: wujin1026@126.com.

Received on Aug. 9, 2022

2015, people used traditional image recognition algorithms for feature extraction and recognition of micro expressions. It was not until 2016 that deep learning was applied to the recognition algorithm of micro-expressions. Buhari et al. [7] used deep learning to exercise the representation of spatial and temporal features of micro-expression recognition, and spatial characteristics were encoded by convolutional neural network (CNN) and temporal characteristics were encoded by long short-term memory (LSTM). Li et al. [8] used Euler motion amplification method to amplify vertex pictures and fine-tuned pattern for further identify. Peng et al. [9] put forward a two-stream spatiotemporal network to capture the time and space information of micro-expressions, in which the spatial stream extracted the spatial information of vertex frames by using the ResNet-10 network, however the temporal flow extracted the information by using the LSTM network. Pan et al. [10] proposed a dynamic segmentation sparse imaging module. Segmented motion participating in the spatiotemporal network can capture the long-distance spatial relationship of facial micro-expression and enhance the robustness to feature-level subtle motion changes.

Although the existing algorithms have achieved certain performance improvements in the research of micro-expression recognition, the recognition rate of micro-expression still can be improved. Graph convolution network (GCN), as an extension of CNN, can process complex graph structure data, and has reached favorable application effects in terms of computer vision[11-13]. Kipf and Welling[14] proposed GCN in 2017, which provided a new idea for the processing of graph structure data, and applied action unit (AU) image CNN commonly used in deep learning to graph data. Then, the GCN-based method was used for facial AU detection, but no application of micro-expression recognition was found. Graphic structure refers to the connection of face identifiers to take shape a construction including node values and edge weights, it is possible to represent information about the texture features surround the nodes and information about the geometry variations among these nodes[15]. However, the parameter values of the graph data are manually calculated, which makes it impossible to apply to any problems. Thence, a network named micro-expression recognition graph convolutional networks (MER-GCN) is designed in this work. The parameter values of graph data, including node and edge weights, can be obtained through training, so that the final graph representation result can be obtained. The proposed facial image structure fused with shape information can better analyze facial muscle movement information and is more discriminative.

1 Relevant work

1.1 Face AU detection

AU is a basic movement that reflects the movement of facial muscles. AU is an observable part of facial movement, in which different combinations of subtle facial movements are associated with changes in facial expression. Different types of micro-expressions can be represented by different combinations of AU, as shown in Fig. 1. According to the statistical calculation of facial anatomical information, different combinations of AU have a strong relationship with different facial micro-expressions. Due to certain deviation of label information in the current micro expression dataset, mapping the extracted feature information to the corresponding emotion class is a difficult issue in micro expression recognition. Although AU detection cannot directly reflect emotion categories, it is the key to achieve efficient micro expression recognition tasks.

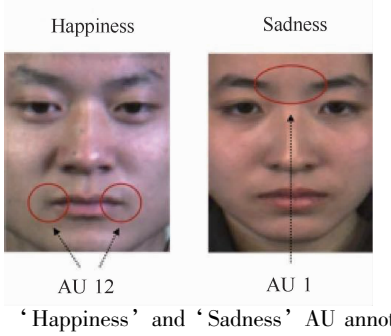

Fig. 1 'Happiness' and 'Sadness' AU annotation instructions in CASME II dataset

Research believes that recognizing well-defined facial muscles (i.e. AU) can reduce the recognition error rate, which is better than the recognition rate of discrete emotion categories. Therefore, many AU detection models extract facial texture information according to general characteristics in image processing assignments. In this paper, by introducing GCN into AU detection model, it can learn to enrich facial features to catch motion information and obtain good testing performance[16].

1.2 Network architecture design for micro-expression recognition

The annotated data of the micro expression recognition task sometimes contains deviations, so even if the trained CNN structure obtains effective frame features, the mapping process is still very difficult. The purpose of facial AUs is to observe the physical movement of facial muscles, so it is objective. Using this capability, a network architecture for AU detection is introduced, which

recognizes micro expressions on the basics of the assumption that facial muscles movements are continuous. The structural model of the micro expression recognition GCN used in this paper consists of 9 layers of graph convolutional units. The number of output channels of the three parts are 64, 128 and 256 respectively. The proposed MER-GCN architecture not only obtains image features but also captures the hidden physical relationships between facial muscles.

The output layer of MER-GCN is used to learn AU information of facial muscles by spread function. The construction of adjacency matrix is an important step in the construction of GCN. In research, the common appearance of per pair of AUs in this dataset is used as the correlation to construct the adjacency matrix in the way of data mining. In order to contain the occultation information between different AUs, stacked GCNs are used to input X from the original coding node to obtain the unified representation H^L. The learned information expressed by AU is applied to the sequence-level features extracted by GCN, and then the obtained vector is input into a full connection layer to take shape the ultimate identification consequence.

