Multimedia Systems (2023) 29:3069–3079
https://doi.org/10.1007/s00530-022-00907-9

SPECIAL ISSUE PAPER

Robust facial expression recognition with global-local joint representation learning

Chunxiao Fan[1,2,3] · Zhenxing Wang[1,3] · Jia Li[1,3] · Shanshan Wang[4] · Xiao Sun[1,3]

Received: 26 August 2021 / Accepted: 15 February 2022 / Published online: 7 March 2022

Abstract
As an important part in computer vision, facial expression recognition (FER) has received extensive attention, but it still has lots of challenges in this area. One of the important difficulties is to remain the topological information in the feature extraction operation. In this paper, we propose a novel facial expression recognition method with lite dual channel neural network based on graph convolutional networks (DCNN-GCN). In the proposed method, (1) the topological structure information and texture feature of regions of interest (ROIs) are modeled as graphs and processed with graph convolutional network (GCN) to remain the topological features. (2) The local features of ROIs and global features are extracted with dual channel neural networks, which can improve the performance of features extraction and reduce the complexity of networks. The proposed method is evaluated on CK+, Oulu-CASIA and MMI data sets. Experiment results show that the proposed method can significantly improve the accuracy of facial expression recognition. In addition, the network is much lite and suitable for application.

Keywords Facial expression recognition · Dual channel neural network · Graph convolutional network

1 Introduction

Facial expression is the intuitive response of human inner emotion, so it is a very important way to analysis the emotion and intention. Nowadays, facial expression recognition (FER) has played a crucial role in lots of applications, such as human–robot interaction systems [1], driver-assistance system [2], and detection of neurological disorders [3]. It has been an important research area for decades, and lots of famous methods have been proposed.

The studies of facial expression analysis can be traced back to the work of Ekman et al. [4]. They studied and summarized six basic human expressions (anger, disgust, fear, happiness, sadness, and surprise). Then many methods tried to model the facial expression recognition as a classification problem, and accomplish it based on facial features and machine learning algorithms.

Some works recognize different expressions based on the features extracted from whole facial image, such as Principal Component Analysis (PCA) [5], Independent Component Analysis (ICA) [6], and Fisher Linear Discrimination (FLD) [7]. Some methods divides the face into several parts, and give more attention on the parts with higher importance. The features extraction methods include Facial Action Coding System (FACS) [8], Gabor [9], Local Binary Patterns (LBP) [10, 11] and so on.

But the methods above are based on manual features and easily interfered by human factors. In recent years, deep learning has shown great information processing ability and better robustness, which does not rely on the accurate design of manual features [12, 13]. Many works propose to introduce the classical network structures to facial expression recognition, such as [14–16]. To improve the performance, some works propose to fuse multiple kinds of features for a comprehensive representation [17–19]. Although the accuracy of recognition is

✉ Xiao Sun
sunx@hfut.edu.cn

1 Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education, Hefei University of Technology, Hefei 230009, Anhui, China

2 Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, Anhui, China

3 School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, Anhui, China

4 Institute of Physical Science and Information Technology, Anhui University, Hefei 230039, Anhui, China

improved, the complexity of the networks in these methods also increases.

As the expressions are caused by motion of facial muscles, the features have structure and topological properties. However, few methods have take the topological structure features into recognition methods, and it is very hard to remain the topological information in the feature extraction operation. In addition, almost all these methods learn features with one network from the whole images. So the network needs to extract the global facial features and local features, which are in different forms and make the network complex.

In this paper, we propose a novel facial expression recognition method with lite dual channel neural network based on GCN (DCNN-GCN). In the proposed method, the features are extracted by two channels, which focus on the global and local ROI features, respectively. The facial key points are detected to extract the the structure features and texture features in ROI. These features are modeled as graphs and processed by graph convolutional networks (GCN), which can aggregate the features into some higher level features with remaining the structure property. With dual channel neural networks, the performance of global and local features extraction can be more efficient and the complexity of networks can be reduced.

Our contributions can be summarized as follows.

1. The topological structure feature of human face is modeled with GCN, which can remain the topological information and extract high level structural feature and texture feature.
2. The global and local ROI features are extracted with dual channel neural networks instead of one unified network, so the performance of feature extraction increases and can make the network lite.
3. The proposed method can improve the recognition performance of expressions recognition on three widely used data sets.

The remainder of this paper is organized as follows. The relative works are introduced in Sect. 2. In Sect. 3, we give the overview and detailed description of our proposed method. Sect. 4 presents the implementation and experiment to verify the proposed algorithm. Our work is concluded in Sect. 5.

2 Relative works

Facial expression recognition can be divided into two kinds: traditional approaches and deep learning based approaches.

2.1 Traditional approaches

Traditional methods usually use manual features for facial expression recognition, such as PCA [5], ICA [6], and FLD [7]. PCA [5] is a common data analysis method, which is always used for dimensionality reduction of high-dimensional data to extract the main feature components of data. ICA [6] can effectively extract expression features with high-order statistical characteristics and analyze them from high-order correlation. FLD [7] can extract the most discriminative low dimensional features from high-dimensional features, so that the same samples can be close and different samples are separated.

To improve the accuracy of expression recognition, some complex features are adopted for facial expression recognition, including FACS [8], Gabor [9], LBP [10, 11], Local Phase Quantization (LPQ) [20], Active Shape Models (ASM) [21] and sparse learning [22]. FACS [8] defines the basic motion unit of human face according to different facial expressions, and the expressions can be divided into various basic units. Gabor [9] is a linear filter to extract edges, and it has a nice adaptability to deformation, which is very important in expression feature recognition. LBP [10, 11] has the invariance of gray and rotation, and can well describe the texture information of expression features. In [23], the multi-layer perceptron (MLP) and Support Vector Machine (SVM) are used to complete expression classification with the facial features. Desrosiers et al. [24] propose to use geometric features based on facial landmarks trajectories, and the effectiveness is verified on UvA-NEMO and CK+ datasets.