The GCN framework designed in this paper is shown in Fig. 2. The network includes three layers of models. The first layer is graph filtering based on content analysis and statistical learning methods. In the second layer, GCN layer is used to extract features from images with different parts and similarity sizes. The third layer is the fusion of image features and time features. The micro-expression image sequence was used as input, and image features were extracted by GCN layer after graph filtering for each frame. Finally, feature vectors with output length of 256 were obtained by graph coarsening. Experimental results show that the proposed framework can effectively reduce the noise interference and obtain better facial details under the premise of maintaining high resolution, and it can easily complete the fusion processing of image and video sequences. At the same time, it has strong robustness and anti-interference.

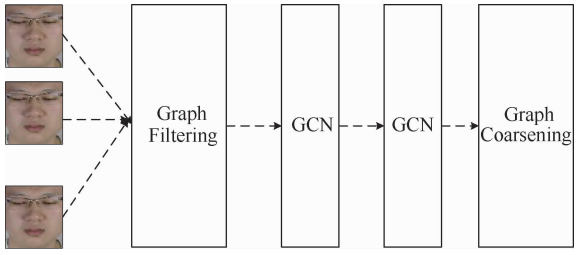

Fig. 2 GCN structure for micro-expression recognition

1.3 Subset division of facial muscle nodes

In micro-expression recognition, it is usually necessary to find more distinguishing characteristics. Researches in Ref. [17] show that eyebrows and mouth regions make significant contributions to micro-expression recognition when extracting features from specific facial area. As a result, more attention is paid to key areas like the eyebrows and mouth. However, this type of region division is still very difficult for micro-expression recognition. For the purpose of refining, a small window based on landmarks is used in this experiment to locate features. Since the GCN processes graph structure data, it is necessary to convert micro-expression sequence images into undirected graphs. Each node represents the kinematic relationships within a muscle flock, and an edge represents the kinematic relations between these muscle flocks. Changes in micro expression cause facial muscles to move. For micro expressions with different emotional types, facial muscles have different movement patterns, node values and edge weights are also different.

In GCNs for micro-expression recognition, it is very important to design the partition rules of facial muscle nodes as label graphs. There are many kinds of division structures of graph data, such as single division, spatial structure division, distance division and so on. Due to the spatial locality of the face, the movement of facial muscles is closely related to micro-expression. For example, the contraction of the levator muscle of the upper eyelid, the attempt of the upper eyelid to lift, the contraction of the orbicularis oculi muscle, and the tightness of the lower eyelid can be judged as angry micro-expression. Therefore, in this paper, the facial muscle nodes in one neighborhood were divided into three subsets, including: (1) the root node itself; (2) the centripetal node set is closer to the key area of the face than the root node; (3) centrifugal node set, the root node is farther from the adjacent nodes of the key area of the face. The so-called key area of the face is the average coordinate of all facial muscle nodes of the face.

In the network structure design of this paper, the label map is obtained by analyzing 1 neighborhood node, and other neighborhood ranges (such as 2 neighborhoods and 3 neighborhoods) can also be used for dividing the node set to contain the micro-expression tag map.

2 Transformer model for micro-expression recognition

Vaswani et al. [18] pointed out that the performance index of Transformer model was better than cyclic neural network model verification and CNN model in machine translation tasks. And it requires less computational resources in the training process. While the Transformer model has become a de facto standard for natural language processing tasks, its use in computer vision is still limited. Recent studies show that the pure Transformer model, which is fully applied to image sequences, can complete the task of image classification well. When a large amount of data is pre-trained and transferred to multiple medium and small image recognition datasets for preliminary testing, Transformer achieves good results compared with the most advanced convolutional networks, and the training requires less computational resources. Based on the GCN, this paper introduces a Transformer model for micro-expression recognition, which consists of an Encoder group and a Decoder group, both of which are stacked by multiple modules. Each module consists of multiple attention layers and fully connected feedback layers. The Transformer model uses position insertion to add relative position information to all elements of the input sequence and then to the image sequence, where each input microexpression signal is represented as a vector. Multi-head attention can be input by mapping multiple different linear transformations. The calculation is shown in Eq. (1).

$$Attention(\boldsymbol{Q}, \boldsymbol{S}, \boldsymbol{V}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{S}^T}{\sqrt{d_s}}\right)\boldsymbol{V} \qquad (1)$$

where, $\boldsymbol{Q}$, $\boldsymbol{S}$ and $\boldsymbol{V}$ represent Query, Source and Attention Value vectors in attention, respectively. In the Attention module composed of Encoder and Decoder, the Query vector is obtained by decoding, while the Source and Attention Value vector are obtained by encoding. In Eq. (1), the weight of each vector in attention is redistributed by Softmax function. The larger the weight value is, the more concentrated the extraction of feature vectors is.