2.2 Deep learning based approaches

However, the local features extracted by these methods are easily interfered by human factors, which may result in the loss of facial expression information and lead to inaccurate classification. In recent years, the deep learning based methods have attracted more attention and achieved great success [25, 26]. Cheng et al. [14] optimize the network structure and parameters based on VGG19 [27], and adopt transfer learning to improve the accuracy of expression recognition. An efficient neural network is proposed in [15], which is based on ResNet [28] and adds SE block [29] to achieve a high accuracy. Mollahosseini et al. [16] construct a CNN structure with Inception layer, and combine facial movements for facial expression recognition. Liu et al. [30] propose a 3D convolutional neural network with variable action constraints, which can detect specific facial part actions and obtain easily distinguishable characterization features. Gan et al. [31] propose a

multi attention network to deal with the problem of facial expression recognition under complex conditions. Meng et al. [32] propose an identity-aware convolutional neural network (IACNN), and an identity-sensitive contrastive loss is used to learn identity-related information. In [33], a self-cure network (SCN) is proposed, and a self-attention mechanism over mini-batch is adopted to prevent deep network from over-fitting uncertain facial images. Yang et al. [34, 35] use GAN to generate neutral face images and take residuals to reduce the impact of identity information. Agrawal et al. [36] construct an efficient CNN model with different parameters for facial expression recognition by analyzing the size of convolution kernel and the number of filters.

To improve the robustness in complex scenes, some works take multiple types of features to get a comprehensive representation, which can improve the recognition accuracy. Hamester et al. [17] use multi-channel CNN (MCCNN) to recognize facial expressions. In MCCNN, CNN channel and Convolutional Auto-Encoders (CAE) channel share the same topology. The recognition accuracy of this method is better than the traditional feature-based method. Zhao et al. [18] propose a peak-piloted depth network (PPDN), which embeds the expression evolving process from non-peak expression to peak expression into the network. It uses a special-purpose back-propagation procedure, which are peak gradient suppression (PGS), to avoid degrading the recognition capability for samples of peak expression caused by interference from their non-peak expression counterparts. Xie and Hu [19] propose the Deep Comprehensive Multi-Patch Aggregation CNN (DCMA-CNN), which is a dual branch CNN framework. One branch is to extract global features from the complete facial expression image, and the other branch extracts local features from a set of overlapped patches. After feature extraction, the global features and local features are connected for final prediction. As the aggregation of local and global features can represent facial expressions on different scales, the recognition accuracy is better than other competitive facial expression recognition methods.

Chen et al. [37] propose an Inter-class Relational Learning method, which learns the relationship between different expressions by distinguishing the mixed features obtained from multiple features, and expands the Fisher criterion between different classes to improve the discrimination ability of different expression categories. Wang et al. [38] propose to use global and regional features and establishing Bayesian network model to improve the accuracy of expression feature classification. Xu et al. [39] propose to combine LBP and convolutional neural network, and combine the two features extracted through two branches to reduce the impact of image rotation on expression recognition. Nguyen et al. [40] propose a multi feature level fusion method based on residual network, which uses the fusion of low-level features and high-level features to help improve its accuracy. Li et al. [41] propose a CNN structure based on attention mechanism. Combined with the features of key areas of the face, each feature is weighted, so as to focus on the recognizable non occlusion area.

Although the large-scale deep learning model achieves high recognition accuracy, the limited number of samples in facial expression data sets can limit the performance. Without enough training samples, the large-scale deep learning model is prone to over fitting. In order to achieve a balance between structural complexity and recognition accuracy, researchers try to design a lightweight deep learning model with compact structure and strong feature extraction ability. Jung et al. [42] propose a lite deep temporal appearance-geometry network (DTAGN). They construct two complementary small-scale depth networks: CNN based deep temporal appearance network (DTAN) and fully connected DNN based deep temporal geometry network (DTGN). DTAN is used to extract the temporal appearance features required for facial expression recognition. DTGN obtains the geometric information of facial landmark motion. To improve the performance of facial expression recognition, a new joint fine-tuning method is used to fuse features. The quantity of parameters in this network is only $5.85M$.

2.3 Graph convolutional network

Graph convolutional network (GCN) [43] has excellent performance in graph classification, and it has been used in many areas [44]. Liu et al. [45] propose a GCN based dynamic facial expression recognition task framework, named as facial expression recognition GCN (FER-GCN), to learn more useful facial expression features and capture the dynamic change on expression. They introduce the GCN layer into a general CNN-RNN based video FER model. GCN layer is used to learn an adjacency matrix which represents the dependency of inter frame. As the features of GCN learning are concentrated in the same region, LSTM layer is applied to learn their long-term dependence. In addition, a weight allocation mechanism is designed to represent the expression strength of each frame and weight the output of different nodes. This method achieves 99.54% excellent performance on CK+ dataset, but it is a video based method. GCN has also been applied in facial micro-expression recognition. Kumar et al. [46] designed a graph that uses a triplet of frames to extract temporal information, in which a two-streams graph attention convolutional network are used and fused for classification.

3 Proposed DCNN-GCN approach

As shown in Fig. 1, the proposed method consists of two channels, Global Feature Channel and Local Feature Channel. The global features of facial expression image are extracted in the Global Feature Channel with CNN module. The Local Feature Channel is used to model the topological structure and texture features and extract higher level local features with GCN. The input image is processed to detect the facial key points, which can represent the structure information on the face. Then the structure and texture information of these points is modeled as graphs. The distances of every two key points form the adjacency matrix, and the pixel values around these key points are the attribution of each point. Aggregated by GCN, some higher level features can be extracted remaining the structure and texture property. The global and local features are set into the classifier to accomplish the expression recognition.

3.1 Global Feature Channel

In the Global Feature Channel, the input images are pre-processed firstly to reduce the noise influence, which includes face detection, uniform cutting and image normalization.

As shown in Fig. 2, five convolution units are contained in Global Feature Channel. Each convolution unit consists of a convolution layer and a maximum pooling layer. The details of this channel is listed in Table 1. The size of convolution is 3×3, and the size of pooling layer is 2×2. ReLU [47] is used as the activation function for each convolution layer. The vectorization layer transforms the multidimensional data into a global feature vector, which is convenient for the connection of the following feature vectors.

Different with the face verification task, multiple individuals may be contained in a category in the facial expression recognition. Images belonging to the same expression may have different appearances, gender, skin color and age, which results in great internal differences. To solve this problem, batch normalization [48] is adopted into the proposed method.

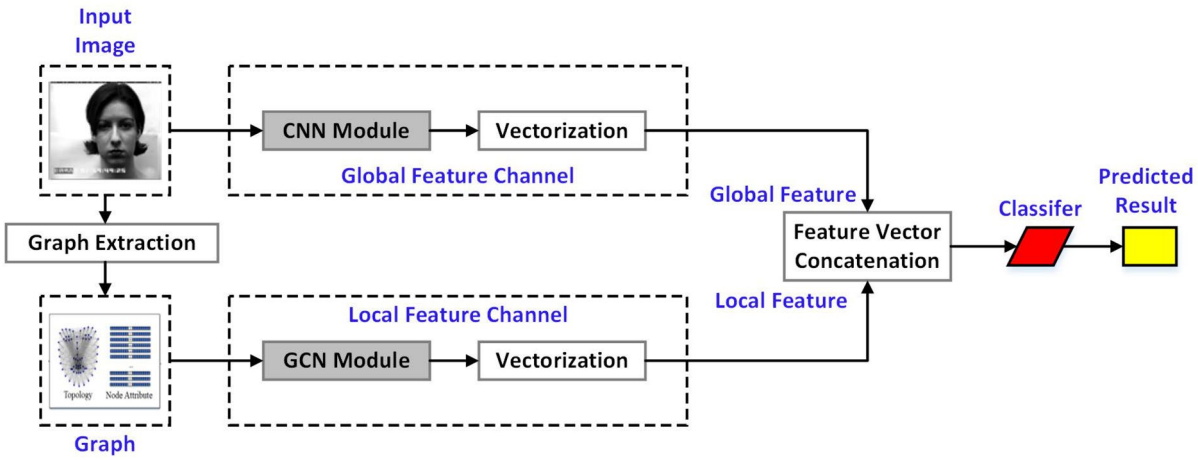

Fig. 1 The structure of our proposed method. The Global Feature Channel extracts global features from facial expression images based on CNN module, and the Local Feature Channel extracts local features from Graph based on GCN module. Global and local features are vectorized and concatenated to obtain the concatenate feature vector, which is input into the classifier for feature fusion and classification

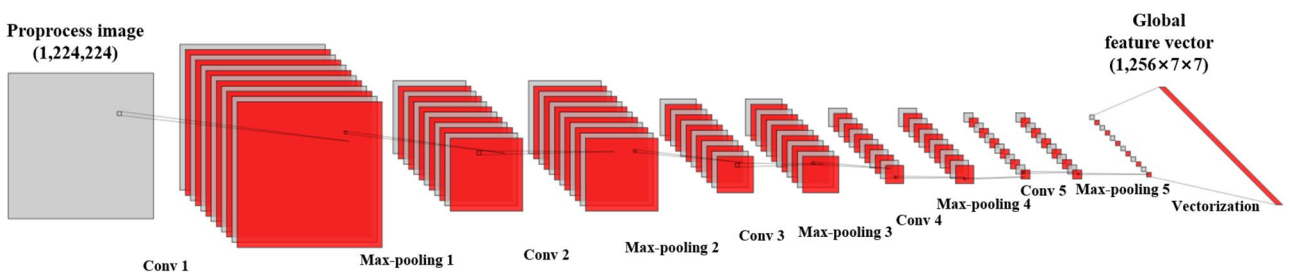

Fig. 2 The architecture of the Global Feature Channel. CNN module consists of five convolution units, and each of them contains a convolution layer and a maximum pooling layer. The vectorization layer transforms the multidimensional data into a global feature vector

Table 1 The details of Global Feature Channel

Layer	Size/Stride	Output
Data	–	(1,224,224)
Conv 1	$3\times3/1$	(32,224,224)
BN 1	–	(32,224,224)
Max-pooling 1	$2\times2/2$	(32,112,112)
Conv 2	$3\times3/1$	(64,112,112)
BN 2	–	(64,112,112)
Max-pooling 2	$2\times2/2$	(64,56,56)
Conv 3	$3\times3/1$	(128,56,56)
BN 3	–	(128,56,56)
Max-pooling 3	$2\times2/2$	(128,28,28)
Conv 4	$3\times3/1$	(256,28,28)
BN 4	–	(256,28,28)
Max-pooling 4	$2\times2/2$	(256,14,14)
Conv 5	$3\times3/1$	(256,14,14)
BN 5	–	(256,14,14)
Max-pooling 5	$2\times2/2$	(256,7,7)
Vectorization	–	$(1,256\times7\times7)$

The kernel sizes, strides, and output sizes for each layer are listed. (*Conv* convolution, *BN* batch normalization)

In the batch training, the activation of each small batch is centered on zero mean and unit variance. For M-dimensional input $X=\{x^{(1)},\ldots,x^{(m)}\}$, the regularization of each dimension is as (1):

$$\hat{x}^{(k)}=\frac{x^{(k)}-E\left[x^{(k)}\right]}{\sqrt{\mathrm{Var}\left[x^{(k)}\right]}}, \tag{1}$$

where E and Var are the expectation and variance of the input values. With batch normalization, all samples in a mini-batch are associated and trained together. Therefore, in the training process, the output of the network are not only determined by the the sample itself, but also the other samples in the same batch. As the batches are randomly selected, it can avoid over fitting to a certain extent.