Based on the GCN, this paper introduces a Transformer model for micro-expression recognition to encode the specific position information of the face. It consists of an Encoder group and a Decoder group, which are stacked by multiple modules. Each module consists of multiple attention and fully connected feedback layers. The Transformer model uses position insertion to add relative position information to all elements of the input sequence and then to the image sequence, with each input micro-expression diagram signal represented as a vector. The Transformer model of GCN is introduced to model micro-expression sequences using different network structures, which improves the ability of the model to capture multi-dimensional facial motion information of micro-expression sequences. Similar to the standard Transformer model, the model introduced by GCN also consists of multiple stacks of the same layers. Fig. 3 shows the micro-expression recognition structure of Transformer model introduced in this paper.

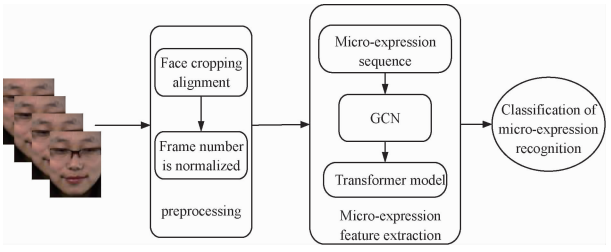

Fig. 3 Micro-expression recognition framework based on GCN and Transformer model

As shown in Fig. 4, Transformer model is composed of two parts, Encoder and Decoder respectively, and its internal parts are unified by multiple modules. The encoder maps the input micro-expression image sequence $(x_1, \cdots, x_n)$ to another sequence $z = (z_1, \cdots, z_n)$ of continuous multiple representations. Given z, the decoder produces an output class sequence $(y_1, \cdots, y_m)$ identical to the various elements in each sequence, and at each step, the model regents automatically, consuming the previously generated sequence image as additional input when generating the next step. Transformer adopts this overall structure, using stacked multi-headed attention and fully connected feedback layers, as shown in the left and right halves of Fig. 4. The encoder in the improved Transformer model in this paper is composed of $N = 6$ identical data layers stacked, each layer has two sub-layers, one data operation sublayer is multi-attentional mechanism, the other is a simple, point-by-point basic connection feedforward network. The output of each Sublayer is a LayerNorm$(x + \mathrm{Sublayer}(x))$, where Sublayer$(x)$ is an internal function call implemented by the Sublayer function itself. To facilitate these residual connections, all the sub-layers and embedding layers in the modeling will produce outputs with dimensions $d_{\mathrm{model}} = 512$. The decoder also includes $N = 6$ identical layers superimposed on each other. In addition to the two sub-layers of each encoder layer, the decoder adds a third sublayer that takes a long time to view the output of the encoder stack. Similar to the encoder, residual tight connections are used around each sub-layer, followed by layer normalization.

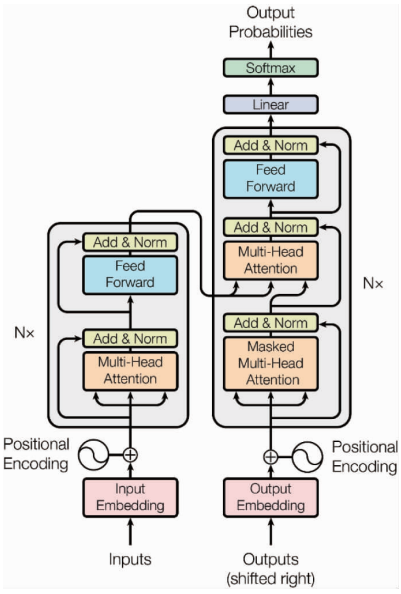

Fig. 4 Transformer model

3 Multi-scale space feature fusion and multi-channel convolutional process

Spatial pyramid pooling (SPP) is a common multi-scale spatial feature extraction method in the field of image segmentation and target detection, which is not limited by the input image size. Pyramid pooling uses pooling layers of different sizes to collect the output features of the graph filter layer, so as to obtain the feature maps of different receptive fields.

In the algorithm design of this paper, a spatial pyramid pooling layer is added between the fully connected layer and the coarse-layer of GCN to extract the multi-scale features of the spatial level of facial actions. pool_1, pool_2 and pool_3 further fused the main spatial multi-scale features extracted through pooled receptive field windows of different sizes, thus improving the spatial feature extraction capability of GCN. Spatial pyramid pooling has two advantages: first, it solves the problem that the dimensions of the node graph of facial micro-expression are inconsistent with the input required by the GCN; second, features are extracted from different angles of the same node graph of facial micro-expression. The operation model of the pooling layer of the pyramid of collection space is shown in Fig. 5.