3.2 Local Feature Channel

In the Local Feature Channel, the local texture feature in ROIs and topological structure feature of the whole face are extracted with GCN.

3.2.1 Construction of graph

As the facial expression is related to the motion of facial muscles, the topological structure of facial key points is a very important feature for the expression recognition. In the proposed method, the facial key points are detected firstly,

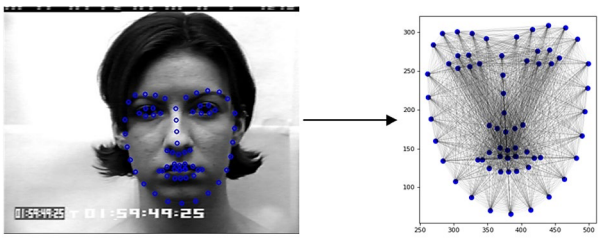

Fig. 3 The construction of graph. The graph is constructed from a facial expression image. The facial key points are detected from the expression image, which constitute the nodes of the graph. And the nodes are joined in pairs to form edges of the graph

which can be regarded as ROI. The structure and texture features of these area are modeled as a graph as shown in Fig. 3. In this graph, every two points are connected and the distance forms the weight of edge, thus a weighted adjacency matrix $A\in\mathbb{R}^{N\times N}$ is obtained, which represents the the structure information. The pixel values around these landmarks are the attribution of these nodes, resulting in a node feature matrix $X\in\mathbb{R}^{N\times M}$ (the feature vector of each node is M-dimensional) that can represent the texture features in ROIs.

To improve the performance of feature extraction, the pixel values in each frame need to be normalized before training. In addition, to accelerate the processing in GCN, the coordinate values of key points in facial images are aligned and normalized into the range of $[-1.0, 1.0]$.

3.2.2 Feature aggregation with GCN

To aggregate the local features, the graph is processed by GCN. The details can be shown as Fig. 4. The key points in the facial image are detected to construct the graph. The locations of these points are extracted to form the adjacency matrix according to the topology of constructed graph, which can reflect the structure information. The pixels around these points are reshaped into several vectors as the node attributes, which are the texture information in ROIs. Processed by GCN, the local structure and texture features can be aggregated to get higher level features with their structure property.

For a node feature matrix X, an adjacency matrix A and F convolution kernels, the feature mapping formula of GCN is:

$$Z=\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}X\Theta, \tag{2}$$

where $\tilde{A}=A+I_N$, I_N is the identity matrix. $\tilde{D}$ is the degree matrix of $\tilde{A}$, which is $\tilde{D}_{ii}=\sum_j\tilde{A}_{ij}$. $\Theta\in\mathbb{R}^{M\times F}$ is the parameter matrix and $Z\in\mathbb{R}^{N\times F}$ is the output matrix after convolution.

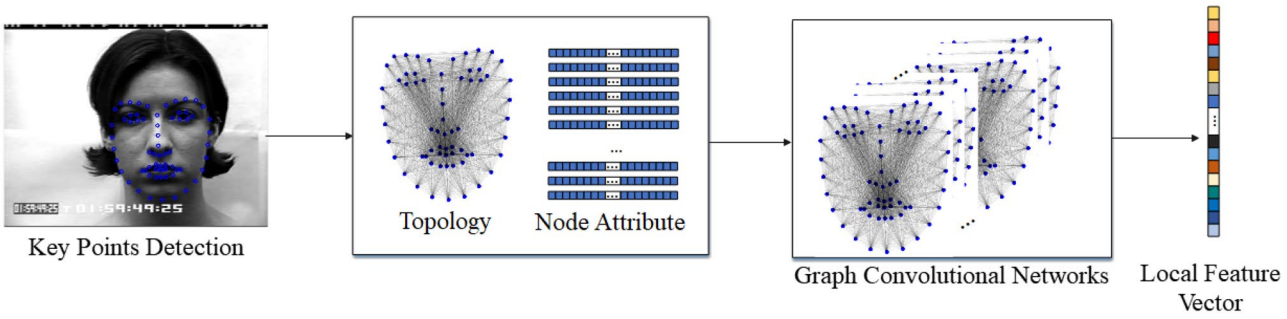

Fig. 4 The feature aggregation with GCN. The graph convolutional network extracts and aggregates two local facial features, structural features and texture features, from graphs

Then We consider an L-layer GCN, the idea of the GCN network is similar to that of the ordinary CNN. Its layer-to-layer propagation mode is as follows:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right), \tag{3}$$

where $W^{(l)} \in \mathbb{R}^{M_l \times M_{l+1}}$ is a trainable weight matrix. $H^{(l)} \in \mathbb{R}^{N \times M_l}$ is the feature of each layer. For the input layer, $H^{(0)} = X$. σ represents the activation function, such as

$$\mathrm{ReLU}(\cdot) = \max(0, \cdot). \tag{4}$$

Through several layers of GCN, the features of each node change from M_0 to M_L, but the adjacency matrix A in different layers are the same, which means the topological structure remains.

3.3 Classification

We directly concatenate the global and local feature vector to obtain a concatenate feature vector, which can be formulated as shown below:

$$v_c = (v_g, v_l), \tag{5}$$

where v_c, v_g and v_l denote concatenate feature vector, global feature vector and local feature vector respectively.