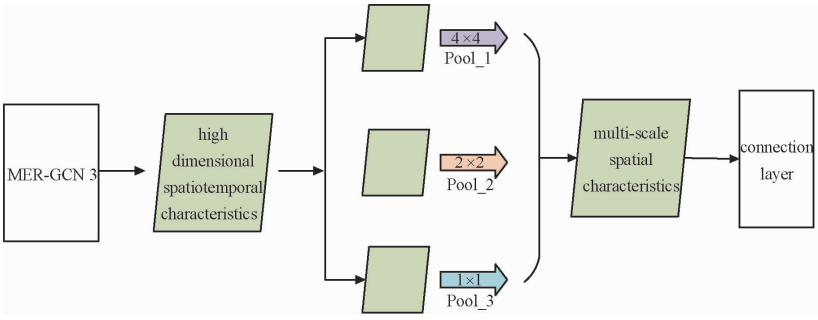

Fig. 5 Pooling layer of spatial pyramid melted by multi-scale spatial features

It is assumed that the extended MER-GCN 3 layer of the model generated feature maps of different sizes after the pooling operation of spatial pyramid, in which the feature map was 128 bits. For the high-dimensional spatio-temporal features, pool_1, pool_2 and pool_3 are pooled at different scales, respectively. Finally, the combination of three pooled feature maps is mapped. It can input different dimensions of facial micro-expression muscle node graph through multiple extended MER-GCN layers, maintain the same feature dimension through spatial pyramid pooling layer, and then input full connection layer for recognition. Assuming that the size of the high-dimensional space-time characteristic graph of the input is $a \times a$, and the size of the pooled

output of the pyramid convolution operation is $n \times n$, the size and step size of the pooled operation can be expressed as Eqs(2) and (3).

$$size = \lceil a/n \rceil \tag{2}$$

$$stride = \lfloor a/n \rfloor \tag{3}$$

Different from other algorithm designs, this paper extends single-channel convolution to multi-channel convolution, and GCN uses multiple convolution kernels to extract richer features. Assume that the input graphic signal is $X \in \boldsymbol{R}^{H \times W \times C}$, where C is the number of channels, and the length H and width W of the convolution kernel are k. Since multi-channel convolution is used here, and each channel has a convolution kernel of the same size $k \times k$, multi-channel convolution can be defined as Eq. (4).

$$H_{m,n,C'} = \sum_{i=-\lfloor \frac{k}{2} \rfloor}^{\lfloor \frac{k}{2} \rfloor} \sum_{j=-\lfloor \frac{k}{2} \rfloor}^{\lfloor \frac{k}{2} \rfloor} X_{m+i,n+j} \cdot G_{i,j}^{C'} \tag{4}$$

where, $X_{m+i,n+j} \in R^C$, $G_{i,j}^{C'} \in R^C$, $H_{m,n,C'} \in R^{H' \times W'}$.

The improved multi-channel convolution process in this paper is shown in Fig. 6. During graph coarsening, the values of corresponding positions on the feature graph are obtained through tensor dot product operation at each sliding position, and the output of a single channel is finally obtained, and then the output values of multiple channels are added.

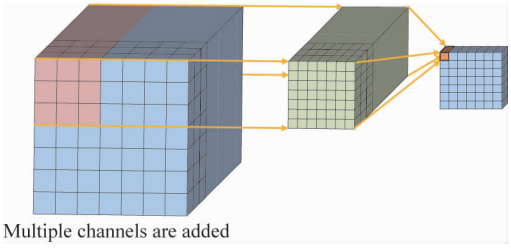

Fig. 6 Multi-channel convolution process

Similar to single-channel convolution, multi-channel convolution uses multiple same convolution kernels to extract richer features. Assuming that the dimension of convolution kernels is $R^{k \times k \times C \times C'}$, where C' represents the number of convolution kernels, the output of multi-channel convolution is $H \in R^{H' \times W' \times C'}$ by combining the input static image data with the results of multiple convolution kernels. After the completion of the convolution operation, a bias is usually added to each feature graph, as shown in Eq. (5).

$$H_{m,n,c'} = b_{c'} + \sum_{i=-\lfloor \frac{k}{2} \rfloor}^{\lfloor \frac{k}{2} \rfloor} \sum_{j=-\lfloor \frac{k}{2} \rfloor}^{\lfloor \frac{k}{2} \rfloor} X_{m+i,n+j} \cdot G_{i,j}^{c'} \tag{5}$$

In the complete process of multi-channel convolution mentioned above, in general, the input image $H \in R^{H \times W \times C}$ is convolved with the convolution kernel $R^{k \times k \times C \times C'}$, and the output image $H \in R^{H' \times W' \times C'}$ is obtained by adding bias. In this process, two parameters, convolution kernel and bias, are introduced, and the total number of parameters involved is $k^2 \times C \times C' + C'$.