Then the concatenate feature vector is fed into a fully-connected layer to realize feature fusion. The fully-connected layer is also a classification layer and directly outputs the classification results. we choose softmax as the classifier, and its formula is as follows:

$$f(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}, \tag{6}$$

z_i is the output of each node of the fully-connected layer, $f(z_i)$ is the probability of various facial expressions, and the expression class with the largest value is regarded as the predicted class. Finally, Dropout [49] with 0.2 probability is added before the full-connected layer to avoid over fitting.

4 Experiments

4.1 Datasets and experiment settings

To evaluate the performance of the proposed model, we conduct experiments on three widely used data sets: CK+ [50], Oulu-CASIA [51] and MMI [52], and compare our model with the most advanced methods.

The extended Cohn–Kanade database (CK+) [50] includes 593 image sequences from 123 subjects, ranging in ages from 18 to 30 years old. 327 sequences of 118 subjects have facial expression tags (Anger, Contempt, Disgust, Fear, Happiness, Sadness and Surprise). All the expression sequences begin with neutral expression and gradually transition to peak expression. There are 5876 pictures in the CK+ dataset, with size of 640×490 pixels and 640×480 pixels, and they are grayscale or color.

The Oulu-CASIA dataset [51] contains 2280 image sequences with 6 basic facial expressions (Anger, Disgust, Fear, Happiness, Sadness and Surprise) from 80 subjects. Oulu-CASIA's video is captured under three lighting conditions (normal, weak and dark) through NIR (near infrared) and VIS (visible light). In the experiments, only 480 image sequences collected by VIS camera under normal conditions are used. Similar to CK + dataset, all expression sequences start from neutral stage and end with peak emotion. In each image sequence, only the last ten frames are selected, a total of 4800 pictures, and the image resolution is 320×240 pixels. Because of the lack of resolution, expression recognition on Oulu-CASIA database is more challenging.

The MMI data set [52] includes 32 subjects, ranging in age from 19 to 62 years old, of European, Asian or South American ethnicity. 213 video sequences are labeled with 6 basic expressions (Anger, Disgust, Fear, Happiness, Sadness and Surprise), of which 166 sequences have only positive

Table 2 The number of selected samples for each expression in the experiment

Dataset	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Contempt
CK+	1022	868	546	1331	547	1329	233
Oulu-CASIA	800	800	800	800	800	800	–
MMI	532	419	500	631	667	571	–

Fig. 5 Example images from CK+ (top), Oulu-CASIA (middle), and MMI (bottom). From left to right: anger, disgust, fear, happiness, sadness and surprise. The contempt expression of CK + is not displayed

faces. Different from the other two data sets, this data set starts with neutral expression, and then returns to neutral expression after the expression peak. We extract 3320 pictures from Oulu-CASIA video randomly. Although the resolution of the MMI data set is 576 × 720 pixels, the data set is also challenging due to the large differences in skin color and age, and some volunteers wear decorations, such as glasses.

The details of CK+, Oulu-CASIA and MMI data sets are listed in the Table 2. Some samples of these data sets are shown in the Fig. 5.

To evaluate the performance of the proposed method, we take the commonly used 10-fold cross-validation on these three data sets, which means the data set is randomly divided into 10 subsets of equal size. Nine of them are trained each time, and the other is tested. A total of 10 experiments are carried out. Finally, the average result of the experiment is taken as the expression recognition rate.

Our neural network is implemented by PyTorch and PyTorch Geometric framework. To detect the key points on the input frames, a machine learning toolkit Dlib library [53] is adopted in the implementation. In this toolkit, the facial key points are detected with the high-performance facial landmark detection methods [54]. The proposed model uses Adam to train 40 epochs, the batch size is 32, and the learning rate is fixed at 0.001.

4.2 Experimental results

The proposed method is compared with the existing state-of-the-art methods as listed in Table 3.

Table 3 The average accuracy of the proposed and existing methods on the CK+, Oulu-CASIA and MMI datasets respectively

Model	CK+ (%)	Oulu-CASIA (%)	MMI (%)
3DCNN-DAP [30]	92.40	–	63.40
DCMA-CNN [19]	93.46	–	–
IACNN [32]	95.37	–	71.55
IA-gen [35]	96.57	88.92	–
DTAGN [42]	97.25	81.46	70.24
DeRL [37]	97.30	88.00	73.23
GCNet [55]	97.93	86.11	81.53
FN2EN [56]	98.60	87.71	–
PPDN [18]	99.30	84.59	–
FER-GCN [45]	99.54	91.54	85.89
SD-CNN [57]	99.70	91.30	–
Proposed	99.78	98.62	97.92

4.2.1 Performance on CK+ dataset

CK+: It can be obtained that the average recognition accuracy of the proposed method can reach 99.78%, which is the highest accuracy among these existing methods. Compared with the FER-GCN, which is also an GCN based method, the performance of proposed method is much better. FER-GCN focuses on dynamic expression changes, and takes the GCN layer to learn an adjacency matrix representing the interdependence between different frames. But our proposed method adopts GCN to extract the local features in a single image, and it focus on the extraction of facial features. Compared with SD-CNN, which also aims at the static image, the proposed method can also improve the the recognition accuracy. The reason is that the proposed method takes the relationship between the expression feature information intensity and different regions of the face into consideration, but SD-CNN ignores the local regions with the most obvious facial movement when expressions occur, such as eyes, mouth and left and right cheeks.

Figure 6 shows the average confusion matrix results of 10-fold cross-validation on CK+ data set. It can be found that almost all expressions can be recognized, and disgust and surprise are relatively difficult to recognize.

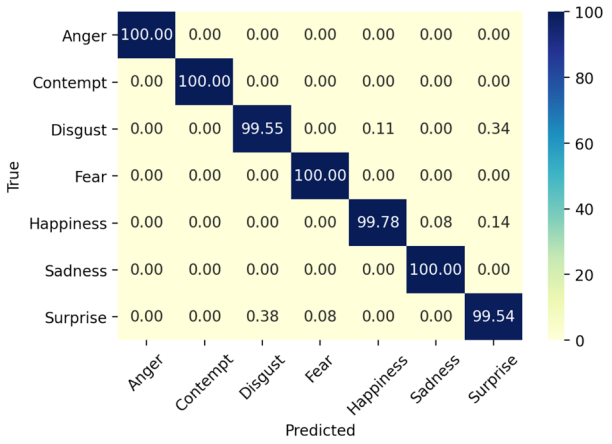

Fig. 6 The confusion matrix of averaged 10-fold cross-validation on CK+ dataset. Values are given in percent

4.2.2 Performance on Oulu-CASIA dataset

Oulu-CASIA: The experimental results of the proposed method on Oulu-CASIA data set are shown in Table 3. When the image resolution is insufficient, the result of our method can get the best recognition performance among all these methods, and the average accuracy is 98.62%, which can prove that the fusion of local features and global features represents facial expressions on different scales makes the expression more comprehensive. The results show that our two channel method has about 7% improvement over the FER-GCN, which proves the effectiveness and robustness of the proposed method. As shown in Fig. 7, the proposed method performs well on the expression of anger and happiness, but the recognition accuracy on the expression of disgust is relatively low. We can also find 1.06% of the samples in disgust are misclassified as anger. That is because that both of them are negative emotions, which have similar muscle movements, including wrinkling around the nose and upper lip and contraction between eyebrows.

4.2.3 Performance on MMI dataset

MMI: Due to the large differences in skin color and age of volunteers in the MMI data set, the performance on MMI is relatively lower than other data sets. But the proposed method can also obtains 97.92%, which is better than other methods. It can be obtained that the proposed method has wide adaptability and robustness, and can maintain its recognition ability in more strict cases. As shown in Fig. 8. The average accuracy of recognizing happiness expression is also the highest, reaching 99.55%. The average recognition accuracy of surprise is relatively low, but it still reaches 95.7%. In the experiment, nearly 2% of the surprise expression samples are incorrectly classified as fear. That is because that the behavior of surprise and fear expressions in several areas of the face (such as inner eyebrow and upper eyelid) is similar.

4.2.4 Performance and parameters quantity

Table 4 has listed the performance and quantity of parameters of the existing methods. It is obvious that the proposed method reaches the best performance among these methods with a small amount of parameters. The number of parameters for the proposed method is only larger than that in DCMA-CNN, but the accuracy is improved greatly compared with DCMA-CNN. The average recognition accuracy of the proposed method on CK+ is more than 6% higher than that of DCMA-CNN. Compared with DTAGN, the proposed method reduces the number of parameters by about 81%, but reaches higher performance, especially on MMI data, with

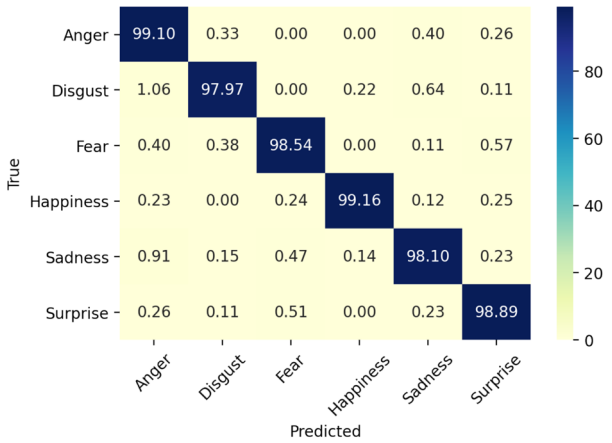

Fig. 7 The confusion matrix of averaged 10-fold cross-validation on Oulu-CASIA dataset

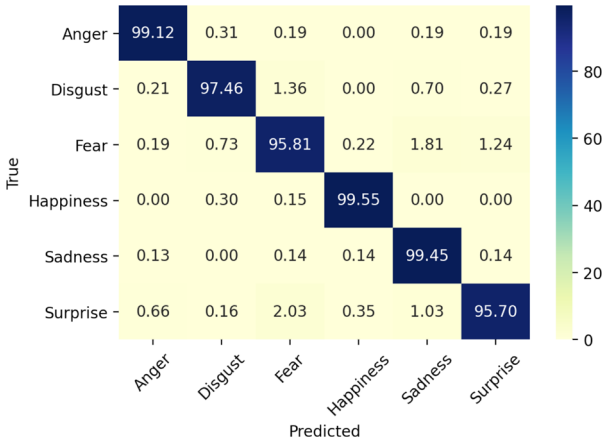

Fig. 8 The confusion matrix of averaged 10-fold cross-validation on MMI dataset

Table 4 Performance and parameters quantity of the proposed and existing methods

Model	CK+(%)	Oulu-CASIA(%)	MMI(%)	Parameters(M)
DCMA-CNN [19]	93.46	–	–	0.05
DTAGN [42]	97.25	81.46	70.24	5.85
FN2EN [56]	98.60	87.71	–	1.19
Proposed	99.78	98.62	97.92	1.09

a huge performance improvement of 27.68%. Although the number of parameter in the proposed method is only 0.1M less than that in FN2EN, but the average accuracy is much higher. This is because an additional GCN channel is used in the proposed method to extract local features, which can reduce the number of parameters in CNN channel and highlight the importance of local detail information in expression recognition.

4.3 Effectiveness analysis

In our proposed DCNN-GCN method, two channels are used to extract the global and local features, which are achieved with the CNN and GCN, respectively. Two models are constructed to evaluate the contribution of each channel to recognition. The model only utilizes global features to make recognition is denoted as GF-CNN. The model that recognizes expressions only with local features is denoted as LF-GCN. The recognition performances on the three expression data sets are experimented, and the results are listed in Table 5.