4 Experiment and result analysis

4.1 Experimental environment

The experimental parameters in the research process are shown in Table 1.

Table 1 Hardware and software environment

Name	Model(version)	Description
Operating system	Linux	Ubuntu 18.04
CUDA	CUDA10.1	The underlying software platform for GPU acceleration
PyTorch	1.0.0	Dynamic graph characteristics
GPU	GTX 1080Ti	Video memory 11 GB
RAM	Kingston HyperX Savage DDR4	Main frequency 2400 MHz and 8 GB
Hard disk	Seagate	1 TB

Network structure: because the single feature input affects the network model to learn information from multiple features, the recognition accuracy is not high. In order to learn multiple features to promote the recognition accuracy of the network model, the GCN structure constructed in this paper for micro-expression recognition uses dual features as input, one is facial features representing static features, the other is facial muscle motion features of micro-expression sequence representing motion features.

The network structure in this model consists of nine layers of graph convolution units, and the cascade of graph convolution is used to construct the single-stream network. The first image convolution layer inputs the micro-expression sequence image for input operation, and passes the generated graph data to subsequent layers.

Parameter configuration: the number of frames sampled from the video is 150. Regularize the number of video image frames; set the batch size as 32; set stochastic gradient descent (SGD) as model optimizer; set the primary learning rate as 0.001; set the learning rate decay as le-10; set momentum to 0.9; set the number of epochs to 400. When the quantity of iterations continues to increase, in order to effectively avoid the divergence of the loss function, the number of iter-

ations is particularly limited. When the number of iterations reaches 20 000, the learning rate decays to one-half of the original.

4.2 Experimental results and analysis

This paper uses the torch. save function in PyTorch to save the parameter dictionary of the model when recording the experimental data. When saving the model for inference, only the learning parameters of the trained model need to be saved, so that the model can be restored later. Before running, call model. eval to set the batch normalization layer to evaluation mode. The method of reading the training data is the same as that of calling the pre-training parameters, using the load_state_dict function. This paper saves the training data into a CSV file, reads the CSV file, and finally draws a graph of the training process.

According to the overall network structure design, the network is trained and tested based on the deep learning PyTorch framework. The training set for network training uses the augmented data of the CASME II and CAS (ME)^2 datasets. In order to better recognize the micro-expression, this experiment uses alignment and image size adjustment, and finally adjusts the processed dataset image size to 256, which meets the input requirements of image processing. Because of the unbalanced and small sample dimension of the dataset, it is necessary to perform data enhancement for the two datasets used. First, crop the image, zoom in to 256, and use the four angles and the core of the image as the cutting out core respectively. The network configuration parameters during training are determined after multiple experiments based on various factors such as the experimental software and hardware environment. The experiment in this paper performed 400 epochs to record the recognition accuracy and its loss function value in detail.

As shown in Fig. 7 and Fig. 8, the accuracy curve and loss function curve of the proposed algorithm trained on CASME II dataset are respectively. When the number of epochs is 240, the network begins to be stable. For different types of training samples, the accuracy is not the same, and two cases can be roughly distinguished: (1) with the increase of training times, the accuracy will gradually improve; (2) when the training times reach a certain value, the accuracy will stop declining.

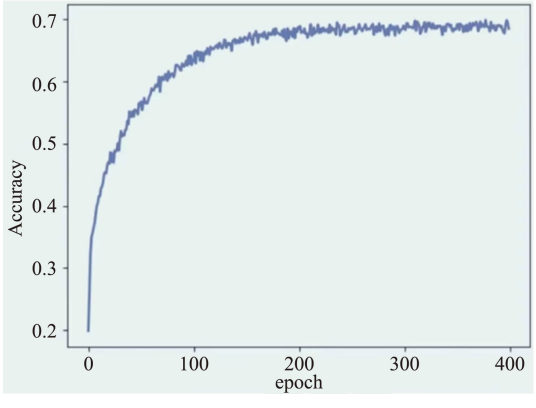

Fig. 7 Accuracy curve during CASME II training process

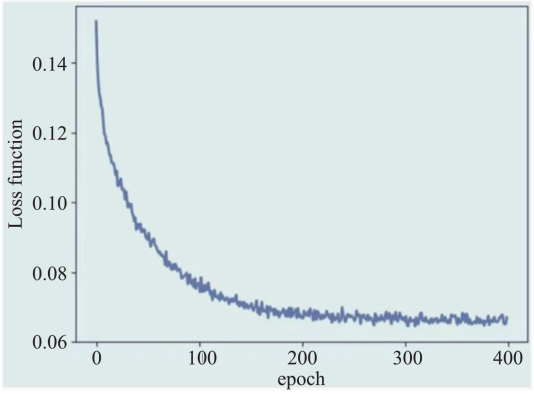

Fig. 8 Loss function curve during CASME II training process

As shown in Fig. 9 and Fig. 10, the accuracy curve and loss function curve of the proposed algorithm trained on CAS(ME)^2 dataset are respectively. According to the experimental result graph, as the number of network iterations increases, the convergence speed and convergence status of the network are stable. According to the experimental results, it can be concluded that the recognition effect on the CAS(ME)^2 dataset is better, which is mainly related to the quantity and quality of the dataset samples.