As it shown in Table 5, the average accuracy of total DCNN-GCN structure is higher than the other two networks. It is obvious that the performance of LF-GCN is quite low on expressions as the lack of global facial information, and the feature aggregation does improve the recognition of expressions. This is reasonable because global or local features only focus on expressing information on a specific scale. The improvement of recognition accuracy by aggregation shows that the two features are complementary.

Table 5 Comparison of performance and model sizes between single channel network and dual channel network

Model	CK+ (%)	Oulu-CASIA (%)	MMI (%)	Parameters (*M*)
LF-GCN	23.99	18.81	20.72	0.01
GF-CNN	99.48	97.66	97.07	1.07
DCNN-GCN	99.78	98.62	97.92	1.09

The average accuracy of DCNN-GCN using both global and local features is higher than LF-GCN and GF-CNN using only one feature

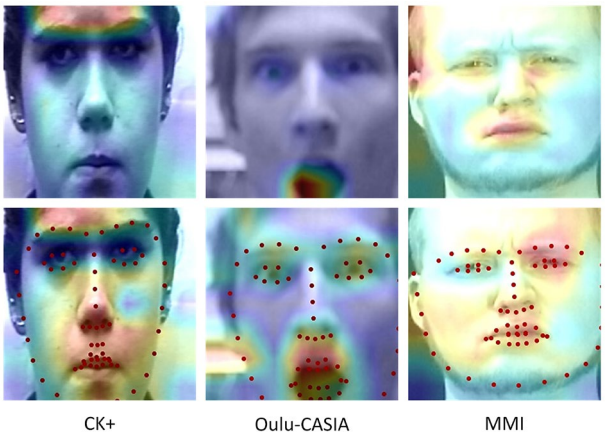

Fig. 9 Visualization of high-level features. High-level features extracted by GF-CNN are shown in top line, which focuses on different areas in the frame. The High-level features extracted by proposed DCNN-GCN are listed in bottom line, which can combine the features extracted in the two channels and target the most contributing expression areas (like mouth and eyes here)

We further provide visualization to prove the effectiveness of our method. As shown in the Fig. 9, we select some samples from the three datasets to visualize the high-level features extracted from the images by GF-CNN and our proposed DCNN-GCN. GF-CNN (top line) focuses on different areas in the frame, but our method (DCNN-GCN in bottom line) can combine the features extracted in the two channels and target the most contributing expression areas (like mouth and eyes here). The results show that our method aggregates global and local features which makes the model pay more attention to the corresponding expression region.

5 Conclusion

In this paper, we propose a lite dual channel facial expression recognition method, which focuses on the global and local ROI features to improve the performance of expression recognition. The proposed method consists of two channels, which aim to extract the global and local features respectively. The local features are modeled as a graph, which can extract the local ROI features with remaining the topological structure property. The facial key points are detected as the ROI, modeled as nodes of the graph for expression recognition. The distance between every two nodes are regarded as weights of edges, and the pixel values around these key points are attributes. The local features are aggregated into some higher level features with GCN with the structure property. Then the global and local features are processed by a classifier to achieve the expression recognition. The proposed method is implemented and evaluated on three

widely used data sets. Experimental results show that the proposed method can efficiently improve the accuracy of facial expression recognition, and the network is much more lite, which is more suitable for application.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (No. 61802105, 61976078), the University Synergy Innovation Program of Anhui Province (No. GXXT-2021-005, GXXT-2020-014), and Natural Science Foundation of Anhui Province (No. 1908085QF265, 2108085MF203).

References

1. Chen, L., et al.: Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction. Inf. Sci. **428**, 49–61 (2018)
2. Jabon, M., Bailenson, J., Pontikakis, E., Takayama, L., Nass, C.: Facial expression analysis for predicting unsafe driving behavior. IEEE Pervasive Comput. **10**(4), 84–95 (2010)
3. Chen, J., Wang, G., Zhang, K., Wang, G., Liu, L.: A pilot study on evaluating children with autism spectrum disorder using computer games. Comput. Hum. Behav. **90**, 204–214 (2019)
4. Ekman, P., Friesen, W.: Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. **17**(2), 124–129 (1971)
5. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis[J]. Chemometr. Intell. Lab. Syst. **2**(1–3), 37–52 (1987)
6. Hyvarinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Netw. **13**(4), 411–430 (2000)
7. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Recognition using class specific linear projection. In: European Conference on Computer Vision, pp. 43–58 (1966)
8. Ekman, P., Friesen, W.V.: Facial action coding system (facs): a technique for the measurement of facial actions. Riv. Psichiatr. **47**(2), 126–38 (1978)
9. Bartlett, M. S., et al.: Recognizing facial expression: machine learning and application to spontaneous behavior. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2**, 568–573 (2005)
10. Shan, C., Gong, S., Mcowan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis. Comput. **27**(6), 803–816 (2009)
11. Huang, M.-W., Wang, Z.-w., Ying, Z.-L.: A new method for facial expression recognition based on sparse representation plus lbp. In: 2010 3rd International Congress on Image and Signal Processing **4**, 1750–1754 (2010)
12. Liu, X., Yang, X., Wang, M., Hong, R.: Deep neighborhood component analysis for visual similarity modeling. ACM Trans. Intell. Syst. Technol. (TIST) **11**(3), 1–15 (2020)
13. Yang, X., Zhou, P., Wang, M.: Person reidentification via structural deep metric learning. IEEE Trans. Neural Netw. Learn. Syst. **30**(10), 2987–2998 (2018)
14. Cheng, S., Zhou, G.: Facial expression recognition method based on improved vgg convolutional neural network. Int. J. Pattern Recognit. Artif Intell. **34**(07), 2056003 (2020)
15. Zhong, Y., Qiu, S., Luo, X., Meng, Z., Liu, J.: Facial expression recognition based on optimized resnet. In: 2020 2nd World Symposium on Artificial Intelligence, pp. 84–91 (2020)
16. Mollahosseini, A., Chan, D., Mahoor, M.H.: Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter Conference on Applications of Computer Vision, pp. 1–10 (2016)
17. Hamester, D., Barros, P., Wermter, S.: Face expression recognition with a 2-channel convolutional neural network. In: 2015 International Joint Conference on Neural Networks, pp. 1–8 (2015)
18. Zhao, X., et al.: Peak-piloted deep network for facial expression recognition. In: European Conference on Computer Vision, pp. 425–442 (2016)
19. Xie, S., Hu, H.: Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. IEEE Trans. Multimed. **21**(1), 211–220 (2018)
20. Wang, Z., Ying, Z.: Facial expression recognition based on local phase quantization and sparse representation. In: 2012 8th International Conference on Natural Computation, pp. 222–225 (2012)
21. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Comput. Vis. Image Underst. **61**(1), 38–59 (1995)
22. Zhong, L., et al.: Learning active facial patches for expression analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2562–2569 (2012)
23. Bandrabur, A., Florea, L., Florea, C., Mancas, M.: Emotion identification by facial landmarks dynamics analysis. In: 2015 IEEE International Conference on Intelligent Computer Communication and Processing, pp. 379–382 (2015)
24. Desrosiers, P.A., Daoudi, M., Devanne, M.: Novel generative model for facial expressions based on statistical shape analysis of landmarks trajectories. In: 2016 23rd International Conference on Pattern Recognition, pp. 961–966 (2016)
25. Meng, L., et al.: Learning using privileged information for food recognition. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 557–565 (2019)
26. Dong, J., et al.: Fine-grained fashion similarity prediction by attribute-specific embedding learning. IEEE Trans. Image Process. **30**, 8410–8425 (2021)
27. Simonyan, K., Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2015)
28. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European Conference on Computer Vision, pp. 630–645 (2016)
29. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
30. Liu, M., Li, S., Shan, S., Wang, R., Chen, X.: Deeply learning deformable facial action parts model for dynamic expression analysis. In: Asian Conference on Computer Vision, pp. 143–157 (2014)
31. Gan, Y., Chen, J., Yang, Z., Xu, L.: Multiple attention network for facial expression recognition. IEEE Access **8**, 7383–7393 (2020)
32. Meng, Z., Liu, P., Cai, J., Han, S., Tong, Y.: Identity-aware convolutional neural network for facial expression recognition. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition, pp. 558–565 (2017)
33. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6897–6906 (2020)
34. Yang, H., Ciftci, U., Yin, L.: Facial expression recognition by de-expression residue learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2168–2177 (2018)
35. Yang, H., Zhang, Z., Yin, L.: Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition, pp. 294–301 (2018)

36. Agrawal, A., Mittal, N.: Using cnn for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. Vis. Comput. **36**, 405–412 (2019)
37. Chen, Y., Hu, H.: Facial expression recognition by inter-class relational learning. IEEE Access **PP**(99), 1–1 (2019)
38. Wang, F., Shen, L.: Expression recognition using region features and facial action units. In: 2019 15th International Conference on Intelligent Environments, pp. 9–15 (2019)
39. Xu, Q., Zhao, N.: A facial expression recognition algorithm based on cnn and lbp feature. 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference **1**, 2304–2308 (2020)
40. Nguyen, D.H., Kim, S., Lee, G.S., Yang, H.J., Na, I.S., Kim, S.H.: Facial expression recognition using a temporal ensemble of multi-level convolutional neural networks. IEEE Trans. Affect. Comput. **13**(1), 226–237 (2022). https://doi.org/10.1109/TAFFC.2019.2946540
41. Li, Y., Zeng, J., Shan, S., Chen, X.: Occlusion aware facial expression recognition using cnn with attention mechanism. IEEE Trans. Image Process. **28**(5), 2439–2450 (2018)
42. Jung, H., Lee, S., Yim, J., Park, S., Kim, J.: Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2983–2991 (2015)
43. Kipf, T. N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
44. Yang, X., Du, X., Wang, M.: Learning to match on graph for fashion compatibility modeling. In: Proceedings of the AAAI Conference on Artificial Intelligence **34**(01), 287–294 (2020)
45. Liu, D., Zhang, H., Zhou, P.: Video-based facial expression recognition using graph convolutional networks. In: 2020 25th International Conference on Pattern Recognition, pp. 607–614 (2021)
46. Kumar, A.J.R., Bhanu, B.: Micro-expression classification based on landmark relations with graph attention convolutional network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1511–1520 (2021)
47. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural. Inf. Process. Syst. **25**, 1097–1105 (2012)
48. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456 (2015)
49. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
50. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp. 94–101 (2010). https://doi.org/10.1109/CVPRW.2010.5543262
51. Zhao, G., Huang, X., Taini, M., Li, S.Z., Pietikainen, M.: Facial expression recognition from near-infrared videos. Image Vis. Comput. **29**, 607–619 (2011)
52. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: 2005 IEEE international conference on multimedia and Expo, p. 5 (2005)
53. King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
54. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867–1874 (2014)
55. Kim, Y., Yoo, B., Kwak, Y., Choi, C., Kim, J.: Deep generative-contrastive networks for facial expression recognition. arXiv preprint arXiv:1703.07140 (2017)
56. Hui, D., Zhou, S.K., Chellappa, R.: Facenet2expnet: regularizing a deep face recognition net for expression recognition. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE, pp. 118–126 (2017)
57. Liu, L., Jiang, R., Huo, J., Chen, J.: Self-difference convolutional neural network for facial expression recognition. Sensors **21**(6), 2250 (2021)