After the network training is completed, for the purpose of verifying the model is good or bad, the trained model needs to be tested, and the test data uses specimen from the original dataset. Through the network test, the confusion matrix on the CASME II and CAS(ME)^2 datasets are obtained, as shown in Fig. 11 and Fig. 12 respectively. The results of the confusion matrix show that the model is most sensitive to the other classes in CASME II, with 86% accuracy, and less accurate when identifying expressions of repression and disgust. In CAS(ME)^2, the model has a good classification ability for positive and negative micro-expressions, and the recognition rate of surprise expression is up to 76%, but the recognition rate of depression microexpressions is low. This may be due to too few training samples in these categories, resulting in insufficient dynamic feature extraction capabilities, and the imbalance of training samples will also lead to different recognition rates.

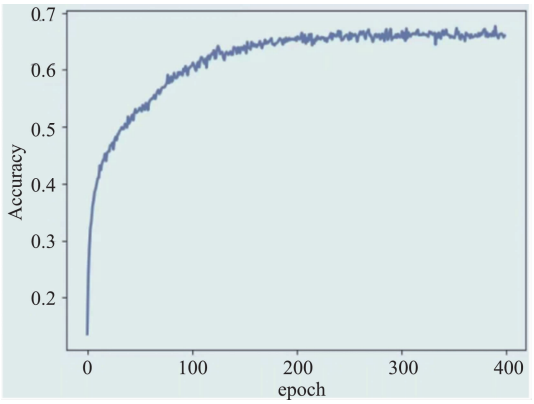

Fig. 9 Accuracy curve during CAS(ME)^2 training process

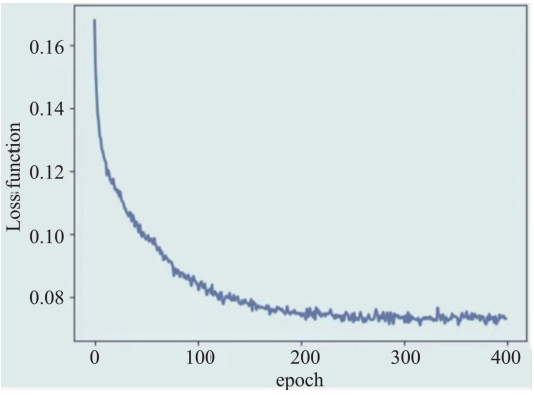

Fig. 10 Loss function curve during CAS(ME)^2 training process

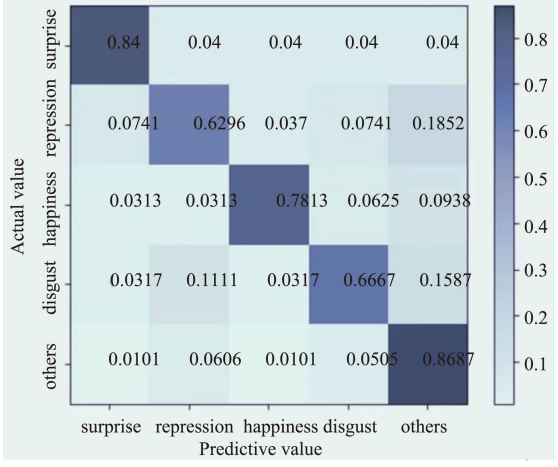

Fig. 11 Confusion matrix in CASME II

To validate the model, two different methods are used: leave-one-subject-out (LOSO) validation and K-fold intersect verification. Regarding LOSO verification, all the data of one category are randomly left for verification in each training process, so as to reduce the chance of category deviation of micro-expression theme. For K-fold verification, the dataset is divided into K parts, and a random part is used for validation in different training runs. Both strategies prevent different types of deviations in the model.

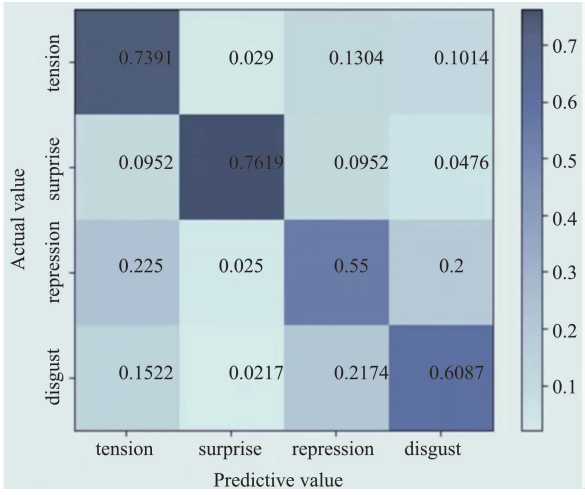

Fig. 12 Confusion matrix in CAS(ME)^2

As shown in Table 2, the ablation experiment is conducted to verify the effectiveness of the algorithm designed in this paper. The recognition rate of GCN or CNN combined with Transformer model is low, but the recognition rate of GCN combined with Transformer model structure is high, which proves the effectiveness of the algorithm in this paper.

Table 2 Ablation expriments

Structure method	CASME II/%	CAS(ME)^2/%
GCN	59.65	56.25
Transformer model	62.35	60.26
Proposed algorithm	69.85	69.11

The model trained by GCN combined with Transformer model is used to train each category on CASME II and CAS(ME)^2 datasets, and the recognition rate results are shown in Fig. 13. It can be seen that the recognition rate of other categories in the CASME II dataset is the highest at 0.86, while the recognition rate of disgust, depression and surprise categories is slightly higher than that of CAS(ME)^2. This is because CASME II has developed a stricter general standard for direct sample selection. Addresses expressionless facial movements based on participants' self-reported emotions. In addition, the sample size of other types of micro-expressions is the largest, indicating that the larger the sample size, the better the results of network training test.

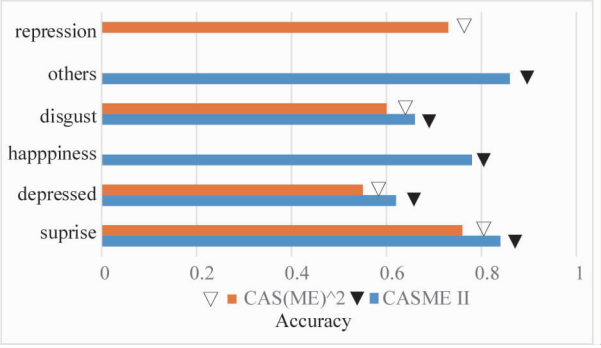

Fig. 13 Classification recognition rate of microexpression dataset

The experiment evaluated the GCN on CASME II and CAS(ME)^2 datasets. In order to quantitatively evaluate the designed network structure, the performance of micro-expression recognition is measured through classification accuracy. As shown in Table 3, it is a comparison of recognition rates on three different micro-expression datasets. It can be concluded that the recognition influences of the MER-GCN network on the CASME II dataset is better. This is because more stringent sample selection criteria have been established in the current database, and the participants' self-reported emotions are eliminated. Facial movements have no emotional significance.

Table 3 Comparison of the recognition rate of MER-GCN on different training datasets

Method	Training dataset	Accuracy/%
MER-GCN	CASME	66.47
MER-GCN	CAS(ME)^2	69.11
MER-GCN	CASME II	69.85

The recognition accuracy of the micro-expression recognition algorithm in this paper is compared with other deep learning algorithms, which is used to test the effectiveness of GCN structure, multi-channel convolution and multi-scale space fusion. The recognition accuracy of CAS(ME)^2 dataset is evaluated. As shown in Table 4, 2D-CNN is a micro-expression recognition algorithm proposed by Mayya et al.[19] combining temporal interpolation model (TIM) and CNN, and 3D-CNN is a three-dimensional CNN firstly designed by Ji et al.[20]. It can be concluded that the accuracy of network model is higher than other models. The recognition rate of the micro-expression recognition algorithm based on GCN and Transformer model in CAS(ME)^2 dataset is 69.85%, which is 9.6% higher than that of S-LRCN algorithm, 5% higher than that of 2D-CNN, 1.39% higher than that of 3D-CNN, and 0.69% higher than that of GCN. It realizes the effective extraction of micro-expression sequence information and excellent recognition accuracy.

Table 4 The accuracy of the algorithm in this paper is compared with other algorithms

Algorithms	Accuracy/%
S-LRCN	60.25
2D-CNN[19]	64.85
3D-CNN[20]	68.46
GCN[21]	69.16
Proposed algorithm	69.85

5 Conclusion

In view of the low detection rate of current micro-expression recognition algorithms, especially the small amount of sample data, this paper proposes a micro-expression recognition algorithm based on GCN and Transformer model. Based on GCNs, Transformer models are introduced to enrich the location features of each facial action, thus containing more sequence information before calculating the overall correlation of each region. By changing the size of learning parameters, the risk of overfitting is reduced. Finally, the constructed MER-GCN network is trained and tested on CASME II, and CAS(ME)^2 datasets. Laboratory results show that superimposed additional information of GCN can promote the recognition of micro-expressions, and the maximum recognition rate of experimental results reaches 69.85%, which verifies the feasibility of the algorithm. Although this paper ameliorates the problems of dataset scarcity and low recognition rate in micro-expression recognition, the work can be extended by adding data in the future through integrated strategies, thus increasing the size of the dataset and building deeper and more complex models.

References

[1] LI J, WANG Y, SEE J, et al. Micro-expression recognition based on 3D flow convolutional neural network[J]. Pattern Analysis and Applications, 2019, 22(4):1331-1339.

[2] HE J, HU J F, LU X, et al. Multi-task mid-level feature learning for micro-expression recognition[J]. Pattern Recognition, 2017, 66(3):44-52.

[3] LI Y, HUANG X, ZHAO G. Joint local and global information learning with single apex frame detection for micro-expression recognition[J]. IEEE Transactions on Image Processing, 2021, 30(1):249-263.

[4] ZHU W, CHEN Y. Micro-expression recognition convolutional network based on dual-stream temporal-domain information interaction[C]//2020 13th International Symposium on Computational Intelligence and Design (ISCID).

Hangzhou: IEEE, 2020: 396-400.
[5] NING J, WANG R. Application of psychological analysis of micro-expression recognition in teaching evaluation[C]// International Conference on Education Studies: Experience and Innovation (ICESEI 2020). Paris: Atlantis Press, 2020: 71-77.
[6] BEN X, JIA X, YAN R, et al. Learning effective binary descriptors for micro-expression recognition transferred by macro-information[J] Pattern Recognition Letters, 2018, 107(5): 50-58.
[7] BUHARI A M, OOI C P, BASKARAN V M, et al. FACS-based graph features for real-time micro-expression recognition[J]. Journal of Imaging, 2020, 6(12): 130.
[8] LI Y, HUANG X, ZHAO G. Can micro-expression be recognized based on single apex frame? [C]// 2018 25th IEEE International Conference on Image Processing (ICIP). Athens: IEEE, 2018: 3094-3098.
[9] PENG M, WANG C, BI T, et al. A novel apex-time network for cross-dataset micro-expression recognition[C]// 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). Cambridge: IEEE, 2019: 1-6.
[10] PAN H, XIE L, LV Z, et al. Hierarchical support vector machine for facial micro-expression recognition[J] Multimedia Tools and Applications, 2020, 79(3): 1-15.
[11] SU W C. Facial activity unit detection and micro-expression analysis[J]. Beijing: Beijing University of Posts and Telecommunications, 2019. (In Chinese)
[12] NIE X, TAKALKAR M A, DUAN M, et al. GEME: dual-stream multi-task GEnder-based micro-expression recognition[J] Neurocomputing, 2021, 427(5): 13-28.
[13] JIN D, XU Z, HARRISON A P, et al. 3D convolutional neural networks with graph refinement for airway segmentation using incomplete data labels[C]// International workshop on machine learning in medical imaging. Quebec City: Springer, 2017: 141-149.
[14] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J]. (2017-02-22) [2022-08-09]. https://arxiv.org/pdf/1609.02907.pdf.
[15] ABOUSALEH F S, LIM T, CHENG W H, et al. A novel comparative deep learning framework for facial age estimation[J]. EURASIP Journal on Image and Video Processing, 2016, 2016(1): 1-13.
[16] ZHU W J, CHENG Y. Recognition of micro-expression under the information interaction mechanism of dual-stream network[J] Journal of Computer Aided Design and Graphics, 2021, 33(4): 545-552.
[17] XIAO R X. Research on micro-expression recognition method based on macro-expression transfer learning[J] Jinan: Shandong University, 2020. (In Chinese)
[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//The 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach: Curran Associates Inc, 2017: 1-15.
[19] MAYYA V, PAI R M, PAI M M M. Combining temporal interpolation and DCNN for faster recognition of micro-expressions in video sequences[C]// International Conference on Advances in Computing, Communications and Informatics. Jaipur: IEEE, 2016: 699-703.
[20] JI S, XU W, YANG M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1): 221-231.
[21] MA Z, MEI G, PREZIOSO E, et al. A deep learning approach using graph convolutional networks for slope deformation prediction based on time-series displacement data[J]. Neural Computing and Applications, 2021, 33(21): 14441-14457.

WU Jin, born in 1975. She received her B. S degree from Xi'an Jiaotong University in 1998, and she also received her M. S. degree from Xi'an Jiaotong University in 2001. Her research focuses on key techniques for signal and information processing.